

Review

Open Access

Application of machine learning in non-targeted analysis for environmental organic pollutants

Yuwei Liu[#], Haoyang Xiong[#], Jinhua Liu, Huaijun Xie^{*}  and Jingwen Chen^{*} 

Citation: Liu, Y.-W.; Xiong, H.-Y.; Liu, J.-H.; et al. Application of machine learning in non-targeted analysis for environmental organic pollutants. *AI Environ.* 2026, 1(1): 11-22.

Received: 11 December 2025

Revised: 6 January 2026

Accepted: 30 January 2026

Published online: 11 February 2026

Abstract

The vast diversity of organic pollutants in the environment, coupled with the scarcity of authentic standards and reference spectra, poses major challenges to the reliable identification and quantification of these compounds. Machine learning (ML), a powerful tool for data processing, has garnered growing attention for its value in non-targeted analysis. This review summarizes recent advances in the application of ML to pollutant identification and quantification. In qualitative analysis, ML accelerates key steps in structure elucidation, including predicting MS/MS spectra to expand in silico libraries, inferring molecular formulas and substructures from spectral data, generating novel candidate structures beyond existing databases, and predicting retention time and collision cross section values to enhance identification confidence. In quantitative analysis, ML facilitates standard-free methodologies through optimized surrogate compound selection and prediction of ionization efficiency and response factors. Despite these achievements, salient challenges remain, including model generalizability across instrumental platforms, improved interpretability, and the establishment of unified frameworks for uncertainty quantification and reproducibility. Future advances toward more accurate, integrated, automated, and transferable ML systems are expected to enable scalable and intelligent analysis of environmental pollutants.

Keywords: Non-targeted analysis, High-resolution mass spectrometry, Machine learning, Structure generation, Quantification method

Highlights

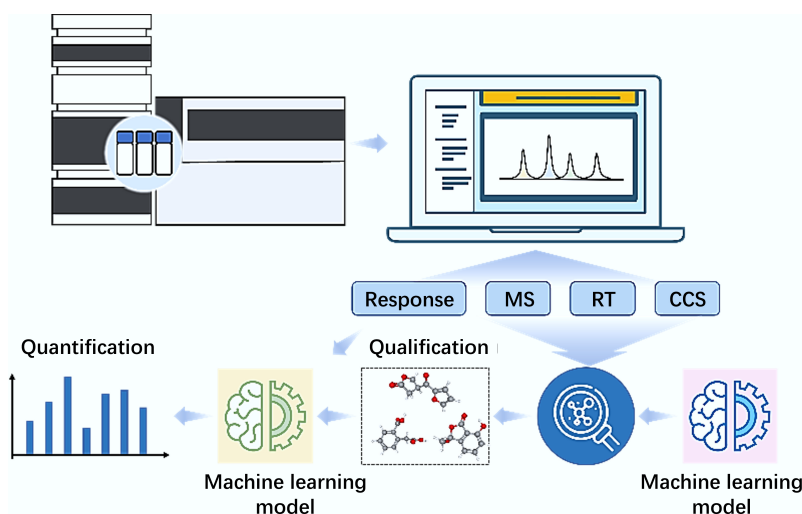
- Reviews recent advances of machine learning (ML) methods in qualitative and quantitative non-targeted analysis of environmental pollutants.
- Summarizes ML-assisted strategies for structure elucidation, including spectra prediction, feature inference, and de novo molecular generation.
- Highlights ML-enabled semi-quantification via ionization efficiency and response factor prediction.
- Discusses key challenges and future directions toward transferable, interpretable, and automated ML-driven screening frameworks.

Authors contributed equally: Yuwei Liu and Haoyang Xiong

^{*} Correspondence: Huaijun Xie (huaijunxie@mail.dlut.edu.cn); Jingwen Chen (jwchen@dlut.edu.cn)

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

Graphical abstract



Introduction

Organic pollutants in the environment exhibit remarkable diversity and pose substantial threats to ecosystems and human health^[1,2]. The continual expansion of industrial and agricultural activities has led to the persistent release of numerous emerging organic pollutants, such as pharmaceuticals and personal care products, pesticides, industrial additives, and their transformation products, into the environment^[3–5]. Accurate identification and quantification of these pollutants are critical prerequisites for reliable exposure assessment and ecological risk evaluation^[6–8]. However, most organic pollutants lack commercially available reference standards, making their structural elucidation and concentration determination central challenges in environmental analytical chemistry^[7,8].

In targeted analysis, reference standards are employed to establish the retention times (RT) and characteristic fragment ions of target pollutants, allowing their specific identification in environmental samples^[9,10]. Nevertheless, the analytical throughput of targeted analysis is inherently constrained and cannot capture compounds outside the scope of available reference standards^[11]. In contrast, non-targeted analysis (NTA) based on liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) enables comprehensive detection of ionizable compounds within environmental samples^[12,13]. By acquiring RT, exact mass, isotopic patterns, fragmentation spectra, and—when ion mobility spectrometry is employed—collision cross section (CCS) values, and subsequently matching these features to spectral libraries, NTA allows the structural annotation of compounds without reference standards^[14–16]. This approach has considerably broadened the analytical scope of environmental analysis and facilitated the discovery of a wide range of pollutants, including endocrine-disrupting chemicals and other emerging pollutants^[17,18].

Despite its extensive data acquisition capability, NTA faces a substantial data interpretation bottleneck. A single analytical run can yield tens of thousands of spectral features, yet only a small fraction can be confidently annotated using existing libraries^[13,19]. The vast majority of detected signals remain unidentified. This limitation primarily arises from the fact that experimental spectral libraries cover only a minute fraction of the relevant chemical space^[20]. Current estimates suggest that less than 3% of environmentally relevant compounds can be identified through conventional NTA workflows^[13]. Furthermore, even for tentatively annotated compounds,

the absence of reference standards hampers accurate quantification, representing another key obstacle to exposure and risk assessment^[21].

Machine learning (ML) offers a transformative pathway to address these analytical challenges^[22,23]. As a data-driven approach capable of extracting complex, nonlinear relationships from high-dimensional data, ML enables the transition from expert-dependent manual interpretation toward automated and high-throughput analysis^[24,25]. Within the NTA workflow, ML has been increasingly applied in multiple analytical stages, from qualitative to quantitative analysis (Fig. 1). In qualitative applications, ML models trained on structure-property relationships can predict tandem mass (MS/MS) spectra, RT, CCS, and other properties to enrich existing libraries^[22,26]. Alternatively, inverse modeling strategies allow the inference or generation of the most probable molecular structures directly from experimental spectra, thereby reducing dependence on limited experimental libraries^[14,22,27,28]. In quantitative contexts, ML-based quantitative structure-property relationship (QSPR) models have been developed to estimate ionization efficiencies (IE) or mass spectrometric response factors (RF), facilitating large-scale standard-free quantification of organic pollutants^[29–31].

Recent advances in ML have increasingly influenced both qualitative and quantitative NTA by addressing key analytical bottlenecks in LC-HRMS-based workflows (Fig. 1). Building on these developments, this review provides a comprehensive and integrative overview of ML-assisted NTA for environmental organic pollutants. Recent methodological advances across key stages of the NTA workflow are systematically summarized, with particular emphasis on their practical limitations. In addition, emerging trends and future research directions are highlighted to guide the development and application of ML-assisted NTA.

Machine learning-assisted identification in non-targeted analysis

ML has emerged as a powerful tool for the qualitative analysis of organic pollutants, and its applications can be broadly categorized into four main domains. The first domain involves training ML models to predict MS/MS spectra from molecular structures in reference databases, thereby providing auxiliary spectral information to support compound annotation^[22,32]. The second domain focuses on inferring

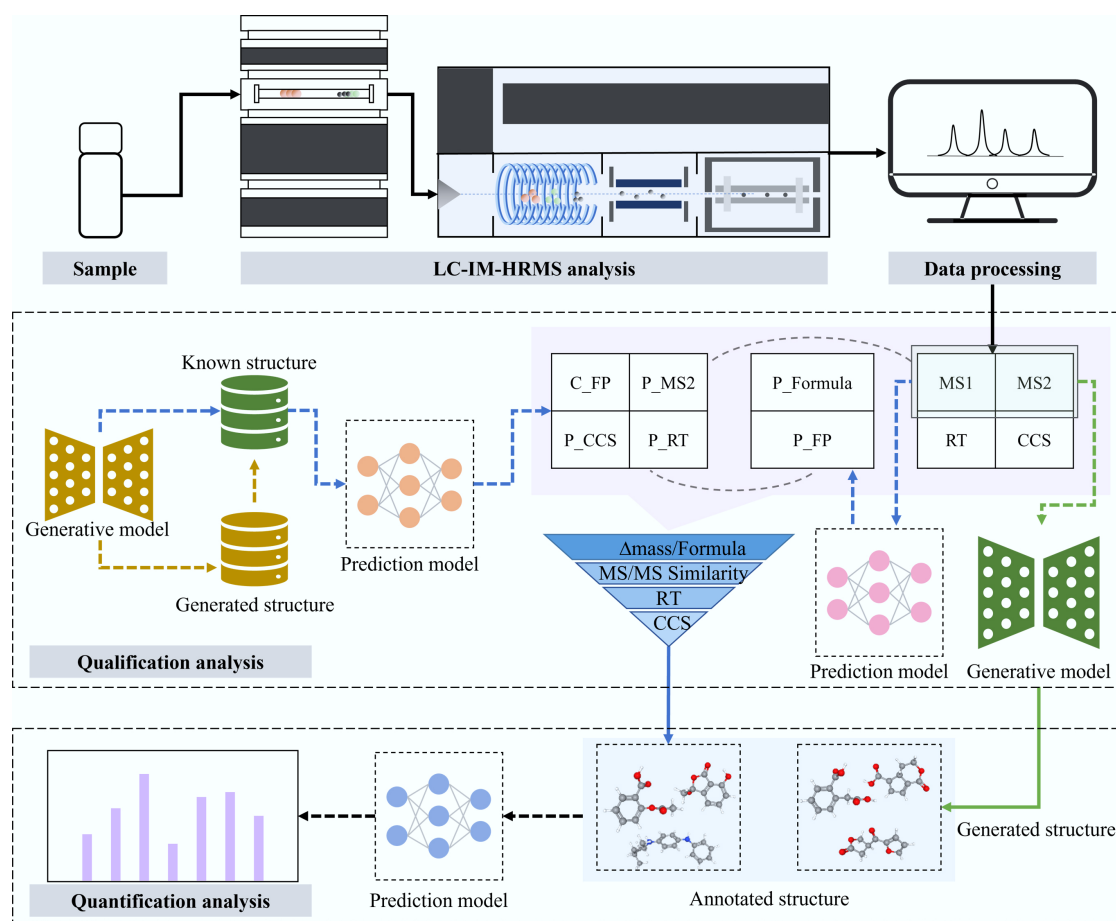


Fig. 1 Workflow of non-targeted analysis based on machine learning. (Variables prefixed with C_ denote calculated values, whereas those prefixed with P_ represent predicted values)

structural attributes directly from mass spectral data, such as molecular formulas, substructures, and molecular fingerprints, which can then be matched against chemical structure libraries for identification^[33–35]. The third domain comprises the prediction of chromatographic and ion-mobility-related properties, including RT and CCS^[36], from molecular structures, providing orthogonal information to prioritize candidate structures and improve confidence in qualitative identification. These three domains are primarily applied to compounds whose structures are already represented in existing databases. The fourth domain extends beyond database-dependent identification by employing ML models to directly generate candidate structures from spectral inputs or to expand chemical space based on known pollutants, thereby facilitating the discovery of truly unknown pollutants outside current databases^[37,38].

MS/MS spectra prediction

Conventional spectra matching is constrained by the incomplete coverage of experimentally acquired spectra. MS/MS prediction aims to computationally generate MS/MS spectra directly from known molecular structures. This capability is essential in LC-HRMS-based NTA, serving as a bridge between chemical structure and spectra to enable structure elucidation and in silico spectral-library expansion (Fig. 2)^[24,39].

Traditional in silico approaches rely on expert-defined bond-cleavage rules coupled with heuristic scoring strategies^[28]. These methods enumerate plausible fragmentation trees and assign intensities based on chemistry-driven heuristics supported by statistical

learning. Representative tools include MetFrag, MAGMa, and CFM-ID^[28,40,41]. Although rule-based models provide mechanistic interpretability, they require substantial expert curation, entail high computational cost, and tend to over-predict candidate fragments, ultimately compromising accuracy.

To overcome the trade-off between mechanistic interpretability and computational efficiency inherent in purely rule-based fragmentation models, hybrid strategies that integrate expert-defined fragmentation knowledge with machine learning have emerged. These approaches aim to reduce exhaustive enumeration of fragmentation pathways while retaining chemically meaningful fragmentation constraints.

CFM-ID 4.0 exemplifies this paradigm by combining curated fragmentation rules with probabilistic modeling of molecular topological features, resulting in substantially improved spectral prediction performance compared with earlier versions^[22,42]. Nevertheless, its reliance on explicit enumeration of candidate fragmentation trees remains computationally demanding, particularly for large or structurally complex molecules^[43].

To further alleviate this bottleneck, subsequent methods have progressively shifted fragmentation modeling from explicit rule enumeration toward learned fragment representations. ICEBERG adopts a two-stage neural-network architecture that separately predicts fragment ions and their intensities, thereby reducing the need for exhaustive pathway enumeration and improving computational efficiency^[44]. Extending this idea, FIORA leverages graph neural networks (GNNs) to directly learn local fragmentation

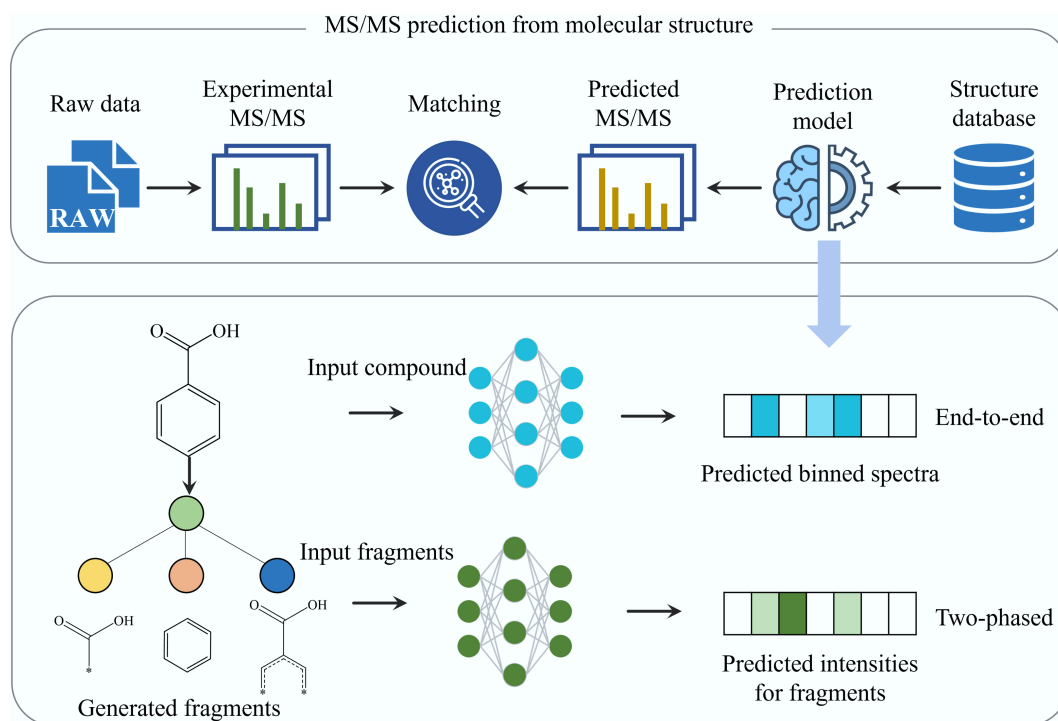


Fig. 2 Compound annotation based on MS/MS spectra predicted from molecular structures.

patterns and associated intensities from molecular neighborhoods, achieving superior performance relative to both CFM-ID and ICEBERG while simultaneously enabling joint prediction of orthogonal properties such as RT and CCS^[45].

Beyond two-stage designs, recent end-to-end models dispense with explicit fragmentation modeling altogether and instead learn a direct mapping from molecular structure to MS/MS spectra. 3DMolMS predicts spectra directly from three-dimensional molecular conformations, capturing stereochemical and spatial effects on fragmentation^[46]. Massformer further advances this paradigm by employing a graph-transformer architecture that integrates collision energy, ionization mode, and instrument parameters, enabling accurate spectrum prediction at 1 Da resolution across large benchmark datasets^[47]. Complementary GNN-based frameworks incorporate chemically informed inductive biases, such as recurrent structural motifs^[48] or quantum-chemical bond dissociation energies as edge features^[49], allowing fine-grained fragmentation patterns to be captured and enabling high-resolution (0.01 Da) MS/MS prediction.

Structure feature prediction from MS/MS spectra

ML models for inferring molecular information from MS/MS spectra have emerged as powerful tools for structural annotation (Fig. 3). Accurate molecular formula determination is a critical step toward elucidating the structures of unknown pollutants. Among the most widely applied tools, SIRIUS generates candidate formulas from precursor m/z via exhaustive enumeration^[27]. For each candidate, it constructs fragmentation trees to explain observed fragment ions, and subsequently evaluates tree plausibility using a Bayesian scoring framework. While effective, such approaches can be computationally intensive, particularly for large-scale screening tasks.

To reduce computational complexity and improve generalization beyond previously observed molecular formulas, recent ML approaches have shifted from explicit fragmentation-tree construction toward end-to-end spectral representations. MIST-CF exemplifies

this transition by directly mapping MS/MS spectra to candidate molecular formula scores^[50]. Rather than enumerating fragmentation trees, the model annotates fragment ions with plausible subformulas and extracts chemically meaningful relationships such as neutral losses (Fig. 3). These local spectral features, together with precursor information, are encoded using a Transformer architecture, and an energy-based model subsequently ranks candidate formulas. By bypassing explicit tree construction, MIST-CF substantially reduces computational burden while enhancing its ability to recognize previously unseen molecular formulas.

Despite these advances, molecular formula prediction intrinsically limits candidate retrieval to sets of isomeric structures and therefore cannot achieve unambiguous compound identification. To further increase structural specificity, subsequent ML methods have progressively moved toward predicting higher-level structural descriptors directly from MS/MS fragmentation patterns. In this context, molecular fingerprint prediction has emerged as a key intermediate representation that bridges spectral information and full molecular structures (Fig. 3).

CSI:FingerID^[51] represents an early and influential implementation of this paradigm. It combines fragmentation-tree analysis with support vector machine (SVM) classifiers to predict substructure-associated molecular fingerprints using kernel-based embeddings. The predicted fingerprints are then matched against reference structure databases, enabling systematic linkage between MS/MS spectra and known compounds. Building on this framework, Wang et al. developed per- and polyfluoroalkyl substances (PFAS)-specific SVM models that predict 254-bit molecular fingerprints directly from MS/MS spectra, tailoring the representation to a chemically constrained class and thereby enabling high-throughput annotation of PFAS in NTA workflows^[35].

With advances in ML, end-to-end Transformer architectures have further improved fingerprint prediction. MIST treats spectra as peak sequences or matrices and leverages self-attention to capture global

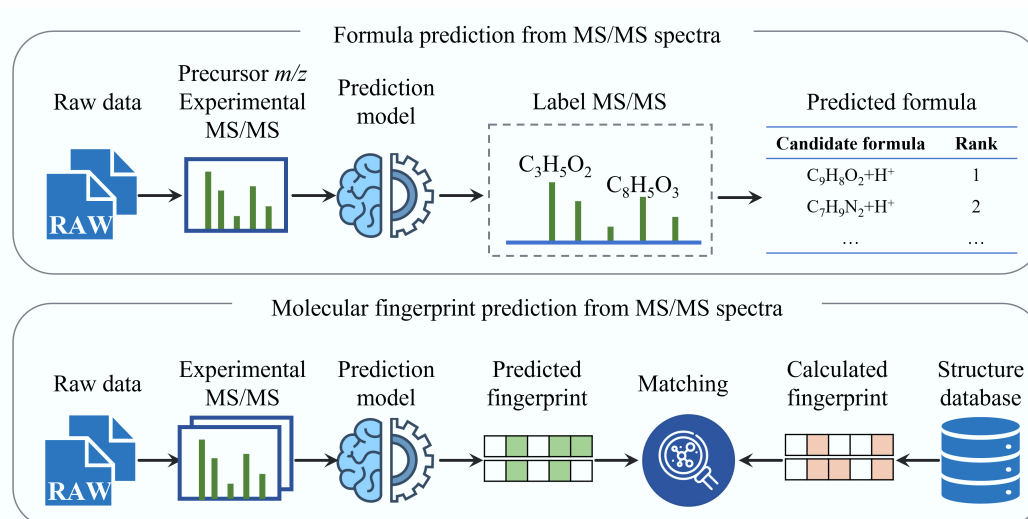


Fig. 3 Compound annotation based on structural features predicted from MS/MS spectra.

dependencies among fragment ions, enabling direct mapping from spectra to molecular fingerprints or substructure vectors without individual feature engineering^[52]. Extending this paradigm, IDSL_MINT introduces a flexible Transformer-based framework that tokenizes peaks and precursor m/z values and applies intensity-weighted positional encoding. This end-to-end model supports user-defined spectral libraries and fingerprint types, allowing rapid adaptation to specific chemical domains (such as lipidomics or environmental contaminants), thus providing a scalable strategy for high-confidence annotation across specific chemical classes^[53].

Structure identification based on *de novo* molecular generation

ML-based methods can either predict MS/MS spectra from known structures or infer structural information from MS/MS spectra. However, both approaches ultimately rely on matching experimental spectra with compounds recorded in structural databases. For pollutants not covered by these databases, particularly emerging pollutants such as transformation products or byproducts, these approaches often fail to provide effective structural annotations^[24,54]. The need to overcome the limitations of databases and to cover a broader chemical space of pollutants is increasingly pressing.

Molecular generation, as a subfield of ML offers a promising avenue for this challenge, enabling the generation of novel molecules by sampling from latent space or reconstructing molecules from latent vectors, thereby facilitating the exploration of the chemical space of unknown pollutants^[55–57]. Representative generative architectures include variational autoencoders, generative adversarial networks, autoregressive models, normalizing flows, and diffusion models, each offering unique mechanisms for sampling chemical space^[57,58].

Molecular generative frameworks have gained increasing traction in the structure elucidation of unknown compounds. These approaches can be classified into two conceptual paradigms. The first seeks to reconstruct molecular structures from MS/MS spectra, thereby enabling *de novo* structure generation. The second focuses on expanding the accessible chemical space by generating novel theoretical compounds, thus breaking through the coverage limitations of existing structure databases and facilitating the annotation of unknown pollutants (Fig. 4).

As generative models rely on extensive and high-quality training data, the limited availability of high-resolution MS/MS spectra-structure paired datasets presents a significant barrier to direct model development^[37,38]. To address this constraint, current research has predominantly explored two complementary strategies: one employs data-augmented end-to-end frameworks that directly generate molecular structures from MS/MS data^[22,59,60], while the other adopts an indirect latent-space paradigm, where spectra are first mapped to latent representations and subsequently decoded into molecular structures, thereby reducing reliance on extensive paired datasets^[37,61,62].

Among these strategies, direct methods leverage MS/MS prediction tools to generate computational MS/MS spectra for known compounds, thereby augmenting MS-structure pairs and enabling end-to-end training. In these models, MS information is mapped to an implicit latent space and subsequently decoded directly into Simplified Molecular Input Line Entry System (SMILES) sequences. Representative examples include MassGenie^[59] and MS2Mol^[60]. MassGenie employs the in-house FragGenie tool to computationally fragment millions of known compounds, producing approximately 21 million MS-structure pairs, which are then used to train a model that translates peak lists directly into SMILES^[59]. On the CASMI 2017 challenge dataset, MassGenie achieved a top-1 accuracy of 53%, demonstrating the feasibility of end-to-end generation^[59]. MS2Mol, on the other hand, uses CFM-ID 4.0 to predict spectra for natural products to augment training data and employs a BART-based Transformer architecture for end-to-end learning^[22,60,63]. Evaluated on the EnvedaDark dataset for unknown compound identification, MS2Mol successfully generated similar structures for 62% of unknown natural products, outperforming traditional database search approaches^[60].

In contrast to end-to-end training, indirect methods avoid the need to simultaneously optimize spectral feature extraction and structure reconstruction. These methods learn spectral embeddings and structural generation separately. This decoupled strategy lowers the difficulty of model training and reduces MS/MS spectra-structure data requirements. Notable examples include MSNovelist^[37], Mass2SMILES^[61], and Spec2Mol^[62]. MSNovelist integrates SIRIUS and CSI:FingerID to predict molecular formulas and high-dimensional fingerprints from MS/MS spectra, encodes the fingerprints together with the formula into latent vectors, and autoregressively generates

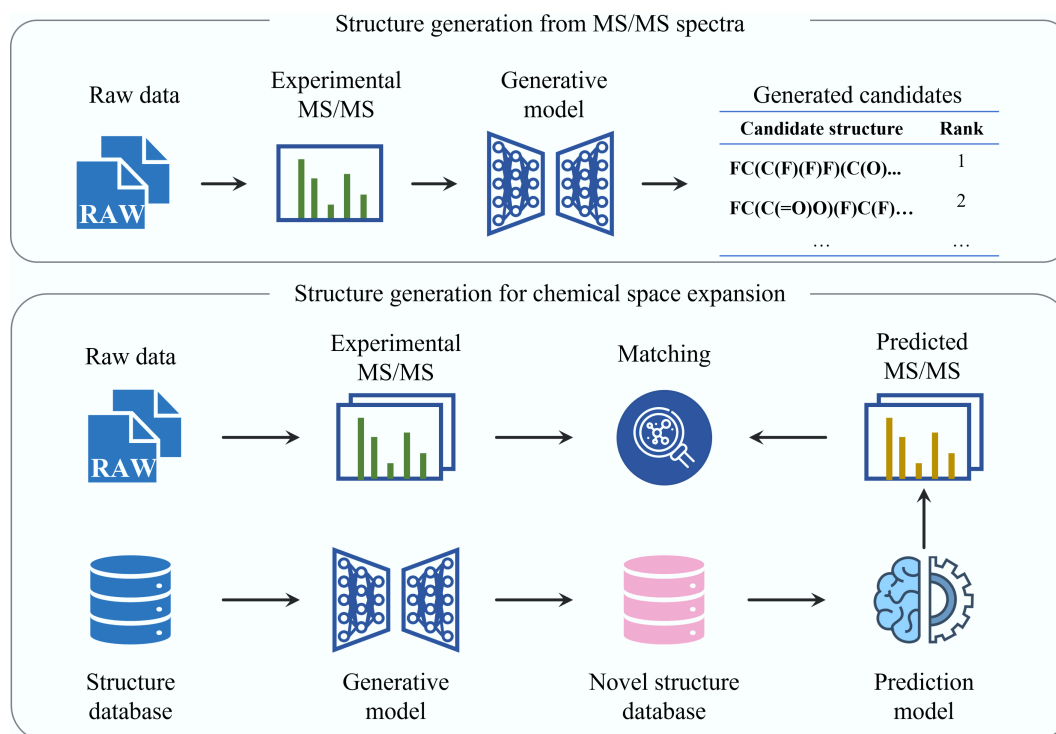


Fig. 4 Compound annotation based on generative models for structure generation or chemical space expansion.

SMILES using a pretrained recurrent neural network decoder^[27,51]. On the Global Natural Products Social Molecular Networking (GNPS) dataset^[64], it successfully recovered the true structure for approximately 45% of molecules among 128 candidates, with a top-1 hit rate of 61%. Mass2SMILES further maps MS/MS spectra to continuous molecular descriptors (CDDD embeddings), reframing structure generation as a combined regression problem^[61,65]. Subsequently, the CDDD embeddings can be decoded to SMILES using the pretrained CDDD decoder^[65]. Spec2Mol adopts a different latent-space strategy, using a convolutional neural network (CNN) encoder to map spectra into vectors constrained within a latent space pretrained on molecular structures using a gated recurrent unit (GRU) autoencoder, from which molecular structures are reconstructed via the GRU decoder^[62].

In addition to structure inference, molecular generation provides a promising means to explore unknown chemical space and generate potential structures of pollutants to expand chemical space coverage^[55,56]. A notable example is DarkNPS, developed to accelerate the identification of novel psychoactive substances (NPS)^[38]. Trained on approximately 1,700 known NPS, DarkNPS generates large ensembles of NPS-like candidate structures conditioned on exact mass. The frequency with which a given structure is sampled is interpreted as a structural prior, reflecting its plausibility within the learned NPS chemical space. By combining exact mass constraints with this sampling-based structural prior, DarkNPS prioritizes candidate structures and enables structure annotation without reliance on reference databases. Using exact mass alone, correct structures were ranked first in approximately 35% of cases, increasing to 51% when MS/MS information was further incorporated^[38].

Extending this concept beyond NPS, generative models have been trained on broader chemical spaces, including natural products, industrial chemicals, and environmental pollutants^[66]. When conditioned on exact mass, these models generate extensive sets of candidate structures and rank them by sampling frequency^[38,66]. In

evaluations on unknown environmental pollutants, the true structure was ranked first in 12.6% of cases and appeared within the top ten candidates in 33.1% of cases. Importantly, even when the true structure was not generated, top-ranked candidates often displayed strong structural similarity to the true molecule, thereby providing testable hypotheses for subsequent targeted validation experiments^[66].

Retention time prediction

As an orthogonal parameter to mass spectral information, RT substantially reduces false positives in NTA and plays a pivotal role in structure confirmation and candidate prioritization in LC-HRMS workflows^[6,24,67]. The integration of RT with MS/MS information for joint scoring has therefore become a key trend^[68].

Despite its utility, RT prediction has long been challenged by data scarcity and limited cross-system transferability due to heterogeneous chromatographic conditions. Emerging machine-learning approaches offer improved scalability and demonstrate a certain degree of transferability across chromatographic systems, positioning them as a promising solution for future RT-assisted identification frameworks. Several representative approaches have been proposed recently to realize these advantages.

Domingo-Almenara et al.^[69] established the METLIN-SMRT dataset, comprising 80,038 small molecules, which provides a standardized foundation for ML-based RT prediction. Building on this resource, Ju et al.^[70] proposed DNNpwa-TL, a deep neural network pretrained with weighted autoencoders on the METLIN-SMRT set and subsequently fine-tuned to new chromatographic systems, achieving a median relative error of 3.1% and a mean relative error of 4.9%, outperforming other ML baselines across 17 target datasets. Similarly, Kwon et al.^[71] introduced an improved transfer-learning strategy, adapting a GNN pretrained on METLIN-SMRT to target chromatographic systems, enabling accurate RT prediction even with limited labeled data. Fan et al.^[72] developed the MTSCAM

model, which incorporates a functional group-based classification system and trains specialized sub-models for 141 compound subclasses. Through SMILES enumeration and topological similarity expansion, the model achieved enhanced accuracy and generalization for RT prediction under specified conditions. Kumari et al.^[73] further applied transfer learning to construct multi-target models that facilitate knowledge sharing across reversed-phase LC platforms. Most recently, the ABCoRT model^[74] leveraged atom-bond co-learning in an end-to-end graph framework, predicting RTs with an average relative error of only 3.24% on METLIN-SMRT and demonstrating strong transferability on the PredRet dataset, highlighting the universality of graph-based approaches for RT prediction.

Collision cross section prediction

Ion mobility spectroscopy (IMS) separates ions in the gas phase based on their mobility through a buffer gas under an electric field^[75–77]. The resulting mobility differences are quantified as CCS, which reflects ion size, shape, and mass and provides valuable structural information for compound identification^[78]. When combined with MS/MS and RT data, CCS enables isomer differentiation, reduces false positives, and enhances confidence in NTA^[79,80]. However, obtaining extensive experimental CCS values remains costly, motivating the development of machine-learning-based CCS prediction methods as an effective complementary solution^[36].

Early work by Zhou et al.^[81] introduced support vector regression (SVR) with molecular descriptors for CCS prediction in metabolites, generating 35,203 CCS values. Subsequent efforts extended this strategy to lipids, resulting in > 60,000 predicted CCS values and demonstrating improved candidate ranking and reduced false positives^[82].

Building on these traditional approaches, artificial neural network (ANN)-based models have further enhanced CCS prediction accuracy and broadened applicability^[24]. DeepCCS employed CNNs with SMILES and ion type inputs, achieving an R^2 of 0.97 and a median absolute relative error (MedARE) of ~ 2.7 ^[83]. Later frameworks, such as AllCCS2, incorporated graph convolutional networks (GCNs), multilayer perceptrons (MLPs), and enhanced molecular representations to improve transferability and enable accurate CCS prediction across different instrument platforms^[84]. SigmaCCS introduced 3D molecular information through edge-conditioned convolutions, achieving $R^2 = 0.9945$ and $MedARE = 1.175\%$ on CCSBase with improved interpretability^[85].

To address analyte class heterogeneity, MGAT-CCS introduced class-guided multimodal attention for lipids and metabolites, reducing lipid prediction error to 0.47% (positive mode) and lowering false positives by $\sim 25\%$ ^[86]. More recently, foundation-model-based architectures such as HyperCCS^[87] integrated pretrained chemical language models with cross-modal fusion of fingerprints, mass, and adduct information, demonstrating strong isomer discrimination and cross-domain generalization.

While substantial progress has been made in metabolomics and lipidomics, mainstream CCS predictors remain less effective for environmental organic pollutants due to greater structural diversity and limited available training data^[88,89]. Between 2016 and 2022, only 9,407 CCS values across 4,170 organic micropollutants were reported, a number that remains insufficient for broad-scale environmental analysis^[90].

To bridge this gap, pollutant-focused CCS models have emerged. Bijlsma et al.^[91] demonstrated CCS prediction for environmental chemicals using a genetic-algorithm-optimized ANN, achieving < 6% error in 95% of protonated molecules and enabling pesticide

screening in spinach extracts. Mollerup et al.^[26] expanded the ANN approach and introduced a joint CCS-RT predictor, showing robust performance across 357 pharmaceutical CCS values and demonstrating simultaneous prediction of orthogonal parameters for high-throughput screening. Foster et al.^[92] developed a halogen-focused SVR model for PFAS and other halogenated compounds, achieving median error $\leq 1.0\%$ and enabling discovery of novel PFAS in human serum. Song et al.^[93] trained SVM models for extractables and leachables from plastics using > 1,000 curated chemicals, significantly improving confidence in river-water screening workflows.

Machine learning-assisted quantification in non-targeted analysis

Quantification in LC-HRMS relies on establishing a functional relationship between signal intensity and known concentration. For identified compounds, a calibration-derived RF enables the conversion of peak area to concentration. RF represents the effective signal produced per unit concentration and is governed by IE and matrix suppression/enhancement. The response can vary by several orders of magnitude across analytes due to the high sensitivity of IE to molecular properties, mobile-phase composition, and matrix effects^[94–96]. In the absence of analytical standards, conventional quantification cannot be implemented, making reliable quantification a major challenge in NTA^[97]. ML has recently been introduced to provide alternative, data-driven solutions through surrogate selection, IE modeling, and RF prediction, enabling robust semi-quantification of compounds without available standards.

In the absence of reference standards, structurally or chromatographically similar compounds are often selected as surrogates for quantification. However, surrogate selection is inherently subjective and relies heavily on expert judgment, introducing uncontrolled variability and unclear sources of quantification bias, ultimately increasing uncertainty^[21]. To address this issue, a recent data-driven surrogate selection strategy enables the quantification of unknown chemicals without standards by employing multimodal learning to integrate structural, chromatographic, and MS-derived features to predict ionization similarity^[8]. Each modality contributes an ionization score, which is fused via a transformer model to cluster compounds with comparable ionization behavior. A representative surrogate in each cluster provides a calibration curve, enabling proxy-based quantification and reducing subjectivity in surrogate selection for NTA. This approach demonstrated robust performance, achieving an average error of 4.03-fold in internal cross-prediction and a median 3.21-fold error in external validation, indicating reliable semi-quantification without authentic standards^[8].

IE is the primary determinant of inter-compound response differences. Predicting logIE is functionally equivalent to RF normalization and thus forms the foundation of standard-free quantification^[30,97]. Numerous models have been developed for logIE prediction (Table 1).

Early logIE prediction relied on multiple linear regression (MLR) within restricted chemical classes and identified hydrophobicity, molecular weight, and charge as dominant drivers, yielding R^2 values of 0.7–0.8. However, the predictive accuracy dropped sharply when chemical diversity increased^[103]. Ensemble models capture nonlinear structure-IE relationships more effectively. The random-forest model developed by Krue et al.^[30] was trained on 353 cations and 101 anions across 106 mobile-phase conditions, achieving an error within a factor of two in negative ionization mode. Subsequent studies reported eXtreme Gradient Boosting models (XGB) that improve predictions in under-represented regions of chemical

Table 1 Comparison of the logIE models from the literature

Source	Mode	Number of compounds	Type of compounds	Datapoints	Algorithm	Performance on test set	
						R^2_{test}	$E_{\text{RMS-test}}$
Nei et al. ^[98]	ESI+	62	Metabolites	815	ANN	0.84	0.40
Kaskow et al. ^[99]	ESI+	241	Peptide	241	RF	0.844	/
Aalizadeh et al. ^[100]	ESI+	103	Emerging contaminants	103	MLR	0.803	0.309
Cramer et al. ^[101]	ESI+	77	Drug-like compounds	77	MLR	0.905	/
Liigand et al. ^[30]	ESI+	353	Emerging contaminants	3319	RF	/	0.477
Hermans et al. ^[102]	ESI−	127	Natural Products	127	MLR	0.7	/
Liigand et al. ^[30]	ESI−	101	Emerging contaminants	1286	RF	/	0.362
Lauria et al. ^[29]	ESI−	132	Emerging contaminants	1319	XGB	/	0.79

ESI+: positive mode of electrospray ionization source; ESI−: negative mode of electrospray ionization source; ANN: artificial neural network; RF: random forest; MLR: multiple linear regression; XGB: eXtreme Gradient Boosting; R^2_{test} : the coefficient of determination for the test set; $E_{\text{RMS-test}}$: the root mean squared error for the test set; "/": no related metric.

space, such as PFAS^[29]. Ensemble learning models, such as XGB and random forests, consistently achieve superior IE prediction across diverse chemical classes and currently represent the state of the art in environmental applications. Although ANN-based models have been explored for IE modeling, their application remains limited, primarily due to the small size and sparsity of available IE datasets.

RF prediction can be viewed as an extension of IE modeling, wherein ML regression reconstructs the concentration–response function and can further harmonize response across instruments. RF models typically use molecular descriptors (such as fingerprints, polarity indices, acid-base constants) along with chromatographic and MS conditions as inputs. For example, both random-forest and ANN models have been trained to predict RF from physicochemical features, enabling concentration estimation by predicted RF values^[31].

Due to the susceptibility of response factors to instrumental and experimental conditions, the transferability of predicted RFs remains a general limitation. To address this issue, Tisler et al.^[104] proposed a tandem machine learning strategy to achieve high-accuracy semi-quantification in NTA. First, a quantitative QSPR model was developed to correlate 2D molecular descriptors with the total MS response factor to predict responses for unknown compounds. Crucially, a second transferability QSPR (t-QSPR) model was introduced to correct for instrumental and temporal variations by predicting compound-specific response adjustment factors using a small set of standards. This dual-model framework successfully confined concentration prediction errors within a factor of 2.4–3.3 for most compounds, significantly enhancing the reliability and transferability of semi-quantitative NTA.

Perspective

Despite substantial progress in applying ML to LC-HRMS-based NTA, several fundamental challenges continue to limit its robustness, transferability, and practical adoption in environmental applications. Future advances in this field would benefit from prioritizing the following key research directions.

Representation learning for improving model performance

In tasks involving the prediction of MS/MS spectra, RT, CCS, and IE from molecular structures, model inputs have traditionally relied on handcrafted molecular features including molecular fingerprints and physicochemical descriptors. These descriptors are typically constructed based on expert knowledge and predefined rules and often exhibit strong task dependence. As a result, they may fail to

capture complex structural characteristics. Recent advances in molecular representation learning offer a promising alternative. Representation learning methods based on molecular sequences or graphs, especially when combined with self-supervised learning strategies, enable models to automatically learn informative features directly from raw molecular data without relying on manually designed descriptors. Future studies may explore the use of learned molecular representations to replace or complement conventional features in order to enhance predictive performance and generalization.

For mass spectral data, an additional challenge arises from the discrepancy between reference spectra and real environmental spectra. High-quality mass spectral libraries are typically constructed using pure reference standards measured under controlled experimental conditions, resulting in spectra with low background noise and well-defined fragmentation patterns. In contrast, mass spectra acquired from environmental samples are often affected by complex matrices and co-eluting compounds, leading to elevated noise levels and distorted fragment intensity distributions. These differences can substantially degrade model performance in practical applications.

Spectral representation learning has emerged as a promising strategy to address these challenges. By employing self-supervised learning approaches such as masked learning or contrastive learning, models can automatically extract informative spectral features and encode them into compact vector representations. For example, DeepMASS adopts a Word2Vec-based embedding strategy in which fragment ions or neutral losses are treated as tokens and their co-occurrence patterns are learned to capture latent structural information^[105]. Similarly, DreaMS employs a contrastive self-supervised learning framework to learn spectral representations from millions of MS/MS spectra from samples, enabling them to accurately reflect underlying molecular structures^[106]. Integrating spectral representation learning with downstream tasks such as molecular fingerprint prediction is expected to further improve prediction accuracy and robustness, thereby enhancing compound identification in environmental NTA.

Multimodal learning to enhance model transferability

Qualitative and quantitative parameters are typically influenced by experimental and instrumental conditions. MS/MS spectra are determined by the fragmentation behavior of precursor ions and are strongly influenced by collision energy, fragmentation method, and ionization mode. RT reflects the partitioning of analytes between the stationary and mobile phases and is therefore highly sensitive to chromatographic conditions such as column type, mobile phase

composition, pH, and temperature. Although CCS is often regarded as a quasi-intrinsic property, it depends on instrumental conditions and exhibits variability across platforms. IE is governed by charge acquisition and stabilization processes in the ion source and is strongly affected by experimental and instrumental conditions such as solvent composition, ion source parameters, and pH.

However, existing prediction models have predominantly been based on molecular structural features, while experimental and instrumental factors have been insufficiently considered, resulting in limited cross-platform generalization. To improve model transferability, future research should adopt multimodal learning frameworks that explicitly integrate molecular structural features with auxiliary information describing experimental conditions and instrumental configurations. By jointly modeling these complementary sources of information, multimodal approaches may better capture systematic variability across platforms. In addition, transfer learning and domain adaptation strategies can enhance robustness and reproducibility across laboratories and instruments by fine-tuning pretrained models using limited target domain data.

Improving the applicability of machine learning models to environmental chemical space

A major limitation in applying ML to environmental NTA is the insufficient representation of environmentally relevant compounds in existing datasets. Most publicly available spectral, RT, and CCS databases are dominated by small molecule metabolites commonly studied in metabolomics, such as organic acids, sugars, and lipids, which occupy a relatively constrained chemical space with high functional group redundancy. In contrast, environmental organic pollutants span a much broader and more heterogeneous chemical space, including many structurally unusual compounds that are rare or absent in natural biochemical systems, such as perfluorinated compounds, halogenated aromatic compounds, and multifunctional industrial additives. The limited representation of such compounds can introduce systematic bias into ML models, leading to degraded performance for environmental pollutants. This issue is particularly pronounced for molecular generative models. The scarcity of structural data for emerging pollutants restricts the ability of generative models to learn their characteristic features.

To address these challenges, multitask learning may be employed to improve model generalization by sharing information across related tasks. Data augmentation strategies, such as the generation of simulated spectra or the perturbation of molecular structures, may also increase data diversity. In addition, pretraining models on large datasets from related domains followed by fine-tuning with limited environmental data represents a practical strategy to enhance model applicability within the environmental chemical space.

Bias in molecular generative models and detectability constraints

Molecular generative models inherently learn the statistical distribution of their training data and therefore tend to generate structures that are similar to those that are well-represented in the training set. As a result, chemical classes that are underrepresented or structurally distinct may be inadequately sampled in the chemical space. Enhancing the structural diversity of training sets with environmentally relevant compounds is therefore critical for improving the exploratory capacity of generative models in environmental applications.

In addition, not all generated molecules are equally amenable to detection by LC-HRMS. Some structures may exhibit poor ionization

efficiency, be detectable only under specific ionization modes. Furthermore, some structures have not yet been synthesized or released into the environment. Incorporating constraints related to ionization feasibility, detectability, and environmental relevance into molecular generation workflows may help reduce false-positive candidates and improve the practical utility of generated structures for real-world environmental screening.

Toward automated machine learning-driven screening frameworks

Despite the growing number of ML tools developed for environmental NTA, most existing approaches are applied in isolation and lack integration into unified analytical frameworks. Differences in predictive accuracy, applicability domains, and uncertainty characteristics across models, together with the absence of systematic strategies for result integration and confidence assessment, limit the reliability of composite predictions.

Future efforts should focus on the development of integrated screening platforms that combine qualitative and quantitative ML models within a coherent framework. Multi-model ensemble approaches that weight and fuse predictions from multiple sources, coupled with uncertainty quantification methods, represent a promising direction for improving robustness and reproducibility. Ultimately, the development of highly automated, interpretable, and transferable ML-driven systems has the potential to enable high-throughput and reliable screening of organic pollutants in complex environmental samples.

Author contributions

The authors confirm their contributions to the paper as follows: Yuwei Liu: Draft manuscript preparation; Haoyang Xiong: Draft manuscript preparation; Jinhua Liu: Draft manuscript preparation; Huaijun Xie: Review and editing; Jingwen Chen: Review and editing. All authors have read and agreed to the published version of the manuscript.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Funding

This study was supported by the National Natural Science Foundation of China (22376017) and the National Key Research and Development Program of China (2022YFC3902100).

Conflict of interest

The authors declare no competing financial interest.

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] Landrigan PJ, Fuller R, Acosta NJR, Adeyi O, Arnold R, et al. 2018. The Lancet Commission on pollution and health. *Lancet* 391:462–512

- [2] Wang Z, 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* 54:2575–2584
- [3] Liu N, Jin X, Johnson AC, Zhou S, Liu Y, et al. 2025. Pharmaceutical and personal care products (PPCPs) in global surface waters: risk and drivers. *Environmental Science & Technology* 59:19146–19159
- [4] Budd R, Teerlink J, Alaimo C, Wong L, Young TM. 2023. Sub-sewer-shed monitoring to elucidate down-the-drain pesticide sources. *Environmental Science & Technology* 57:5404–5413
- [5] Zhao HN, Hu X, Gonzalez M, Rideout CA, Hobby GC, et al. 2023. Screening p-phenylenediamine antioxidants, their transformation products, and industrial chemical additives in crumb rubber and elastomeric consumer products. *Environmental Science & Technology* 57:2779–2791
- [6] Wang H, Zhong L, Su W, Ruan T, Jiang G. 2024. Machine learning-assisted identification of environmental pollutants by liquid chromatography coupled with high-resolution mass spectrometry. *TrAC Trends in Analytical Chemistry* 180:117988
- [7] Liu J, Xie H, Lu J, Xie Q, Chen J. 2025. Integrated approach for comprehensive screening of indoor semi-volatile organic compounds via passive air sampling coupled with non-targeted analysis. *Environmental Science & Technology* 59:22701–22712
- [8] Cheng F, Li H, Lou X, He L, Wu X, et al. 2025. Event-driven taxonomy (EDT) screening: leveraging effect-based spectral libraries to accelerate semiquantitative nontarget analysis of ahr agonists in sediment in the era of big data. *Environmental Science & Technology* 59:14359–14371
- [9] Xie H, Dong Y, Chen J, Wang X, Fu M. 2021. Development and evaluation of a ceramic diffusive layer based DGT technique for measuring organic micropollutants in seawaters. *Environment International* 156:106653
- [10] Izquierdo-Sandoval D, Sancho JV, Hernández F, Portoles T. 2025. Approaches for GC-HRMS screening of organic microcontaminants: GC-APCI-IMS-QTOF versus GC-EI-Qorbitrap. *Environmental Science & Technology* 59:2436–2448
- [11] Escher BI, Stapleton HM, Schymanski EL. 2020. Tracking complex mixtures of chemicals in our changing environment. *Science* 367:388–392
- [12] Hollender J, Schymanski EL, Singer HP, Ferguson PL. 2017. Nontarget screening with high resolution mass spectrometry in the environment: ready to go? *Environmental Science & Technology* 51:11505–11512
- [13] Hulleman T, Turkina V, O'Brien JW, Chojnacka A, Thomas KV, et al. 2023. Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis. *Environmental Science & Technology* 57:14101–14112
- [14] Aicheler F, Li J, Hoene M, Lehmann R, Xu G, et al. 2015. Retention time prediction improves identification in nontargeted lipidomics approaches. *Analytical Chemistry* 87:7698–7704
- [15] Charbonnet JA, McDonough CA, Xiao F, Schwichtenberg T, Cao D, et al. 2022. Communicating confidence of per- and polyfluoroalkyl substance identification via high-resolution mass spectrometry. *Environmental Science & Technology Letters* 9:473–481
- [16] Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, et al. 2014. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental Science & Technology* 48:2097–2098
- [17] Alvarez-Mora I, Muratuly A, Johann S, Arturi K, Jünger F, et al. 2025. High-throughput effect-directed analysis of androgenic compounds in hospital wastewater: identifying effect drivers through non-target screening supported by toxicity prediction. *Environmental Science & Technology* 59:24513–24525
- [18] Liu M, Xiong G, Ke Y, Wang X, Zhu X, et al. 2025. Integrating suspect and nontarget screening with spatial autocorrelation to prioritize emerging contaminants in a coastal urban river. *Environmental Science & Technology* 59:17222–17234
- [19] Samanipour S, Reid MJ, Thomas KV. 2017. Statistical variable selection: an alternative prioritization strategy during the nontarget analysis of LC-HR-MS data. *Analytical Chemistry* 89:5585–5591
- [20] Cai Y, Zhou Z, Zhu Z. 2023. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. *TrAC Trends in Analytical Chemistry* 158:116903
- [21] Mccord JP, Groff LC, Sobus JR. 2022. Quantitative non-targeted analysis: bridging the gap between contaminant discovery and risk characterization. *Environment International* 158:107011
- [22] Wang F, Liigand J, Tian S, Arndt D, Greiner R, et al. 2021. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Analytical Chemistry* 93:11692–11700
- [23] Wei JN, Belanger D, Adams RP, Sculley D. 2019. Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Central Science* 5:700–708
- [24] Jin Z, Chen L, Wang Y, Shi C, Zhou Y, et al. 2025. Application of machine learning in LC-MS-based non-targeted analysis. *TrAC Trends in Analytical Chemistry* 189:118243
- [25] Hong Y, Ye Y, Tang H. 2025. Machine learning in small-molecule mass spectrometry. *Annual Review of Analytical Chemistry*, 18:193–215
- [26] Møllerup CB, Mardal M, Dalsgaard PW, Linnet K, Barron LP. 2018. Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. *Journal of Chromatography A* 1542:82–88
- [27] Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, et al. 2019. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* 16:299–302
- [28] Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. 2016. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* 8:3
- [29] Lauria MZ, Sepman H, Ledbetter T, Plassmann M, Roos AM, et al. 2024. Closing the organofluorine mass balance in marine mammals using suspect screening and machine learning-based quantification. *Environmental Science & Technology* 58:2458–2467
- [30] Liigand J, Wang T, Kellogg J, Smedsgaard J, Cech N, et al. 2020. Quantification for non-targeted LC/MS screening without standard substances. *Scientific Reports* 10:5808
- [31] Abrahamsson D, Park J, Singh RR, Sirota M, Woodruff TJ. 2020. Applications of machine learning to in silico quantification of chemicals without analytical standards. *Journal of Chemical Information and Modeling* 60:2718–2727
- [32] Chen B, Li H, Huang R, Tang Y, Li F. 2024. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. *Nature Communications* 15:8396
- [33] Dührkop K, Nothias L, Fleischauer M, Reher R, Ludwig M, et al. 2021. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* 39:462–471
- [34] Fan Z, Alley A, Ghaffari K, Ransom HW. 2020. Metfid: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics* 16:104
- [35] Wang X, Yu N, Jiao Z, Li L, Yu H, et al. 2024. Machine learning-enhanced molecular network reveals global exposure to hundreds of unknown pfas. *Science Advances* 10:eadn1039
- [36] Li X, Wang H, Jiang M, Ding M, Xu X, et al. 2023. Collision cross section prediction based on machine learning. *Molecules* 28:4050
- [37] Stravs MA, Dührkop K, Böcker S, Zamboni N. 2022. MSNovelist: de novo structure generation from mass spectra. *Nature Methods* 19:865–870
- [38] Skinnider MA, Wang F, Pasin D, Greiner R, Foster LJ, et al. 2021. A deep generative model enables automated structure elucidation of novel psychoactive substances. *Nature Machine Intelligence* 3:973–984
- [39] Hupatz H, Rahu I, Wang W, Peets P, Palm EH, et al. 2025. Critical review on in silico methods for structural annotation of chemicals detected with LC/HRMS non-targeted screening. *Analytical and Bioanalytical Chemistry* 417:473–493
- [40] Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RCH, van Schaik R, et al. 2012. Substructure-based annotation of high-resolution multi-stage msn spectral trees. *Rapid Communications in Mass Spectrometry* 26:2461–2471

- [41] Djoumbou-Feunang Y, Pon A, Karu N, Zheng J, Li C, et al. 2019. CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification. *Metabolites* 9:72
- [42] Wang F, Allen D, Tian S, Oler E, Gautam V, et al. 2022. CFM-ID 4.0—a web server for accurate MS-based metabolite identification. *Nucleic Acids Research* 50:165–174
- [43] Allen F, Greiner R, Wishart D. 2015. Competitive fragmentation modeling of esi-ms/ms spectra for putative metabolite identification. *Metabolomics* 11:98–110
- [44] Goldman S, Li J, Coley CW. 2024. Generating molecular fragmentation graphs with autoregressive neural networks. *Analytical Chemistry* 96:3419–3428
- [45] Nowatzky Y, Russo FF, Lisek J, Kister A, Reinert K, et al. 2025. Fiora: local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nature Communications* 16:2298
- [46] Hong Y, Li S, Welch CJ, Tichy S, Ye Y, et al. 2023. 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* 39:btad354
- [47] Young A, Röst H, Wang B. 2024. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence* 6:404–416
- [48] Park J, Jo J, Yoon S. 2024. Mass spectra prediction with structural motif-based graph neural networks. *Scientific Reports* 14:1400
- [49] Overstreet R, King E, Clopton G, Nguyen J, Ciesielski D. 2024. QC-GN2OMS2: a graph neural net for high resolution mass spectra prediction. *Journal of Chemical Information and Modeling* 64:5806–5816
- [50] Goldman S, Xin J, Provenzano J, Coley CW. 2024. Mist-cf: chemical formula inference from tandem mass spectra. *Journal of Chemical Information and Modeling* 64:2421–2431
- [51] Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. 2015. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proceedings of the National Academy of Sciences* 112:12580–12585
- [52] Goldman S, Wohlwend J, Stražar M, Haroush G, Xavier RJ, et al. 2023. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence* 5:965–979
- [53] Baygi SF, Barupal DK. 2024. IdsL_mint: a deep learning framework to predict molecular fingerprints from mass spectra. *Journal of Cheminformatics* 16:8
- [54] Evich MG, Davis M, Weber EJ, Tebes-Stevens C, Acrey B, et al. 2022. Environmental fate of Cl-PPPECAS: predicting the formation of PFAS transformation products in new jersey soils. *Environmental Science & Technology* 56:7779–7788
- [55] Du Y, Jamasb AR, Guo J, Fu T, Harris C, et al. 2024. Machine learning-aided generative molecular design. *Nature Machine Intelligence* 6:589–604
- [56] Anstine DM, Isayev O. 2023. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society* 145:8736–8750
- [57] Pang C, Qiao J, Zeng X, Zou Q, Wei L. 2024. Deep generative models in de novo drug molecule generation. *Journal of Chemical Information and Modeling* 64:2174–2194
- [58] Sousa T, Correia J, Pereira V, Rocha M. 2021. Generative deep learning for targeted compound design. *Journal of Chemical Information and Modeling* 61:5343–5361
- [59] Shrivastava AD, Swainston N, Samanta S, Roberts I, Wright Muelas M, et al. 2021. MassGenie: a transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules* 11:1793
- [60] Butler T, Frandsen A, Lightheart R, Bargh B, Taylor J, et al. 2023. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv*
- [61] Elser D, Huber F, Gaquerel E. 2023. Mass2SMILES: deep learning based fast prediction of structures and functional groups directly from high-resolution MS/MS spectra. *bioRxiv*
- [62] Litsa EE, Chenthamarakshan V, Das P, Kavraki LE. 2023. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry* 6:132
- [63] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, et al. , 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv: 1910.13461 <https://doi.org/10.48550/arXiv.1910.13461>
- [64] Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, et al. 2016. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology* 34:828–837
- [65] Winter R, Montanari F, Noé F, Clevert D. 2019. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science* 10:1692–1701
- [66] Skinnider MA. 2024. Invalid smiles are beneficial rather than detrimental to chemical language models. *Nature Machine Intelligence* 6:437–448
- [67] Xue J, Wang B, Ji H, Li W. 2024. RT-Transformer: retention time prediction for metabolite annotation to assist in metabolite identification. *Bioinformatics* 40:btac84
- [68] Hu M, Müller E, Schymanski EL, Ruttkies C, Schulze T, et al. 2018. Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS. *Analytical and Bioanalytical Chemistry* 410:1931–1941
- [69] Domingo-Almenara X, Guijas C, Billings E, Montenegro-Burke JR, Uritboonthai W, et al. 2019. The metlin small molecule dataset for machine learning-based retention time prediction. *Nature Communications* 10:5811
- [70] Ju R, Liu X, Zheng F, Lu X, Xu G, et al. 2021. Deep neural network pretrained by weighted autoencoders and transfer learning for retention time prediction of small molecules. *Analytical Chemistry* 93:15651–15658
- [71] Kwon Y, Kwon H, Han J, Kang M, Kim J, et al. 2023. Retention time prediction through learning from a small training data set with a pretrained graph neural network. *Analytical Chemistry* 95:17273–17283
- [72] Fan M, Sang C, Li H, Wei Y, Zhang B, et al. 2025. Development of an efficient and generalized MTSCAM model to predict liquid chromatography retention times of organic compounds. *Research* 8:607
- [73] Kumari P, Guilherme MSR, Choudhary P, Van Laethem T, Fillet M, et al. 2024. Transfer learning approach to multitarget QSRR modeling in RPLC. *Journal of Chemical Information and Modeling* 64:7447–7456
- [74] Cheng G, Wang B, Bai N, Li W. 2025. Abcort: retention time prediction for metabolite identification via atom-bond co-learning. *Journal of Chemical Information and Modeling* 65:1419–1427
- [75] George AC, Schmitz-Afonso I, Marie V, Colsch B, Fenaille F, et al. 2022. A re-calibration procedure for interoperable lipid collision cross section values measured by traveling wave ion mobility spectrometry. *Analytica Chimica Acta* 1226:340236
- [76] D'Atri V, Causon T, Hernandez Alba O, Mutabazi A, Veuthey JL, et al. 2018. Adding a new separation dimension to MS and LC-MS: what is the utility of ion mobility spectrometry? *Journal of Separation Science* 41:20–67
- [77] Lanucara F, Holman SW, Gray CJ, Evers CE. 2014. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nature Chemistry* 6:281–294
- [78] Celma A, Sancho JV, Schymanski EL, Fabregat-Safont D, Ibáñez M, et al. 2020. Improving target and suspect screening high-resolution mass spectrometry workflows in environmental analysis by ion mobility separation. *Environmental Science & Technology* 54:15120–15131
- [79] Castellanos A, Benigni P, Hernandez DR, Debord JD, Ridgeway ME, et al. 2014. Fast screening of polycyclic aromatic hydrocarbons using trapped ion mobility spectrometry – mass spectrometry. *Analytical Methods* 6:9328–9332
- [80] Rüger CP, Le Maître J, Maillard J, Riches E, Palmer M, et al. 2021. Exploring complex mixtures by cyclic ion mobility high-resolution mass spectrometry: application toward petroleum. *Analytical Chemistry* 93:5872–5881
- [81] Zhou Z, Shen X, Tu J, Zhu Z. 2016. Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical Chemistry* 88:11084–11091

- [82] Zhou Z, Tu J, Xiong X, Shen X, Zhu Z. 2017. Lipidccs: prediction of collision cross-section values for lipids with high precision to support ion mobility–mass spectrometry-based lipidomics. *Analytical Chemistry* 89:9559–9566
- [83] Plante P, Francovic-Fontaine É, May JC, Mclean JA, Baker ES, et al. 2019. Predicting ion mobility collision cross-sections using a deep neural network: DeepCCS. *Analytical Chemistry* 91:5191–5199
- [84] Zhang H, Luo M, Wang H, Ren F, Yin Y, et al. 2023. Allccs2: curation of ion mobility collision cross-section atlas for small molecules using comprehensive molecular representations. *Analytical Chemistry* 95:13913–13921
- [85] Guo R, Zhang Y, Liao Y, Yang Q, Xie T, et al. 2023. Highly accurate and large-scale collision cross sections prediction with graph neural networks. *Communications Chemistry* 6:139
- [86] Wang C, Yuan C, Wang Y, Shi Y, Zhang T, et al. 2024. Predicting collision cross-section values for small molecules through chemical class-based multimodal graph attention network. *Journal of Chemical Information and Modeling* 64:6305–6315
- [87] Zhu Z, Xie C, Lin S, Xu X, Niu S, et al. 2025. Large language model-empowered compound collision cross-section prediction. *Analytical Chemistry* 97:19791–19800
- [88] Righetti L, Dreolin N, Celma A, McCullagh M, Barknowitz G, et al. 2020. Travelling wave ion mobility-derived collision cross section for mycotoxins: investigating interlaboratory and interplatform reproducibility. *Journal of Agricultural and Food Chemistry* 68:10937–10943
- [89] Song X, Canellas E, Dreolin N, Goshawk J, Nerin C. 2022. A collision cross section database for extractables and leachables from food contact materials. *Journal of Agricultural and Food Chemistry* 70:4457–4466
- [90] Song X, Canellas E, Dreolin N, Goshawk J, Lv M, et al. 2023. Application of ion mobility spectrometry and the derived collision cross section in the analysis of environmental organic micropollutants. *Environmental Science & Technology* 57:21485–21502
- [91] Bijlsma L, Bade R, Celma A, Mullin L, Cleland G, et al. 2017. Prediction of collision cross-section values for small molecules: application to pesticide residue analysis. *Analytical Chemistry* 89:6583–6589
- [92] Foster M, Rainey M, Watson C, Dodds JN, Kirkwood KI, et al. 2022. Uncovering pfas and other xenobiotics in the dark metabolome using ion mobility spectrometry, mass defect analysis, and machine learning. *Environmental Science & Technology* 56:9133–9143
- [93] Song X, Dreolin N, Canellas E, Goshawk J, Nerin C. 2022. Prediction of collision cross-section values for extractables and leachables from plastic products. *Environmental Science & Technology* 56:9463–9473
- [94] Kiontke A, Oliveira-Birkmeier A, Opitz A, Birkemeyer C. 2016. Electrospray ionization efficiency is dependent on different molecular descriptors with respect to solvent pH and instrumental configuration. *PLOS One* 11:e167502
- [95] Krueve A. 2020. Strategies for drawing quantitative conclusions from nontargeted liquid chromatography–high-resolution mass spectrometry analysis. *Analytical Chemistry* 92:4691–4699
- [96] Raji MA, Schug KA. 2009. Chemometric study of the influence of instrumental parameters on esi-ms analyte response using full factorial design. *International Journal of Mass Spectrometry* 279:100–106
- [97] Malm L, Liigand J, Aalizadeh R, Alygizakis N, Ng K, et al. 2024. Quantification approaches in non-target LC/ESI/HRMS analysis: an interlaboratory comparison. *Analytical Chemistry* 96:16215–16226
- [98] Nie C, Liu H, Huang X, Zhou D, Wang X, et al. 2024. Prediction of mass spectrometry ionization efficiency based on cosmo-rs and machine learning algorithms. *Analyst* 149:3140–3151
- [99] Kaskow JA, Hahnert ET, Porter TK, Lu Y, Stanev V, et al. 2024. Predicting peptide ionization efficiencies for electrospray ionization mass spectrometry using machine learning. *Journal of the American Society for Mass Spectrometry* 35:2297–2307
- [100] Aalizadeh R, Nikolopoulou V, Alygizakis N, Slobodnik J, Thomaidis NS. 2022. A novel workflow for semi-quantification of emerging contaminants in environmental samples analyzed by LC-HRMS. *Analytical and Bioanalytical Chemistry* 414:7435–7450
- [101] Cramer CJ, Johnson JL, Kamel AM. 2017. Prediction of mass spectral response factors from predicted chemometric data for druglike molecules. *Journal of the American Society for Mass Spectrometry* 28:278–285
- [102] Hermans J, Ongay S, Markov V, Bischoff R. 2017. Physicochemical parameters affecting the electrospray ionization efficiency of amino acids after acylation. *Analytical Chemistry* 89:9159–9166
- [103] Golubović J, Birkemeyer C, Protić A, Otašević B, Zečević M. 2016. Structure–response relationship in electrospray ionization-mass spectrometry of sartans by artificial neural networks. *Journal of Chromatography A* 1438:123–132
- [104] Tisler S, Kilpinen K, Pattison DI, Tomasi G, Christensen JH. 2024. Quantitative nontarget analysis of CECs in environmental samples can be improved by considering all mass adducts. *Analytical Chemistry* 96:229–237
- [105] Ji H, Du R, Dai Q, Su M, Lyu Y, et al. 2024. DeepMASS: unknown compound annotation using semantic similarity of mass spectral language and chemical space localization. *bioRxiv*
- [106] Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, et al. 2025. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nature Biotechnology*



Copyright: © 2026 by the author(s). Published by NEW PRESS LIMITED, TSIM SHA TSUI, HK. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0>.