Artificial Intelligence & Environment

https://doi.org/10.66178/aie-0026-0011

Research

Open Access

Multimodal large language models for big data synthesis: Addressing the data bottleneck of plastic chemicals

Yunhan Cui, Yuxuan Zhang, Haobo Wang, Wenjia Liu, Zijun Xiao, Yuwei Liu, Jiale He, Guijia Bi and Jingwen Chen*

Citation: Cui Y; Zhang Y; Wang H; et al. Multimodal large language models for big data synthesis: Addressing the data bottleneck of plastic chemicals. *AI Environ.* 2026, 1(2): 83-92.

Received: 20 March 2026

Revised: 12 April 2026

Accepted: 28 May 2026

Published online: 16 June 2026

Abstract

The global plastic pollution is escalating, driven in part by hazardous plastic chemicals. However, the absence of comprehensive data on chemical exposure and hazards, particularly concerning economic indicators, physicochemical properties, environmental behavior parameters, toxicity profiles, and regulation status, severely constrains effective decision-making and policy formulation. This study developed a multimodal large language model-based framework for extracting economic indicators of plastic chemicals from unstructured sources. From 36,289 images and 969 literature sources, 44,249 entries covering 814 chemicals were extracted across 9 economic indicators, including production volume, production value, production capacity, market size, consumption, and import/export volumes and values. The pipeline achieved an average F_1 score of 92.8% (95% confidence interval: 89.9–95.6%), demonstrating high accuracy and comprehensiveness. Toxicity data for 4,551 chemicals, environmental behavior parameters for 9,770 chemicals, and regulatory information for 13,684 chemicals from 111 global lists were also manually compiled. A graph attention network model was employed to predict 11 physicochemical properties as well as the persistence, bioaccumulation, mobility, and toxicity (PBT/PMT) properties of 9,768 chemicals. This study constructed the most up-to-date and comprehensive dataset with multiple data types for plastic chemicals, providing an AI-based methodological foundation for synthesizing big data to support the sound management of plastic chemicals.

Keywords: Plastic chemicals, Multimodal large language model, Economic indicators, Physicochemical properties, Toxicity, Environmental behavior parameters, Regulatory lists

Highlights

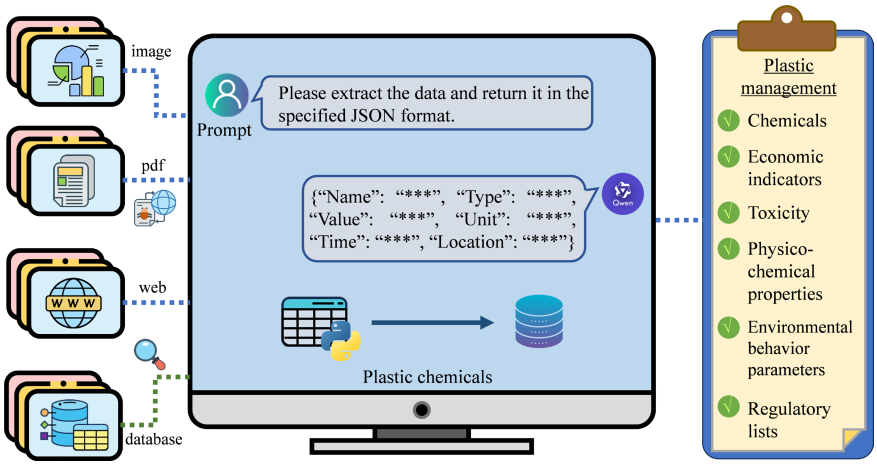
- LLM-based data extraction pipeline achieved 92.8% F_1 score (95% CI: 89.9–95.6%).
- Over 40,000 records across 9 economic indicators for 814 plastic chemicals.
- Toxicity, environmental behavior, and regulatory data for > 13,000 plastic chemicals.
- GAT model predicts 11 physicochemical property/PBT/PMT attributes for > 9,000 chemicals.

* Correspondence: Jingwen Chen (jwchen@dlut.edu.cn)

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

https://doi.org/10.66178/aie-0026-0011

Graphical abstract



Introduction

The accelerating global production of plastics continues to exacerbate plastic pollution worldwide[1]. In response, the United Nations Environment Assembly has adopted a resolution calling for international action to prevent, reduce, and ultimately eliminate plastic pollution[2]. A central pillar of the proposed global plastics treaty is the sound management of plastic chemicals[3]. The "plastic chemicals" encompass not only polymer backbones but also intentionally added substances, such as processing aids and additives[4], as well as non-intentionally added substances (NIAS), including impurities, reaction by-products, and degradation products[4,5].

To date, more than 16,000 chemicals are known or suspected to be present in plastic materials and products[6], many of which have been documented to leach into the environment[7] and pose risks to ecosystems, wildlife, and human health[8]. Strengthening the scientific basis for managing plastic chemicals is therefore essential to addressing plastic pollution at its sources.

Effective management of plastic chemicals requires an integrated understanding of their production and consumption patterns, physicochemical properties, environmental behavior, toxicological profiles, and regulatory status[9,10]. However, such information remains highly fragmented across heterogeneous sources and is often reported in unstructured formats[11], which severely hampers systematic data integration. Existing chemical databases frequently lack quantitative details or key contextual attributes. For example, the OECD high production volume (HPV) chemicals database chemical list[12] classifies substances only as being produced above a threshold level, without specifying exact production quantities, timeframes, geographic distribution, or data provenance. These limitations significantly constrain robust risk assessment and evidence-based decision-making, underscoring the urgent need for a comprehensive and systematic dataset of plastic chemicals.

Historically, the construction of chemical datasets has relied on manual data retrieval or conventional rule-based text mining approaches[13] applied to a limited corpus of literature. Such methods are labor-intensive, difficult to scale, and prone to data incompleteness. Moreover, critical information on chemical production, trade, and market dynamics is frequently presented in graphical formats, such as charts and figures embedded in online reports and webpages[14]. The vast number of such images available online further limits the effectiveness of traditional text mining techniques, highlighting the need for more advanced approaches capable of extracting structured information from diverse, unstructured sources.

Recent advances in large language models (LLMs) have transformed text mining by enabling direct semantic understanding of complex texts. LLMs such as GPT-4[15] have been successfully applied to automated data extraction in materials science[16,17], medicine[18], and toxicology[19], achieving high accuracy. However, most state-of-the-art LLMs remain closed-source, resulting in prohibitive costs for large-scale applications, and are primarily optimized for English-language text, with limited capacity for image data or non-English sources. The emergence of open-source multimodal LLMs offers a promising alternative. In particular, Qwen2-VL[20] demonstrates strong performance in both text and image understanding, with notable proficiency in Chinese-language content, enabling scalable and cost-effective data extraction across multilingual and multimodal sources.

In this study, a workflow was developed for automated, structured, and high-throughput extraction of economic indicator data for plastic chemicals from both textual literature and images, using open-source multimodal LLMs (Qwen2-VL-72B and Qwen2.5-32B)[20]. The workflow enabled the generation of a large-scale, multi-dimensional dataset comprising 44,249 records for 814 chemicals. The dataset includes production volumes, production values, production capacities, market sizes, consumption volumes, and import and export volumes and values, with each record accompanied by detailed temporal, spatial, and source information.

The dataset was further enriched through integration of manually curated information and in silico predictions, incorporating 11 physicochemical properties as well as persistence, bioaccumulation, mobility, and toxicity (PBT and PMT) properties, 67 toxicity endpoints, 12 environmental behavior parameters, and 111 chemical inventories or regulatory lists from 26 countries and regions. Together, the proposed workflow and dataset provide a robust foundation for advancing the sound management of plastic chemicals and supporting evidence-based policy development in the global effort to address plastic pollution.

Materials and methods

The dataset for plastic chemicals constructed in this study integrates four categories of data obtained through different approaches (Figure 1). Economic indicator data were extracted using an LLM-based

workflow. Toxicity data and environmental behavior parameters were compiled through a systematic literature review. Physicochemical properties, as well as PBT and PMT properties, were predicted using established computational toxicology models. Regulatory information was retrieved via systematic searches of official regulatory agency websites.

Economic indicator data

The LLM-based workflow for economic indicator data mining primarily comprises four stages: (1) data source identification, (2) data collection, (3) data extraction, and (4) data processing (Figure 1).

Data source identification

A systematic search was conducted on the Web of Science platform and an industry report website (e.g., the United Nations Environment Programme, UNEP)[21] using the keywords "plastic additives" and "plastics". Two comprehensive datasets on plastic chemicals were identified: the PlastChem dataset[5,6] developed by the Norwegian University of Science and Technology, and the dataset compiled by the UNEP, referred to as "CIP" in the "Chemicals in Plastics: A Technical Report" [11]. As the PlastChem dataset (Supplementary Table S1) fully encompasses all entries included in the CIP dataset (Supplementary Table S8), it was selected as the primary data source. Details on the manual data verification process (Supplementary Fig. S1) are provided in the Supporting Information (SI). The final inventory comprised 17,932 plastic chemicals, of which 16,325 were associated with valid Chemical Abstracts Service (CAS) registry numbers (Supplementary Table S2). Core chemical identifiers, including the CAS numbers, Simplified Molecular Input Line Entry System (SMILES) strings, and chemical names from the PlastChem dataset[5,6], were used for subsequent data integration and analysis.

Economic indicator data for plastic chemicals, including production volume and value, production capacity, market size, consumption, and import/export volumes and values (detailed in the SI), were collected primarily from online images and literature reviews. The images were obtained from three publicly accessible industrial websites (AskCI[22], Fxbaogao[23], and Qianzhan[24]), which provide market statistics and industry data in both Chinese and English. These platforms are not only leaders in industrial consulting but also provide data sourced from authoritative institutions such as the National Information Center and the Bureau of Statistics, and are widely recognized and utilized across the industry. The literature reviews were obtained from an open-access literature database (Cqvip[25]). Bibliographic platforms including Web of Science[26] and PubMed[27] were not used as data sources due to scope limitations (e.g., the absence of review articles reporting market-size statistics for specific chemicals).

Data collection

The economic indicator data were primarily retrieved from websites in image formats (e.g., bar, pie, and line charts). To collect this information systematically, chemical names from the plastic chemical inventory (Supplementary Table S2) were combined with relevant keywords (e.g., "production volume") to manually download image files (JPG format) from selected websites. Similarly, academic literature (PDF format) was downloaded from the Cqvip platform[25] using the same set of keywords, with January 2, 2025 as the cutoff date.

The filenames of the downloaded files retained their original naming conventions, summarizing the content and key indicators covered. To ensure relevance to economic indicators, a keyword-based screening approach was applied, with selected terms including "prospects", "outlook", "progress", "development", "market", "production", "application", "industry", "demand", "current situation", and "trend". In total, 36,289 online images and 969 literature documents were collected related to the economic indicators of plastic chemicals.

Data extraction

To extract structured data from both image-based charts and textual literature, multimodal LLMs, Qwen2-VL-72B and Qwen2.5-32B[20], were deployed locally. The models' outputs were guided by specifically designed prompts that incorporated clear logic, explicit rules, and repeated reminders of common errors to mitigate issues such as machine hallucinations.

Qwen2-VL-72B, developed by Alibaba's DAMO Academy, was a flagship open-source visual multimodal model as of September

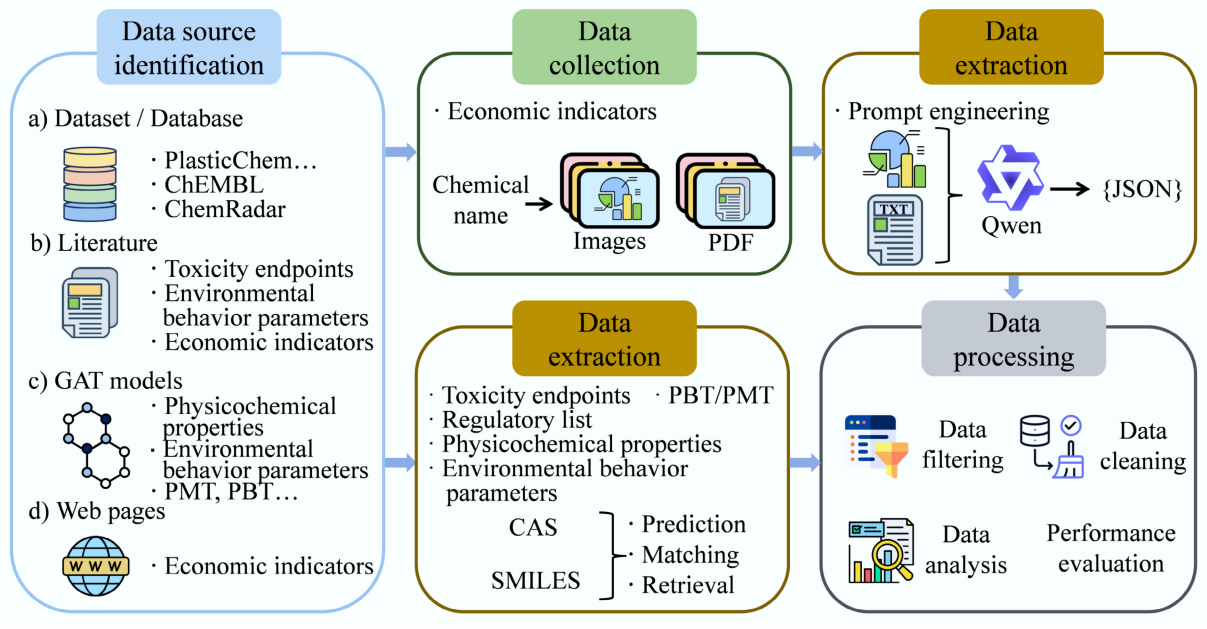


Fig. 1 Schematic overview of the workflow for the dataset construction.

https://doi.org/10.66178/aie-0026-0011

2024[28]. It outperforms the closed-source model GPT-4o in benchmarks such as document and diagram reading[29]. For instance, in a chart-based question-answering task, Qwen2-VL-72B scored 88.3, surpassing GPT-4o's score of 85.7[28]. Qwen2.5-32B, excelling in text comprehension[29], scored 69 in a multidisciplinary knowledge understanding benchmark, outperforming GPT-4o Mini (score 63.1)[29]. Therefore, Qwen2-VL-72B was selected for image data extraction, and Qwen2.5-32B was used for text.

The extraction prompt consisted of three components: extraction requirements, detailed instructions, and critical notes. The extraction requirements specified a standardized output format, with fields for "chemical name", "economic indicator type", "numerical value", "unit", "year", "month", and "location". The detailed instructions clarified rules for generating each output element, with 9 specific economic indicators listed. Missing fields were labeled as "no information", and irrelevant data were filtered out by the critical notes. The full text of the prompts is provided in the SI. To enhance text extraction accuracy, the PDF files were converted to plain text using the PyMuPDF Python package[30]. All the extraction codes are available at https://github.com/johannatsui-cyhan/Ciplastic.

Data processing

The chemical names associated with the economic indicator data were mapped to their CAS numbers using chemical dictionary tools (GuideChem and Ichemistry)[31,32]. The CAS numbers were cross-referenced with the plastic chemical inventory to retain only the relevant data. After screening, there were 6,267 images and 373 literature sources that contained valid information. A thorough manual data cleaning process was followed, including unit standardization, deduplication, and verification, which was detailed in the SI (Supplementary Table S3).

Precision (R_p), recall (R_r), and F_1 score are commonly employed evaluation metrics for assessing the reliability of LLMs in text mining tasks[13,16,19]. These metrics were computed by selecting a sample of images and literature (validation set) and comparing the automatically extracted data records with those obtained through manual annotation. In this study, a validation set comprising 101 images and literature sources (755 entries) containing all 9 economic indicator data was selected by stratified sampling (detailed in the SI). The R_p, R_r, and F_1 score were subsequently calculated:

$$R_p = N_{TP}/(N_{TP} + N_{FP}) \quad (1)$$

$$R_r = N_{TP}/(N_{TP} + N_{FN}) \quad (2)$$

$$F_1 = 2 \times R_p \times R_r/(R_p + R_r) \quad (3)$$

where N_{TP}, N_{FP}, and N_{FN} represent the number of true positives, false positives, and false negatives in the validation dataset, respectively. The F_1 score is the harmonic mean of R_p and R_r, with a higher value indicating better performance of the LLM-based data extraction.

Toxicity data and environmental behavior parameters

To compile toxicity-related and environmental behavior parameters (Supplementary Table S9), manual searches were performed on the Web of Science platform[26] and across multiple open databases (detailed in the SI). The canonicalized SMILES strings (detailed in the SI) of plastic chemicals were manually matched with the corresponding entries in the compiled dataset, resulting in a comprehensive dataset on plastic chemical properties (Supplementary Table S4).

Physicochemical, PBT, and PMT properties

In-house graph attention network (GAT) models were applied to predict physicochemical properties (e.g., vapor pressure, water solubility, and Henry's law constant), environmental behavior parameters (e.g., biodegradation half-lives in water and sediment), and key properties (e.g., PBT and PMT) (Supplementary Table S5, S10, and S11). The SMILES strings of plastic chemicals were used as input for the GAT-based models (detailed in the SI).

Regulatory information

To compile global chemical inventories and regulatory lists relevant to plastic chemicals, ChemRadar[33] was first used as a screening tool to identify national and regional chemical inventories and regulatory lists worldwide. Subsequently, the original official sources of all identified inventories and regulatory lists were manually consulted. The inclusion status of plastic chemicals in each list was determined based on their CAS numbers (Supplementary Table S6). In addition, the most recent update date and geographic scope of each inventory or regulatory list were documented (Supplementary Table S7).

Results and discussion

Performance of the LLM-based data extraction

The performance of the LLM-based data extraction demonstrates robust and reliable results when applied to tasks involving the 9 key economic indicators, particularly with the effective prompt engineering strategies. For images in the validation set containing the target information, the model achieves an R_p of 91.1%, an R_r of 95.0%, and an F_1 score of 92.9% (Supplementary Table S12). For literature-based sources in the validation set, the R_p, R_r, and F_1 score are 95.2%, 88.2%, and 91.6%, respectively (Supplementary Table S12). These outcomes reflect the model's high accuracy and completeness in data extraction. The Qwen2.5-32B and Qwen2-VL-72B models outperform other models in similar text extraction tasks, such as Llama-2 (R_p = 83.6%, R_r = 82.1%) for material-related data[16], and GPT-4 (R_p = 91.0%, R_r = 81.0%) for toxicological data[19].

A retrospective error analysis identified the primary causes of false positives (N_{FP}) and false negatives (N_{FN}) in the workflow (Supplementary Table S13 and S14). In the case of images, the false positives often resulted from vague or overlapping data labels, leading to misclassification. The false negatives were mainly caused by incomplete or ambiguous captions missing essential contextual information, such as geographic regions or time periods. For instance, the title of an image might read "RedSun's chlorantraniliprole industry chain capacity planning table from 2024 to 2027." While the model correctly identified the data, it missed the company name "RedSun Group", which led to the omission of associated location data. These observations suggest that many false negatives were caused by missing entity-level information rather than model reasoning errors, a limitation that is often addressed through external knowledge grounding methods such as Retrieval-Augmented Generation (RAG)[34]. However, the fragmented and weakly structured nature of the available data precluded the construction of a reliable retrieval corpus, preventing the systematic use of RAG in this study.

In textual literature, the false negatives were typically caused by vague phrasing where essential details, such as the type of economic indicator, spatial context, or temporal reference, were implied but not explicitly stated. For example, the statement "the import volume of polyethylene was 4.9 million tons last year" without a specified year may result in missed extraction of the data. The false positives occurred when irrelevant terms, such as "demand volume" or "growth rate", were mistakenly identified as valid data points, leading to the extraction of unnecessary information. Future enhancement may involve the introduction of an automated

feedback loop mechanism[35–39], where the model generates its own feedback and refines its prompts accordingly, or incorporates domain-specific examples to optimize performance. This iterative process would enhance model accuracy and ensure it adapts to evolving data patterns.

Compared with traditional multimodal deep learning methods, LLMs offer key advantages for extracting chemical information from images and literature. They achieve high extraction performance with minimal or even zero labeled data, whereas conventional models often require large annotated datasets[34]. Unlike rule-based approaches[13], LLMs do not rely on manually crafted patterns, providing greater flexibility and robustness for diverse and complex chemical information. Moreover, they enable efficient and accurate information extraction, reducing human effort and potential bias. These features highlight the necessity and practical advantages of using LLMs in constructing plastic chemical datasets.

Crucially, the developed workflow overcomes the limitations of traditional dataset construction by enabling the batch extraction of structured data from a large corpus of unstructured images (36,289 items) and literature (969 documents).

Economic indicator data coverage

Economic indicator data on plastic chemicals are currently limited, with only a few registered production volumes reported in European and American countries. These data often rely on coarse categorical ranges, such as 10, 100, or 1,000 metric tons (Supplementary Table S1), and lack clearly defined temporal or spatial boundaries. Data from other regions and across different scales remain fragmented and sporadic. These limitations impede a comprehensive and accurate assessment of global plastic chemical usage. Consequently, advanced scientific and technological approaches, such as the AI-based methods, are required to address the data gaps.

With the LLM-based workflow, the newly constructed dataset extends beyond production volume to include eight additional economic indicators, namely production capacity, production value, market size, consumption volume, import and export volumes, and their corresponding monetary values (Supplementary Table S3). These indicators are reported as specific values across multiple temporal and geographical scales, substantially expanding the resolution and scope of economic information on plastic chemicals. In total, the dataset comprises 44,249 entries covering 814 plastic chemicals. All the data are standardized in scientific notations with consistent units, unique identifiers, and detailed metadata, thereby conforming to the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles[40]. It should be noted that the reported indicators reflect the total industrial usage across all application sectors. Disaggregation to plastic-specific applications would require additional sector-specific allocation assumptions, which may be addressed in future application-oriented analyses. Additionally, the dataset generated using LLM-based text mining may include some inaccuracies and should be applied with caution in subsequent analyses.

Among all the indicators, production volume is the most frequently reported (16,224 records across 461 chemicals), followed by production capacity (7,896 entries) and consumption volume (7,308 entries), as shown in Figure 2. Seventeen chemicals with available economic data are classified as substances of very high concern[41], among which bisphenol A exhibits the largest number of records (225 entries), reflecting its high production volume and well-documented environmental and health risks[42]. Economic indicator data are also relatively abundant for per- and polyfluoroalkyl substance (PFAS)-related plastic chemicals, such as polytetrafluoroethylene, tetrafluoromethane, and polyvinylidene fluoride, providing a quantitative basis for their exposure and risk assessments.

The economic indicators such as production, consumption, and trade volumes constitute essential inputs for material flow analysis (MFA)[43–45], a core approach in industrial ecology for quantifying anthropogenic chemical cycles[46–49]. MFA applies mass balance principles to estimate chemical stocks, flows, and releases across life-cycle stages. However, traditional MFA studies have relied on manually extracted data from statistical yearbooks and literature,

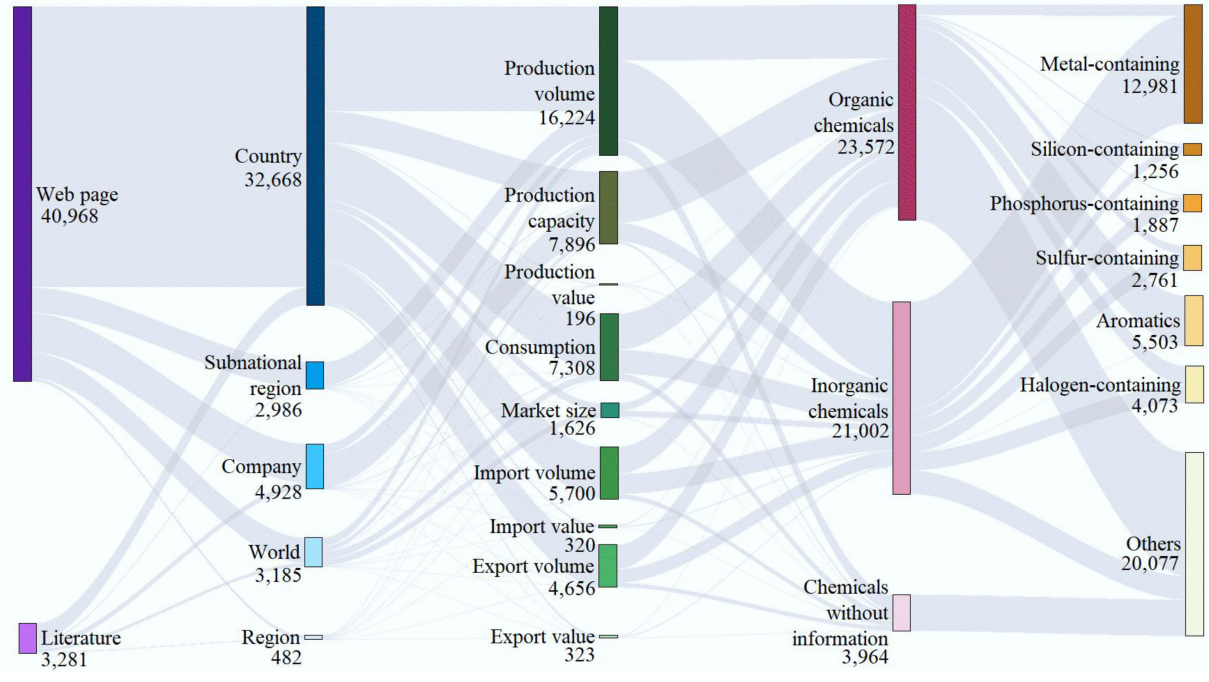


Fig. 2 Data size of the 9 collected economic indicator data in the current dataset (unit: entry).

https://doi.org/10.66178/aie-0026-0011

resulting in substantial labor requirements and limited chemical coverage. A review of published studies indicates that only 269 plastic chemicals were investigated with MFA between 2000 and 2022[43], underscoring the persistent data gaps. The comprehensive economic indicators provided herein are conducive to addressing the data gaps and enabling broader MFA applications.

Dynamic MFA further requires long-term time series data to capture the evolution of in-use stocks and delayed emissions[44]. Such data have been largely absent from existing resources[5,6]. The present dataset may fill this gap by spanning both historical (1886–2024) and projected (2025–2060) periods. The dataset includes annual (70.5%), quarterly (1.3%), and monthly (28.2%) observations (Figure 3), supporting high-resolution temporal analyses and scenario-based projections.

From a spatial perspective, prior MFA studies have predominantly focused on regional[50–52] or global scales[53,54], largely due to the scarcity of spatially resolved chemical data. The current dataset provides multi-scale coverage, including global, regional (e.g., Asia), national (e.g., Australia), subnational (e.g., Beijing), and company-level (e.g., Mitsubishi Chemical Group) indicators (Figure 2). The national-level data dominate the dataset, comprising 32,668 entries from 73 countries. Notably, China, accounting for over 40% of the global chemical consumption in 2023[55], has been substantially underrepresented in earlier datasets. This study addresses this critical gap by incorporating 30,808 entries of Chinese economic indicators across all nine indicators, thereby supplementing necessary data for global material flow and risk assessments of plastic chemicals.

To demonstrate practical applicability, the dataset was used to conduct a dynamic MFA of polytetrafluoroethylene (PTFE) in China from 2001 to 2025, as detailed in the SI (Supplementary Table S15 and S16). The results indicate steadily growing in-use stocks and waste generation of PTFE, reaching 1.40×10^9 kg and 5.41×10^7 kg in 2025, respectively. The emissions are dominated by the waste disposal stage, with substantial releases of both micro- and macroplastics, primarily to soil. This case study illustrates how the dataset enables integrated assessments of sources, flows, and sinks of plastic chemicals.

The finest spatial resolution of the dataset is at the company level, comprising 4,928 entries. The company-level indicators not only support national aggregation but also enable the identification of pollution hotspots and targeted mitigation strategies. Such granularity is particularly relevant for policy design and corporate accountability, especially for producers of environmentally preferable chemicals[56]. Transparent reporting at this scale may facilitate informed decision-making and support sustainable industrial transitions.

Finally, the dataset identifies 423 chemicals with production volumes exceeding 1.00×10^6 kg·year^{-1} at national, subnational, or company levels, meeting the threshold for high-production-volume (HPV) chemicals[12]. Applying the criterion, 522 chemicals exceed this limit in production capacity, 285 in consumption, and 265 in trade volumes. Many of these high-volume substances fall into categories of regulatory priority[57,58], including PFASs, organometallic compounds, and ortho-phthalates. These findings highlight the necessity of incorporating consumption and trade data alongside production figures in chemical risk assessments. The enriched dataset presented herein may provide a robust foundation for such comprehensive evaluations.

It should be noted that the economic indicators of chemicals, including production and trade volumes, are inherently time-dependent variables rather than fixed attributes. They continuously evolve in response to market demand, regulatory interventions, technological substitution, and broader socio-economic conditions. Consequently, there is no single, temporally invariant value that can fully represent real-world chemical production or usage. Even official statistics, such as those reported through the EU-REACH registration system, are typically derived from periodic reporting, declarations, or estimations, and therefore reflect institutionally bounded snapshots rather than comprehensive or real-time market realities.

At the global scale, the chemical economic data are further characterized by substantial sparsity and heterogeneity. Available information is fragmented across regions, regulatory systems, and industrial sectors, and is often collected for purposes unrelated to chemical risk assessment. Variations in reporting boundaries, statistical methodologies, confidentiality constraints, and human reporting practices introduce systematic inconsistencies among data sources. Under these conditions, the primary methodological challenge is not the recovery of an unattainable ground truth, but the integration and structuring of dispersed, low-resolution, and partially overlapping information.

The limitations of point estimates are illustrated by widely used plastic chemicals such as dioctyl phthalate, for which reported production and trade volumes have varied substantially over time. These variations reflect shifting regulatory restrictions, regional industrial demand, and broader economic fluctuations, highlighting the need for approaches that emphasize coverage, comparability, and trend identification rather than precise quantification.

Accordingly, the present study contributes an LLM-based data mining framework that systematically consolidates heterogeneous

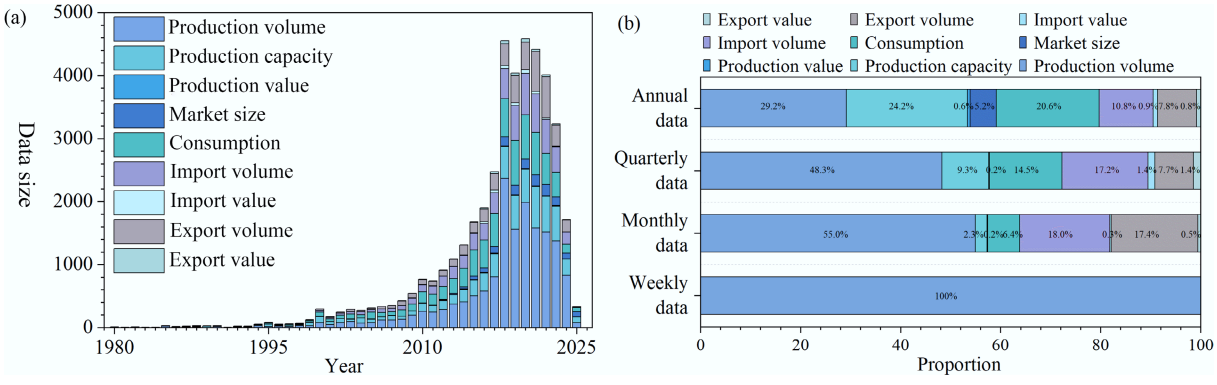


Fig. 3 Temporal coverage of economic indicator data: (a) annual data size for 9 economic indicators from 1980 to 2025 (unit: entry), (b) proportions of the 9 economic indicators across time frames.

and often unstructured economic indicator records into unified, machine-readable datasets. The framework is reproducible, scalable, and highly efficient, enabling levels of coverage and internal consistency that would not be achievable through traditional manual literature reviews. While the resulting datasets do not resolve existing gaps in environmental or health hazard information, they provide a systematic basis for identifying representational imbalances, prioritizing substances for further investigation, and supporting downstream statistical or predictive modeling in data-poor chemical domains.

The number and types of chemicals for which economic information can be extracted are necessarily constrained by the availability of existing reports and records. In general, high-volume or widely used commodity chemicals are better documented, whereas emerging or niche substances remain sparsely characterized. Rather than obscuring this imbalance, the present framework explicitly reveals and quantifies it. For plastic chemicals with remaining data gaps, the structured datasets generated herein provide a foundation for subsequent predictive modeling using machine learning approaches[59]. In addition, integration with emerging data acquisition strategies, including automated sensing[60], digital product passports[61], and robotic sampling technologies[62], may progressively improve data completeness and reduce uncertainty.

Overall, this study should be regarded as a methodological and foundational data contribution rather than a source of definitive or fully authoritative economic indicator data. The primary value of the dataset lies in improving transparency, enabling scalability, and systematically identifying data gaps in domains where perfect data are intrinsically unattainable.

Physicochemical properties, environmental behavior parameters, and PBT/PMT properties

The dataset includes 11 physicochemical properties relevant to plastic chemicals, covering 9,768 substances (Supplementary Table S5). The properties, including boiling point, melting point, vapor pressure, water solubility, and Henry's law constant, are key determinants of chemical fate, controlling volatilization, solubility, degradation, and inter-media partitioning. As essential inputs for multimedia environmental models[42,46,63,64], they underpin the simulation of chemical distribution across the environment at multiple spatial scales.

In addition, 12 environmental behavior parameters were incorporated, covering 9,770 chemicals (Supplementary Tables S4 and S5). Among the parameters, degradation half-lives and ready biodegradability are particularly relevant for regulatory screening, as they directly inform persistence assessments under international frameworks such as the Stockholm Convention on Persistent Organic Pollutants[65]. The integration of these physicochemical and environmental behavior parameters substantially enhances exposure modeling and risk assessments, supporting regulatory intervention.

Using the dataset, 2,144 chemicals were identified as exhibiting PBT properties and 3,951 as PMT substances (Supplementary Table S5), including 4,542 compounds not previously classified within these categories. The omission of these chemicals in earlier evaluations may have led to an underestimation of plastic chemical risks. For example, bis[4-(2-phenyl-2-propyl)phenyl]amine, a widely used plastic antioxidant, exhibits a log bioaccumulation factor of 5.98 in aquatic organisms, exceeding the regulatory threshold of 3.70[66]. Moreover, only 17.0% of its organic carbon was mineralized within 28 days[67] in the OECD-ready biodegradation tests[66], far below the 60% criterion[66], indicating its poor biodegradability and a potential for environmental persistence. This compound has also been associated with skin sensitization[68] and mitochondrial membrane potential disruption[69] (Supplementary Table S4).

Toxicity profiles

The current dataset compiles toxicity information for 4,551 plastic chemicals across 67 endpoints spanning 18 major categories (Supplementary Table S4). The endpoints encompass molecular-level responses (e.g., estrogen receptor α agonism), cellular-level effects (e.g., neurocytotoxicity), and organ-specific outcomes (e.g., hepatotoxicity), thereby enabling characterization of chemical toxicity across biological scales. Such endpoint diversity allows assessment of chemical hazards under different biological and environmental contexts and supports downstream applications in regulatory screening and risk-oriented analyses.

The dataset includes both discrete and continuous toxicity measurements, comprising 48 categorical toxicity endpoints and 19 continuous endpoints (Supplementary Table S4). This structure provides a flexible data foundation for computational toxicology applications, including the development of prediction models using conventional machine learning approaches (e.g., random forests)[70] or deep learning architectures (e.g., graph neural networks)[70]. In addition, the availability of multiple toxicity endpoints facilitates the implementation of multi-task learning frameworks, which can leverage shared toxicological patterns across endpoints to enhance model robustness and generalizability[71,72].

Beyond toxicity prediction, the dataset may also support data-driven chemical design. The integrated toxicity profiles can be incorporated into generative AI approaches[73–75], such as generative adversarial networks[73–75], to guide the exploration of molecular structures with reduced toxicological potential, thereby informing the development of safer chemical alternatives.

Inventory and regulatory scope

The dataset developed in this study mapped the inclusion status of 13,684 plastic chemicals across 111 chemical inventories and regulatory lists (Supplementary Tables S6 and S7). These include 31 existing chemical substance inventories and 80 regulatory lists targeting hazardous or restricted substances (Supplementary Table S7), with broad coverage in regulation, geography, and chemicals (Supplementary Table S6).

The regulatory lists span multiple governance levels, including international (e.g., IARC Monographs on the Identification of Carcinogenic Hazards to Humans)[76], regional (e.g., EU Prior Informed Consent Regulation Annex I Banned Chemicals)[77], and national or subnational regulations[78,79]. The national-level regulations dominate the dataset, accounting for 63 lists across four continents, such as China in Asia (54 lists), France in Europe (5 lists), the United States in North America (3 lists), and New Zealand in Oceania (1 list).

Overall, 9,301 plastic chemicals are included in at least one regulatory list. Regulatory intensity varies widely across jurisdictions, with the "New Zealand Inventory of Chemicals" [80] containing the largest number of hazardous plastic chemicals (7,621), whereas "Indonesia's List of Hazardous and Toxic Substances" [81] includes only 144. Among the regulated substances, PFASs constitute the most heavily regulated category (440 chemicals), followed by organometallic compounds (365 chemicals) and azo dyes (89 chemicals). For instance, perfluoroalkyltriethoxysilane is fully prohibited in Europe and subject to restrictions in several Asian countries, yet remains unregulated in many regions where it is still widely used in packaging, textiles, and architectural materials. This disparity illustrates the lack of international harmonization in plastic chemical regulation, which poses challenges under the increasingly globalized production and trade systems.

In contrast, 7,024 plastic chemicals are not listed on any regulatory inventory. Among these, 1,958 exhibit PBT or PMT properties.

Notably, six unregulated toxic chemicals exceed the HPV threshold of 1.00×10^6 kg·year^{-1}[12]. One example is cis-1,2,3,6-tetrahydrophthalic anhydride, a widely used epoxy resin curing agent and plasticizer with documented respiratory toxicity and corrosive effects[68,82]. Despite an annual production volume of up to 5.1×10^6 kg·year^{-1} in China, this chemical is not included in any national or international regulatory list, representing a significant regulatory blind spot.

At present, only one regulatory framework, the "EU Regulation 10/2011: Authorized Substances for Plastic Food Contact Materials", is specifically designed for plastic-related chemical control. This limited sector-specific coverage underscores the need for a dedicated priority control list for plastic chemicals to support sustainable development of the plastics industry.

Finally, the dataset enables scenario-based modeling of regulatory interventions and their impacts on chemical flows through human and environmental systems. Continued dynamic monitoring of regulatory developments is essential. Strengthened international cooperation and data sharing will be critical for closing regulatory gaps and reducing risks associated with plastic chemical pollution.

Conclusion

This study introduces a multimodal LLM-based framework for large-scale chemical information extraction, enabling accurate retrieval of economic indicators for plastic chemicals from unstructured images and literature. The framework achieves a mean F_1 score of 92.8% (95% confidence interval: 89.9–95.6%) and ensures full reproducibility through open data and code. By integrating manual literature retrieval and computational modeling, the resulting dataset also covers physicochemical properties, toxicity, environmental behavior, and regulatory information, supporting comprehensive chemical risk assessment and sustainable transformation of the plastics industry. Given that more than 350,000 chemicals and mixtures are estimated to be in commercial use worldwide[83], with limited information available for many of them, the workflow proposed here offers a scalable approach to closing the critical data gaps. In the longer term, AI agent-based systems could automate the full pipeline, from data acquisition and extraction to analysis and reporting, substantially reducing manual effort and enabling the construction of comprehensive chemical information infrastructures across industrial sectors.

Despite these advances, several limitations remain. Economic indicators are not uniformly available across chemicals, time periods, or regions, largely due to incomplete statistical reporting, particularly in non-OECD countries and at company-specific levels. Addressing these gaps will require strengthened chemical governance frameworks that mandate transparent disclosure of production capacities and volumes.

The trade-related data are also comparatively limited. A major constraint arises from the coarse resolution of the Harmonized System (HS) codes used in international trade databases, such as UN Comtrade[84], which often rely on 4- or 6-digit classifications that cannot uniquely identify individual chemicals. This lack of specificity hampers the tracing of cross-border chemical flows and the assessment of their contributions to transboundary environmental pollution. Enhancing the granularity and transparency of chemical identifiers in global trade statistics is therefore critical for advancing international chemical monitoring and policy analysis.

The current dataset should be regarded as a baseline rather than a ceiling. Continued progress in non-targeted analytical techniques is expected to substantially expand the recognized chemical space associated with plastics. Moreover, as the dataset is primarily derived from Chinese- and English-language sources, its global representativeness would benefit from the systematic inclusion of multilingual literature and regulatory documents.

Supplementary information

It accompanies this paper at: https://doi.org/10.66178/aie-0026-0011.

Ethical statement

During the preparation of this work, the authors used Qwen2.5-32B and Qwen2-VL-72B for data extraction. 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 research conceptualization, design, implementation, and initial drafting of the manuscript were carried out by Yunhan Cui. Data collection and analysis were performed by Yuxuan Zhang. The development of the machine learning models was conducted by Haobo Wang and Wenjia Liu. Code modification was undertaken by Zijun Xiao and Yuwei Liu. Revisions and critical feedback on the manuscript were provided by Jiale He and Guijia Bi. The manuscript was reviewed by Jingwen Chen, who also contributed to funding acquisition and supervision. All authors approved the final version of the manuscript.

Data availability

All data used or analyzed in this study are publicly available. The compiled dataset supporting the results is provided in Supplementary Information S1.

Funding

This study was supported by the National Natural Science Foundation of China (22136001) and the National Key R&D Program of China (2022YFC3902100).

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author details

Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

References

[1] Global plastics outlook policy scenarios to 2060. 2022. OECD: Paris, France https://www.oecd.org/en/publications/global-plastics-outlook_aa1edf33-en.html

[2] UNEA Resolution 5/14 entitled "End plastic pollution: towards an international legally binding instrument". 2022. United Nations Environment Assembly of the United Nations Environment Programme: Nairobi https://wedocs.unep.org/bitstream/handle/20.500.11822/39812/OEWG_PP_1_INF_1_UNEA%20resolution.pdf
[3] Wang ZY, Praetorius A. 2022. Integrating a chemicals perspective into the global plastic treaty. *Environmental Science & Technology Letters* 9(12):1000–1006
[4] Wiesinger H, Wang ZY, Hellweg S. 2021. Deep dive into plastic monomers, additives, and processing aids. *Environmental Science & Technology* 55(13):9339–9351
[5] Monclús L, Arp HPH, Groh KJ, Faltynkova A, Løseth ME, et al. 2025. Mapping the chemical complexity of plastics. *Nature* 643(8071):349–355
[6] State of the science on plastic chemicals - Identifying and addressing chemicals and polymers of concern. 2024. Norwegian University of Science and Technology: Trondheim https://plastchem-project.org
[7] Gao S, Huang G, Han D. 2025. Plastic leachates disproportionately impair aquatic animals: a multifactor, multieffect, and multilevel meta-analytic model. *Environmental Science & Technology* 59(22):10854–10866
[8] Plastics, EDCs & health: a guide for public interest organizations and policy-makers on endocrine disrupting chemicals & plastics. 2020. Endocrine Society & IPEN: Sweden https://www.endocrine.org/topics/edc/plastics-edcs-and-health
[9] Global plastics outlook economic drivers, environmental impacts and policy options. 2022. OECD: Paris, France https://doi.org/10.1787/de747aef-en
[10] Global Chemicals Outlook II: From legacies to innovative solutions: implementing the 2030 agenda for sustainable development-synthesis report. 2019. United Nations Environment Programme: Geneva, Switzerland https://www.unep.org/resources/report/global-chemicals-outlook-ii-legacies-innovative-solutions
[11] Chemicals in Plastic: A technical report. 2023. United Nations Environment Programme and Secretariat of the Basel, Rotterdam and Stockholm Conventions: Geneva, Switzerland https://www.unep.org/resources/report/chemicals-plastics-technical-report
[12] OECD Existing Chemicals Database. https://hpvchemicals.oecd.org/UI/Search.aspx (accessed Jun 14, 2025)
[13] Wang X, Huang LL, Xu SZ, Lu K. 2024. How does a generative large language model perform on domain-specific information extraction? - A comparison between gpt-4 and a rule-based method on band gap extraction. *Journal of Chemical Information and Modeling* 64(20):7895–7904
[14] Chyxx. com. http://www.chyxx.com/industry/201806/653004.html (accessed Sep 10, 2018)
[15] GPT-4. https://openai.com/index/gpt-4/ (accessed Jun 14, 2025)
[16] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, et al. 2024. Structured information extraction from scientific text with large language models. *Nature Communications* 15(1):1418
[17] Yuan WH, Chen GY, Wang ZL, You FQ. 2025. Empowering Generalist Material Intelligence with Large Language Models. *Advanced Materials (Weinheim)* 37(32):e2502771
[18] Tang YY, Xiao ZY, Li X, Fang QW, Zhang QP, et al. 2025. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: A comparative study of GPT-3.5 and GPT-4. *Medrxiv* http://doi.org/10.1101/2024.03.20.24304572
[19] Liang WQ, Su WY, Zhong LJ, Yang ZD, Li TY, et al. 2024. Comprehensive characterization of oxidative stress-modulating chemicals using GPT-based text mining. *Environmental Science & Technology* 58(46):20540–20552
[20] Qwen. https://www.aliyun.com/product/tongyi (accessed Jun 14, 2025)
[21] United Nations Environment Programme. https://www.unep.org/data-resources (accessed Jan 12, 2025)
[22] AskCI. https://www.askci.com/ (accessed Jul 17, 2024)
[23] Fxbaogao. https://www.fxbaogao.com/ (accessed May 14, 2024)
[24] Qianzhan. https://bg.qianzhan.com/ (accessed Jun 15, 2025)
[25] Cqvip. https://qikan.cqvip.com/ (accessed Sep 19, 2025)
[26] Web of Science. https://webofscience.clarivate.cn/wos/alldb/basic-search (accessed Jun 15, 2025)
[27] PubMed Central. https://pmc.ncbi.nlm.nih.gov/ (accessed Dec 21, 2025)
[28] Wang P, Bai S, Tan S, Wang SJ, Fan ZH, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *Arxiv Preprint* https://doi.org/10.48550/arXiv.2409.12191
[29] Qwen2.5. https://news.qq.com/rain/a/20240919A036GQ00 (accessed Sep 29, 2024)
[30] PyMuPDF. https://pypi.org/project/PyMuPDF/ (accessed Jul 5, 2025)
[31] GuideChem. https://china.guidechem.com/datacenter/namestocas.html (accessed Dec 1, 2023)
[32] Ichemistry. http://www.ichemistry.cn/ (accessed June 6, 2025)
[33] ChemRadar. https://www.chemradar.com/ (accessed Sep 10, 2024)
[34] Schilling-Wilhelmi M, Ríos-García M, Shabih S, Gil MV, Miret S, et al. 2025. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews* 54(3):1125–1150
[35] Yu YN, Yu Y, Wei K, Luo H, Wang HH. 2025. SIPDO: Closed-loop prompt optimization via synthetic data feedback. *Arxiv Preprint*
[36] Choi J. 2025. Efficient prompt optimization for relevance evaluation via LLM-based confusion matrix feedback. *Applied Sciences-Basel* 15(9):5198
[37] Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, et al. 2023. Self-Refine: Iterative refinement with self-feedback, *37th Conference on Neural Information Processing Systems*. Curran Associates, Inc. pp. 46534–46594 http://doi.org/10.48550/arXiv.2303.17651
[38] Li J, Klinger R. 2025. iPrOp: Interactive prompt optimization for large language models with a human in the loop, *The 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. pp. 276–285 http://doi.org/10.48550/arxiv.2412.12644
[39] Karn RR. 2025. Linear feedback control systems for iterative prompt optimization in large language models. *Arxiv Preprint* http://doi.org/10.48550/arxiv.2501.11979
[40] Barker M, Chue Hong NP, Katz DS, Lamprecht A, Martinez-Ortiz C, et al. 2022. Introducing the FAIR Principles for research software. *Scientific Data* 9(1):622
[41] Candidate List of substances of very high concern for Authorisation. https://echa.europa.eu/candidate-list-table (accessed May 6, 2025)
[42] Wang JY, Chan FKS, Johnson MF, Chan HK, Cui YH, et al. 2025. Material cycles, environmental emissions, and ecological risks of bisphenol a (BPA) in China and implications for sustainable plastic management. *Environmental Science & Technology* 59(3):1631–1646
[43] Wang JY, Chan FKS, Johnson MF, Chan HK, Cui YH, et al. 2024. Material flow analysis of chemical additives in plastics: A critical review. *Critical Reviews in Environmental Science and Technology* 54(23):1692–1708
[44] Graedel TE. 2019. Material flow analysis from origin to evolution. *Environmental Science & Technology* 53(21):12188–12196
[45] Wang CY, Liu Y, Chen WQ, Zhu B, Qu S, et al. 2021. Critical review of global plastics stock and flow data. *Journal of Industrial Ecology* 25(5):1300–1317
[46] Cui YH, Chen JW, Wang ZY, Wang JY, Allen DT. 2022. Dynamic material flow, multimedia environmental model, and ecological risk analysis for chemical management: A di(2-ethylhexhyl) phthalate case in China. *Environmental Science & Technology* 56(15):11006–11016
[47] Cui YH, Wang JY, Song GB, Chen JW. 2025. Mapping flows, stocks, plastic emissions, and greenhouse gas emissions of polyurethanes: Decoding challenges and pollution prevention pathways in China. *Environmental Science & Technology* 59(20):9980–9990
[48] Bi MY, Liu W, Luan XY, Li MY, Liu M, et al. 2021. Production, use, and fate of phthalic acid esters for polyvinyl chloride products in China. *Environmental Science & Technology* 55(20):13980–13989
[49] Heller MC, Mazor MH, Keoleian GA. 2020. Plastics in the US: toward a material flow characterization of production, markets and end of life. *Environmental Research Letters* 15(9):94034
[50] Ciacci L, Passarini F, Vassura I. 2017. The European PVC cycle: In-use stock and flows. *Resources Conservation and Recycling* 123:108–116

[51] Kawecki D, Scheeder PRW, Nowack B. 2018. Probabilistic material flow analysis of seven commodity plastics in Europe. *Environmental Science & Technology* 52(17):9874–9888
[52] Eriksen MK, Pivnenko K, Faraca G, Boldrin A, Astrup TF. 2020. Dynamic material flow analysis of PET, PE, and PP flows in Europe: Evaluation of the potential for circular economy. *Environmental Science & Technology* 54(24):16166–16175
[53] Chen CK, Chen A, Zhan FQ, Wania F, Zhang SX, et al. 2022. Global historical production, use, in-use stocks, and emissions of short-, medium-, and long-chain chlorinated paraffins. *Environmental Science & Technology* 56(12):7895–7904
[54] Houssini K, Li JH, Tan QY. 2025. Complexities of the global plastics supply chain revealed in a trade-linked material flow analysis. *Communications Earth & Environment* 6(1):257
[55] Facts & Figures of the European Chemical Industry. https://cefic.org/facts-and-figures-of-the-european-chemical-industry/ (accessed July 30, 2024)
[56] Cathay Biotech Inc. https://www.cathaybiotech.com/projxags.aspx#content (accessed May 6, 2025)
[57] US EPA CompTox Chemicals Dashboard. https://comptox.epa.gov/dashboard/ (accessed Aug 20, 2024)
[58] Assessment of regulatory needs list. https://echa.europa.eu/assessment-regulatory-needs (accessed Aug 20, 2024)
[59] Cottom JW, Cook E, Velis CA. 2024. A local-to-global emissions inventory of macroplastic pollution. *Nature* 633(8028):101–108
[60] Pracucci A, Giovanardi M. 2025. Design of a Sensor-Based Digital Product Passport for Low-Tech Manufacturing: Traceability and Environmental Monitoring in Bio-Block Production. *Sensors* 25(18):5653
[61] Koppelaar RHEM, Pamidi S, Hajósi E, Herreras L, Leroy P, et al. 2023. A Digital Product Passport for Critical Raw Materials Reuse and Recycling. *Sustainability* 15(2):1405
[62] Li W, Liu X, Hu B. 2026. Robotic technologies for environmental mass spectrometry: Recent advances and perspectives. *TrAC Trends in Analytical Chemistry* 195:118614
[63] Chen CK, Li L, Liu JZ, Liu JG. 2019. Global environmental fate of short-chain chlorinated paraffins: Modeling with a single vs. multiple sets of physicochemical properties. *Science of the Total Environment* 666:423–430
[64] Chen ZH, Ma TY, Liu W, Yuan G, Pan X, et al. 2024. Brominated flame retardants (BFRs) in China over the past half-century: Stocks, flows, fates, and ecological risks. *Environmental Science & Technology* 58(31):13613–13623
[65] Stockholm Convention on Persistent Organic Pollutants. https://pops.int/ (accessed Jun 15, 2025)
[66] Test No. 301: Ready Biodegradability. https://www.oecd.org/en/publications/test-no-301-ready-biodegradability_9789264070349-en.html (accessed Aug 13, 2024)
[67] Huang K, Zhang HC. 2022. Classification and regression machine learning models for predicting aerobic ready and inherent biodegradation of organic chemicals in water. *Environmental Science & Technology* 56(17):12755–12764
[68] Borba JVB, Alves VM, Braga RC, Korn DR, Overdahl K, et al. 2022. STopTox: An *in Silico* alternative to animal testing for acute systemic and topical toxicity. *Environmental Health Perspectives* 130(2):027012
[69] Tang W, Chen J, Hong H. 2020. Discriminant models on mitochondrial toxicity improved by consensus modeling and resolving imbalance in training. *Chemosphere* 253:126768
[70] Jeong J, Choi J. 2022. Artificial intelligence-based toxicity prediction of environmental chemicals: Future directions for chemical management applications. *Environmental Science & Technology* 56(12):7532–7543
[71] Wu ZX, Jiang DJ, Wang JK, Hsieh CY, Cao DS, et al. 2021. Mining toxicity information from large amounts of toxicity data. *Journal of Medicinal Chemistry* 64(10):6924–6936
[72] Yang X, Sun JQ, Jin BY, Lu YE, Cheng JY, et al. 2025. Multi-task aquatic toxicity prediction model based on multi-level features fusion. *Journal of Advanced Research* 68:477–489
[73] Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. 2020. Generative molecular design in low data regimes. *Nature Machine Intelligence* 2(3):171–180
[74] Du YQ, Jamasb AR, Guo J, Fu T, Harris C, et al. 2024. Machine learning-aided generative molecular design. *Nature Machine Intelligence* 6(6):589–604
[75] Anstine DM, Isayev O. 2023. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society* 145(16):8736–8750
[76] IARC Monographs on the Identification of Carcinogenic Hazards to Humans. https://publications.iarc.who.int/Book-And-Report-Series/Iarc-Monographs-On-The-Identification-Of-Carcinogenic-Hazards-To-Humans (accessed Dec 1, 2023)
[77] EU PIC Regulation Annex I (Prior Informed Consent - Banned Chemicals). https://echa.europa.eu/regulations/prior-informed-consent/list-chemicals (accessed Jun 1, 2025)
[78] Singapore Ozone Depleting Substances Controlled List. https://sso.agc.gov.sg/SL/EPMA1999-RG9 (accessed Aug 2, 2025)
[79] Measures for Prohibiting, Restricting and Controlling Hazardous Chemicals in Beijing Municipality. https://yjglj.beijing.gov.cn/art/2024/1/24/art_10906_3802.html (accessed Jan 24, 2024)
[80] New Zealand Inventory of Chemicals. https://www.epa.govt.nz/database-search/new-zealand-inventory-of-chemicals-nzioc/ (accessed Jun 1, 2025)
[81] Indonesia List of Limited Use Hazardous and Toxic Substances. https://peraturan.bpk.go.id/Details/53080/pp-no-74-tahun-2001 (accessed Jun 1, 2025)
[82] Wang Z, Zhao P, Zhang X, Xu X, Li W, et al. 2021. In silico prediction of chemical respiratory toxicity via machine learning. *Computational Toxicology* 18:100155
[83] Wang ZY, Walker GW, Muir DCG, Nagatani-Yoshida K. 2020. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories. *Environmental Science & Technology* 54(5):2575–2584
[84] UN Comtrade. http://comtrade.un.org/ (accessed Oct 11, 2021)

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.