OPEN ACCESS

Scholars International Journal of Chemistry and Material Sciences

Abbreviated Key Title: Sch Int J Chem Mater Sci
ISSN 2616-8669 (Print) |ISSN 2617-6556 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: http://saudijournals.com

Review Article

AI-Driven Analytical and Molecular Data Modeling for Environmental and Pharmaceutical Applications

Noman Hassan[1]*, Umar Farooq[1], Shumaila Raheem[1], Ariba Anwar[1], Tasawar Abbas[1], Allah Ditta Shah[1], Areeba Mumtaz[1], Alishba Zaheer[1], Sidra Rehman[1], Laraib Umar[1]

[1]Department of Chemistry, Faculty of Sciences, Thal University Bhakkar, Bhakkar 30000, Punjab, Pakistan

DOI: https://doi.org/10.36348/sijcms.2026.v09i04.002 | **Received:** 03.05.2026 | **Accepted:** 25.06.2026 | **Published:** 11.07.2026

***Corresponding author:** Noman Hassan
Department of Chemistry, Faculty of Sciences, Thal University Bhakkar, Bhakkar 30000, Punjab, Pakistan

Abstract

Artificial intelligence is reshaping chemical research by linking high-dimensional analytical signals with molecular, biological, environmental, and pharmaceutical information. This review synthesizes literature from 2018–2026 on chemometrics, machine learning, deep learning, graph neural networks, transformers, foundation models, multimodal learning, and generative systems. It examines data obtained from spectroscopy, chromatography, mass spectrometry, electrochemical sensors, hyperspectral imaging, process monitoring, molecular descriptors, fingerprints, SMILES, graphs, three-dimensional structures, proteins, and omics. Environmental applications include contaminant detection and quantification, suspect and non-target screening, source attribution, fate and transport prediction, ecotoxicity assessment, wastewater-treatment evaluation, and ecological-risk prioritization. Pharmaceutical applications encompass raw-material authentication, quality control, impurity profiling, formulation and drug-delivery optimization, continuous manufacturing, real-time release testing, virtual screening, molecular design, and ADMET prediction. Across both domains, AI improves nonlinear pattern recognition, structure–signal translation, candidate ranking, and multi-objective optimization; however, sophisticated models do not consistently outperform well-designed chemometric approaches, particularly with small or biased datasets. Major barriers include class imbalance, limited chemical diversity, data leakage, instrument variability, missing metadata, weak external validation, poor uncertainty calibration, limited interpretability, and regulatory concerns. Future progress requires FAIR multimodal datasets, independent validation, applicability-domain analysis, explainable and uncertainty-aware models, digital twins, federated learning, autonomous laboratories, human oversight, and sustainable computing for trustworthy scientific deployment.

Keywords: Artificial intelligence; Analytical chemistry; Molecular data modeling; Environmental contaminant analysis; Pharmaceutical development; Multimodal learning; ADMET prediction.

INTRODUCTION

The evolution of analytical and molecular data modeling reflects a broader transformation in chemical research from the statistical interpretation of relatively structured datasets to the automated learning of complex relationships across heterogeneous and high-dimensional data. Conventional chemometrics established the conceptual and methodological foundation for this transformation by integrating mathematics, statistics, and computation with chemical measurement science. Techniques such as principal component analysis (PCA), partial least-squares regression (PLSR), linear discriminant analysis, cluster analysis, and multivariate curve resolution enabled researchers to reduce dimensionality, recognize patterns, quantify analytes, and separate overlapping chemical signals. These methods remain central to exploratory analysis, calibration, classification, process monitoring, and quality control, particularly when datasets are moderate in size and the relationships between predictors and responses are approximately linear (Biancolillo & Marini, 2018; Brereton *et al.,*2018).

Among these methods, PCA became one of the most widely adopted unsupervised approaches for transforming correlated analytical variables into a smaller number of orthogonal latent variables that capture the principal sources of variance. In spectroscopy, chromatography, metabolomics, and sensor-based analysis, PCA has been extensively used to

visualize sample clustering, identify outliers, examine instrumental drift, and detect systematic differences among experimental groups. PLSR extended latent-variable modeling to supervised prediction by identifying components that maximize covariance between an analytical predictor matrix and a measured response. It therefore became particularly important for quantitative spectroscopic calibration, where hundreds or thousands of wavelength or wavenumber variables may be used to estimate analyte concentrations, physicochemical properties, or pharmaceutical quality attributes. These techniques provided interpretable scores, loadings, regression coefficients, and validation procedures, creating a disciplined framework for converting multivariate measurements into chemically meaningful conclusions (Biancolillo & Marini, 2018; Brereton *et al.,*2018).

However, the analytical landscape has changed substantially with the proliferation of high-throughput and high-resolution technologies. Modern experiments can generate thousands to millions of features from a single study. Vibrational, optical, and magnetic-resonance techniques produce densely sampled spectra containing nonlinear baselines, overlapping bands, scattering effects, matrix interferences, and instrument-dependent variations. Chromatographic systems generate multidimensional retention, peak-shape, and intensity information, while high-resolution and tandem mass spectrometry produce large feature tables, isotope patterns, fragmentation spectra, molecular-formula candidates, and retention-time information. Sensor arrays, imaging systems, and portable devices add spatial, temporal, and multichannel dimensions. At the same time, molecular research increasingly incorporates chemical structures, molecular descriptors, fingerprints, graphs, protein–ligand interactions, toxicological endpoints, omics profiles, biological activities, environmental exposure variables, and pharmaceutical process parameters. The scale and heterogeneity of these data have exceeded the practical capacity of manual interpretation and, in many cases, the representational flexibility of conventional linear models (Beck *et al.,*2024; Xue *et al.,*2023).

This transition does not imply that artificial intelligence has rendered classical chemometrics obsolete. Rather, machine learning can be viewed as an expansion of chemometric reasoning into more flexible computational spaces. Algorithms such as support vector machines, random forests, gradient-boosting methods, k-nearest-neighbour models, Gaussian processes, and shallow artificial neural networks retain the traditional objectives of classification and regression but can capture nonlinear relationships, interactions, thresholds, and hierarchical patterns that may be difficult to represent through PCA or PLSR alone. They are particularly useful when analytical responses depend on complex matrix effects, nonlinear concentration–signal relationships, interacting experimental conditions, or highly structured feature combinations. In practice, effective analytical pipelines frequently combine chemometric preprocessing and dimensionality reduction with machine-learning prediction rather than treating the two traditions as competing paradigms (Brereton *et al.,*2018; Xue *et al.,*2023).

Deep learning marked a further shift by reducing dependence on manually engineered features. Convolutional neural networks can learn local motifs directly from spectra, chromatograms, images, and sensor signals, whereas recurrent and sequence-based architectures can model ordered dependencies across wavelengths, retention times, mass-to-charge ratios, molecular strings, or biological sequences. Autoencoders can compress high-dimensional measurements into latent representations, support denoising, and identify anomalous samples. In mass spectrometry, deep-learning systems increasingly support peak detection, spectral similarity estimation, fragmentation prediction, compound annotation, proteomic analysis, and structure elucidation. For example, Spec2Mol uses an encoder–decoder architecture to learn embeddings from tandem mass spectra and generate candidate molecular structures as simplified molecular-input line-entry system strings. This approach illustrates the transition from conventional spectral matching against predefined libraries to models capable of recommending structures for compounds that may not be represented in reference databases (Beck *et al.,*2024; Litsa *et al.,*2023).

A parallel transformation has occurred in quantitative structure–activity relationship modeling. Conventional QSAR generally depended on predefined physicochemical descriptors, structural fragments, topological indices, or molecular fingerprints that were related to biological activity or toxicity through linear regression, discriminant analysis, or other statistical procedures. Machine learning expanded QSAR by enabling nonlinear mapping between molecular descriptors and biological endpoints, while deep learning progressively shifted the emphasis from descriptor selection to representation learning. Instead of requiring all relevant molecular features to be specified in advance, modern models can learn representations directly from simplified molecular-input line-entry system strings, molecular graphs, three-dimensional conformations, or combinations of structural and experimental information (Soares *et al.,*2022; Yang *et al.,*2019).

Graph neural networks are especially relevant to this development because molecules possess an intrinsic graph structure: atoms can be represented as nodes, chemical bonds as edges, and molecular properties as outcomes emerging from local and global interactions. Message-passing and graph-convolutional architectures update atomic representations by aggregating information from neighbouring atoms and subsequently combine these representations to predict molecular-level properties. Such models have been applied to solubility,

lipophilicity, biological activity, toxicity, metabolic behaviour, permeability, and other environmental and pharmaceutical endpoints. Comparative investigations have shown that learned graph representations can equal or outperform models based exclusively on fixed descriptors in many settings, particularly when adequately large and chemically diverse datasets are available (Wieder *et al.,*2020; Yang *et al.,*2019).

Self-supervised learning has further reduced dependence on costly labelled data. In molecular contrastive learning, a model is pretrained on large collections of unlabelled structures by learning consistent representations of differently augmented versions of the same molecule. MolCLR, for example, used graph neural networks and approximately 10 million unlabelled molecules to learn transferable molecular representations before fine-tuning on classification and regression tasks (Wang *et al.,*2022). This pretraining-and-adaptation strategy represents an important step toward chemical foundation models, which are trained on very large datasets and subsequently adapted to multiple downstream applications. MolE extended this concept by using a transformer architecture and self-supervised pretraining on molecular graphs from approximately 842 million molecules, producing embeddings that could be fine-tuned for absorption, distribution, metabolism, excretion, and toxicity-related prediction tasks (Méndez-Lucio *et al.,*2024).

Transformers have also altered how chemical and analytical information is modeled. Their attention mechanisms allow models to learn long-range relationships without processing inputs solely through local or sequential recurrence. In molecular science, transformers can interpret chemical strings, reaction sequences, spectra, graphs, and textual chemical descriptions. The Molecular Transformer demonstrated that attention-based sequence modeling could be applied to reaction prediction while also providing uncertainty-related information (Schwaller *et al.,* 2019). More recently, transformer architectures have connected experimental analytical signals directly with molecular structures. Alberts *et al.,* (2024), for example, pretrained a transformer on simulated infrared spectra and fine-tuned it using experimental spectra to predict molecular structures from the infrared fingerprint region. This represents a major conceptual advance over conventional functional-group assignment: rather than using a limited number of recognizable peaks, the model attempts to exploit the distributed information contained throughout the spectrum.

Multimodal AI constitutes the next stage of this progression. Traditional analytical and QSAR workflows often treat spectra, molecular structures, biological activities, textual descriptions, imaging information, and experimental metadata as separate sources. Multimodal architectures instead learn relationships across these representations and place them within a shared latent space. MoleculeSTM, for example, jointly learned molecular structures and natural-language descriptions using more than 280,000 structure–text pairs, enabling zero-shot structure–text retrieval and text-guided molecular editing (Liu *et al.,*2023). In environmental and pharmaceutical research, comparable strategies could integrate spectroscopic signatures, chromatographic behaviour, mass fragmentation, molecular graphs, toxicological outcomes, biological pathways, formulation variables, and process conditions. Such integration can potentially support more complete tasks, including unknown-compound identification, hazard prioritization, impurity annotation, activity prediction, formulation optimization, and evidence-guided molecular design.

Generative AI has expanded the role of computation from predicting predefined outputs to producing new data, structures, and hypotheses. Variational autoencoders, generative adversarial networks, autoregressive transformers, normalizing flows, and diffusion models can generate synthetic spectra, augment underrepresented classes, reconstruct missing signals, propose candidate structures, and explore chemical spaces conditioned on desired properties. In spectroscopy, generative systems are being investigated as solutions to limited training data, class imbalance, instrument variation, and the high cost of acquiring labelled experimental measurements. Nevertheless, generated data must preserve chemically meaningful relationships and should not be treated as equivalent to independently measured evidence without rigorous validation (Flanagan *et al.,*2025).

Importantly, the historical progression from PCA, PLSR, and conventional QSAR to graph networks, transformers, foundation models, and multimodal AI should not be interpreted as a simple sequence in which each new method universally outperforms its predecessor. Large-scale benchmarking has demonstrated that fixed descriptors and conventional machine-learning models can remain highly competitive, particularly for small datasets, poorly defined endpoints, or chemical spaces insufficiently represented during training. Deng *et al.,* (2023) showed that representation-learning models did not consistently outperform fixed molecular representations across numerous molecular-property datasets and that performance was strongly influenced by dataset size, splitting strategy, label quality, structural diversity, and activity cliffs. Therefore, model complexity cannot compensate for weak experimental design, unsuitable data, inadequate chemical-space coverage, or inappropriate validation.

The emergence of AI-driven molecular analytics is consequently best understood as an extension of the chemometric tradition rather than its replacement. Conventional chemometrics provides interpretable latent-variable methods, disciplined preprocessing, calibration theory, and validation principles. Machine

learning adds nonlinear predictive flexibility; deep learning enables automated representation learning from high-dimensional raw inputs; graph networks incorporate molecular topology; transformers model long-range and cross-domain dependencies; foundation models transfer knowledge from large unlabelled datasets; multimodal systems connect complementary experimental and molecular representations; and generative AI supports data augmentation and molecular design. Their combined use provides a pathway from isolated analytical measurements toward integrated, automated, and predictive systems. For environmental and pharmaceutical applications, the scientific value of this transition will ultimately depend not only on predictive accuracy but also on data quality, external validation, uncertainty quantification, interpretability, applicability-domain assessment, and the ability of model outputs to support reproducible and chemically defensible decisions.

The increasing convergence of experimental analytical measurements and computational molecular information represents a fundamental shift in chemical data science. Traditionally, instrumental signals and molecular structures were treated as distinct information layers. Spectroscopists interpreted peaks, bands, and chemical shifts; chromatographers evaluated retention profiles and peak areas; mass spectrometrists examined mass-to-charge ratios and fragmentation patterns; and computational chemists separately calculated molecular descriptors, fingerprints, conformations, and structure–activity relationships. Artificial intelligence is progressively dissolving these disciplinary boundaries by enabling analytical signals, chemical structures, biological measurements, and contextual metadata to be processed within a common predictive framework. In this emerging paradigm, experimental observations are no longer regarded merely as final measurements, and molecular representations are not limited to static database entries. Instead, both become complementary views of the same chemical system that can be aligned, transformed, and integrated to support identification, quantification, classification, property prediction, toxicity assessment, formulation development, and molecular design.

A useful conceptual model for this convergence is a signal-to-representation-to-decision framework. The first stage consists of acquiring experimental data from one or more analytical platforms. The second converts raw measurements and chemical structures into standardized numerical representations. The third applies statistical, machine-learning, or deep-learning models to learn relationships among analytical patterns, molecular attributes, experimental conditions, and target outcomes. The fourth produces an actionable result, such as an identified contaminant, a predicted concentration, a classified pharmaceutical sample, an estimated toxicological endpoint, an optimized formulation, or a newly proposed molecular structure. Model outputs may subsequently guide confirmatory experiments, producing a feedback loop in which new analytical observations refine the computational model. This closed relationship between measurement and prediction differentiates contemporary AI-enabled molecular analytics from conventional workflows in which experimental and computational analyses are conducted sequentially and largely independently.

The experimental side of this framework encompasses diverse analytical platforms that encode complementary chemical information. Fourier-transform infrared spectroscopy and related mid-infrared techniques measure molecular vibrational absorption and are particularly informative for functional-group identification, bond environments, and molecular fingerprinting. Near-infrared spectroscopy records overtone and combination vibrations and is widely applied to rapid, nondestructive quantification, moisture analysis, raw-material authentication, and pharmaceutical process monitoring. Raman spectroscopy provides vibrational information that is complementary to infrared spectroscopy and is especially useful for aqueous samples, polymorph discrimination, biological analysis, microplastic characterization, and spatially resolved chemical imaging. Although these techniques generate spectra with strong chemical specificity, their interpretation can be complicated by overlapping bands, baseline drift, fluorescence interference, scattering effects, matrix variability, and instrument-to-instrument differences. AI models can learn distributed spectral patterns that extend beyond individually assigned peaks, thereby supporting classification, regression, denoising, spectral unmixing, and structure inference (Xue *et al.,*2023).

Ultraviolet–visible absorption and fluorescence spectroscopy provide another important class of analytical inputs. UV–visible spectra encode electronic transitions and are commonly used for concentration determination, reaction monitoring, water-quality assessment, nanoparticle characterization, and pharmaceutical assays. Fluorescence measurements can offer high sensitivity and may generate excitation, emission, lifetime, or multidimensional fluorescence landscapes. However, fluorescence responses often depend on pH, solvent environment, quenching, photobleaching, matrix composition, and interactions among analytes. Machine-learning models can exploit the complete spectral response rather than relying on a single analytical wavelength, allowing chemically overlapping signals to be separated and multiple analytes to be quantified simultaneously. This is especially valuable for complex environmental and biological matrices where conventional univariate calibration is insufficient.

Nuclear magnetic resonance spectroscopy contributes detailed information on atomic connectivity, chemical environments, stereochemistry, and molecular

dynamics. One-dimensional proton and carbon spectra, together with multidimensional correlation experiments, provide rich structural constraints but generate complex data that often require extensive specialist interpretation. AI can support chemical-shift prediction, peak assignment, spectral matching, dereplication, candidate ranking, and automated structure elucidation. NMR also complements techniques such as mass spectrometry because it provides structural information that is not dependent on ionization or fragmentation. A particularly important recent advance is the development of systems that integrate multiple spectroscopic modalities rather than interpreting each instrument independently. Mirza et al. (2026), for example, introduced a framework that aligned representations derived from raw NMR, infrared, and mass-spectrometric data, demonstrating how complementary experimental views can be combined for end-to-end molecular structure elucidation.

Chromatographic and mass-spectrometric systems generate an additional layer of chemical and temporal information. Liquid chromatography–mass spectrometry combines retention behavior with accurate mass, isotope patterns, adduct formation, signal intensity, and tandem-mass-spectrometric fragmentation. Gas chromatography–mass spectrometry similarly combines chromatographic separation with electron-ionization or other mass spectra, making it particularly important for volatile and semi-volatile compounds. High-resolution mass spectrometry further increases mass accuracy and resolving power, enabling formula assignment and non-target screening across highly complex environmental, pharmaceutical, metabolomic, and biological samples. These systems can generate thousands of features from a single experiment, many of which remain unidentified because authentic standards and experimental reference spectra are unavailable (Beck *et al.,*2024).

AI can act at nearly every stage of an LC–MS, GC–MS, or HRMS workflow. Relevant tasks include peak detection, retention-time alignment, blank filtering, isotope and adduct grouping, molecular-formula prediction, spectral-similarity estimation, compound classification, fragmentation prediction, database searching, and structural-candidate ranking. Deep-learning approaches have moved mass-spectrometric analysis beyond conventional library matching. Spec2Mol, for example, was developed to translate tandem mass spectra into candidate molecular structures, including recommendations for compounds that were not represented in the training reference library (Litsa *et al.,*2023). Such systems illustrate a bidirectional relationship between signals and structures: spectra can be converted into molecular candidates, while molecular structures can be used to predict expected spectra and construct in silico libraries.

Electrochemical sensors constitute another rapidly expanding analytical data source. Instead of producing a conventional optical spectrum, these systems generate current, potential, impedance, conductance, capacitance, or time-dependent response profiles. Sensor arrays may respond partially and non-selectively to several analytes, producing cross-reactive chemical fingerprints analogous to electronic noses or electronic tongues. Machine learning can distinguish analytes from these multivariate response patterns, compensate for interferents, quantify target concentrations, detect sensor drift, and improve selectivity in complex matrices. The combination of experimental design, chemometrics, and machine learning is therefore becoming integral to electrochemical-sensor development, especially for portable environmental monitoring, point-of-care analysis, and pharmaceutical quality assessment (Puthongkham *et al.,*2021).

Hyperspectral imaging extends spectral analysis into the spatial domain by recording a spectrum at each image pixel. The resulting data cube contains two spatial dimensions and one spectral dimension and may include hundreds of correlated wavelength bands. It can therefore reveal both the chemical identity and spatial distribution of materials. Environmental applications include vegetation-stress assessment, contaminant mapping, waste classification, plastic identification, and water or soil monitoring, whereas pharmaceutical applications include coating analysis, component-distribution mapping, blend-uniformity assessment, counterfeit detection, and solid-state characterization. However, hyperspectral datasets are affected by high dimensionality, spectral redundancy, illumination changes, sensor noise, and spatial heterogeneity. Machine learning and deep learning are used for dimensionality reduction, segmentation, anomaly detection, object classification, spectral–spatial feature learning, and quantitative prediction (Bhargava *et al.,*2024).

Experimental signals describe how a chemical system interacts with an instrument, whereas molecular representations encode the system's composition, topology, geometry, and physicochemical behavior in computational form. The simplest representations are numerical physicochemical descriptors, including molecular weight, lipophilicity, polar surface area, hydrogen-bond counts, formal charge, rotatable-bond count, electronic descriptors, and topological indices. These variables offer interpretability and can be related to properties such as solubility, permeability, retention, adsorption, bioaccumulation, toxicity, and drug release. Their limitations are that they are usually predefined and may not capture all structural features relevant to a specific predictive task.

Molecular fingerprints encode the presence, absence, or frequency of structural fragments and local chemical environments. Examples include molecular access system keys, extended-connectivity fingerprints,

PubChem fingerprints, and pharmacophore fingerprints. Their fixed-length vectors can be used by random forests, support vector machines, gradient-boosting models, and neural networks. Fingerprints are computationally efficient and remain highly competitive for many applications, but their performance may be affected by hash collisions, sparse encoding, and loss of three-dimensional or stereochemical information. Comparative molecular-machine-learning studies have shown that handcrafted descriptors and fingerprints can remain strong baselines and should not be excluded merely because more complex representations are available (Yang *et al.,*2019).

The simplified molecular-input line-entry system, or SMILES, represents a molecular structure as a character sequence. Because SMILES strings resemble linguistic sequences, recurrent neural networks, transformers, and chemical language models can be trained to learn molecular syntax, predict properties, perform reactions, or generate new compounds. SMILES-based modeling is convenient because chemical databases commonly provide this representation, but a single molecule may have multiple valid SMILES strings, and textual proximity does not always correspond to spatial or topological proximity. Models must therefore learn both the grammar of the representation and the underlying chemical meaning. Molecular graphs provide a representation that more directly reflects chemical topology. Atoms are encoded as nodes, bonds as edges, and atom- or bond-level attributes as features. Graph neural networks iteratively propagate information through neighboring atoms, enabling the model to learn task-specific molecular embeddings. These embeddings can support predictions of physical properties, biological activity, environmental fate, ADME characteristics, and toxicity. Graph-based representations avoid some limitations of fixed fingerprints because important structural features can be learned from data rather than specified entirely in advance. Nevertheless, model performance remains dependent on dataset size, chemical diversity, endpoint quality, and appropriate validation against compounds that differ meaningfully from the training structures (Wu *et al.,*2018; Yang *et al.,*2019).

Two-dimensional graphs do not completely represent molecular behavior because conformation, stereochemistry, bond angles, torsions, and intermolecular interactions occur in three-dimensional space. Three-dimensional conformers and atom-coordinate representations are therefore important for modeling geometry-dependent properties, including protein–ligand binding, molecular recognition, crystal packing, reactivity, and optical behavior. Geometry-aware models can encode atomic distances, bond angles, dihedral angles, local coordinate systems, and rotational or translational symmetries. The geometry-enhanced molecular representation framework developed by Fang et al. (2022), for example, incorporated atom–bond and bond-angle relationships into a self-supervised graph-learning architecture, demonstrating how spatial information can improve molecular-property prediction.

Protein structures add another level of complexity. They can be represented through amino-acid sequences, residue-contact maps, molecular surfaces, binding pockets, atomistic graphs, secondary-structure elements, or three-dimensional coordinates. When ligand and protein representations are analyzed jointly, AI models can predict binding sites, ligand poses, affinities, target interactions, and off-target effects. Such systems may combine a molecular graph for the ligand with a residue graph, pocket surface, or learned protein embedding. Their outputs can then be linked with experimental activity, toxicity, or pharmacokinetic data to support structure-based drug discovery and molecular optimization. Learned embeddings provide a more flexible alternative to manually defined molecular features. An embedding is a continuous numerical vector produced by a model after it has learned regularities from large quantities of molecular, spectral, textual, or biological data. Molecules with related structures or properties may occupy neighboring regions of the embedding space even when their original representations differ. Self-supervised models can learn such representations without requiring a labelled endpoint for every molecule. MolE, for example, was pretrained on approximately 842 million molecular graphs and subsequently adapted to multiple absorption, distribution, metabolism, excretion, and toxicity prediction tasks, illustrating the emergence of foundation models for molecular property prediction (Méndez-Lucio *et al.,*2024).

The central function of AI in this convergent framework is to learn correspondences between analytical and molecular representations. These correspondences may operate in several directions. In a signal-to-property model, a spectrum, chromatogram, sensor response, or hyperspectral image is used to predict concentration, identity, quality class, toxicity, or another endpoint. In a structure-to-signal model, molecular information is used to predict infrared bands, NMR chemical shifts, chromatographic retention, or mass-spectral fragmentation. In a signal-to-structure model, experimental measurements are translated into candidate molecular structures. In a structure-to-property model, molecular descriptors, strings, graphs, conformers, or protein complexes are mapped to physicochemical, biological, toxicological, or pharmaceutical outcomes. Multimodal models extend this logic by combining several of these relationships within a shared representation space.

Signal-to-structure modeling is among the most scientifically significant developments because it addresses the inverse problem of recovering molecular identity from experimental measurements. Alberts et al. (2024) demonstrated that a transformer trained using

simulated infrared spectra and adapted with experimental spectra could use the broader infrared fingerprint region to predict molecular structures. This approach moves beyond conventional manual interpretation, in which only selected functional-group bands are assigned, by attempting to exploit distributed information across the entire spectrum. Similarly, mass-spectrometric models such as Spec2Mol infer candidate structures from fragmentation patterns (Litsa *et al.,*2023). The multimodal framework introduced by Mirza et al. (2026) goes further by aligning raw NMR, infrared, and mass-spectrometric embeddings, thereby reproducing the expert practice of integrating orthogonal evidence while automating candidate generation and ranking. Data fusion may be performed at different stages. Early fusion combines preprocessed variables from different sources into a single input matrix, such as concatenating spectral intensities, molecular descriptors, and process conditions. This strategy is straightforward but can be sensitive to incompatible scales, missing modalities, and extreme dimensional imbalance. Intermediate fusion processes each modality through a specialized encoder and combines the resulting latent representations. For example, one encoder may learn a Raman embedding, another a molecular-graph embedding, and a third an experimental-metadata embedding. Late fusion develops separate models and combines their predictions through averaging, stacking, voting, or probabilistic integration. Intermediate fusion is particularly attractive because it allows each data type to retain its own structure before information is shared across modalities.

Multimodal molecular models show how shared embedding spaces can expand the capabilities of predictive systems. MoleculeSTM jointly learned molecular structures and natural-language descriptions through contrastive learning, enabling structure–text retrieval and text-guided molecular editing (Liu *et al.,*2023). Although structure–text integration differs from instrumental data fusion, it demonstrates the broader principle that molecular graphs can be aligned with another information domain. The same principle can be applied to align molecular structures with spectra, chromatographic profiles, toxicological labels, protein targets, formulation compositions, and environmental metadata. For chemical identification, AI can compare unknown analytical patterns with experimental libraries, predicted spectra, molecular embeddings, or database-derived candidate structures. Combining accurate mass, isotopic distribution, retention behavior, fragmentation, vibrational spectra, and NMR constraints can reduce the number of plausible candidates more effectively than any single data source. In environmental non-target analysis, this integration may help prioritize previously unrecognized contaminants, transformation products, and industrial chemicals. In pharmaceutical analysis, it can support impurity profiling, degradation-product identification, counterfeit detection, and raw-material verification.

For quantification, full-spectrum and multichannel models can estimate analyte concentrations even when signals overlap or vary with matrix composition. Inputs may include absorbance values, Raman shifts, fluorescence intensities, electrochemical responses, chromatographic peak areas, and environmental or process variables. Classical PLSR remains effective for many such tasks, whereas nonlinear machine-learning and deep-learning models can capture matrix effects, interactions, and concentration-dependent signal changes. Quantitative models can also be transferred into real-time monitoring systems, although calibration transfer, sensor drift, and instrument variability must be explicitly controlled.

Classification applications include differentiating pollutants, microbial or biological samples, pharmaceutical polymorphs, raw materials, dosage forms, product batches, adulterated samples, and environmental sources. AI can learn classification boundaries from complete analytical fingerprints rather than requiring a unique marker for every class. This is especially useful for sensor arrays and hyperspectral datasets, where individual variables may have limited selectivity but the combined pattern is highly informative.

Toxicity prediction connects molecular representations with biological and toxicological endpoints. Descriptors, fingerprints, SMILES strings, graphs, and learned embeddings can be trained against acute toxicity, hepatotoxicity, cardiotoxicity, mutagenicity, genotoxicity, endocrine activity, and Tox21 assay outcomes. The selection of molecular representation and validation strategy is strongly endpoint-dependent, and no single algorithm performs best across all toxicological tasks (Cavasotto & Scardino, 2022). Analytical measurements can strengthen these models by providing exposure concentrations, metabolite profiles, transformation products, bioavailability indicators, or omics responses. Thus, an integrated environmental-toxicity workflow might combine contaminant spectra, HRMS annotations, molecular graphs, physicochemical properties, and biological-assay data to move from chemical detection to hazard prioritization.

Formulation optimization similarly benefits from combining molecular and experimental variables. Drug descriptors can be integrated with polymer or excipient properties, composition ratios, processing parameters, particle characteristics, dissolution profiles, and spectroscopic measurements. Bannigan et al. (2023) demonstrated this approach for polymeric long-acting injectables by combining physicochemical descriptors of drugs and polymers with formulation and experimental variables to predict drug-release behavior and guide the selection of new drug–polymer combinations. In broader pharmaceutical development, comparable models can support tablet composition, solubility enhancement,

nanoparticle design, encapsulation efficiency, stability, dissolution, and release kinetics.

Molecular design represents the most generative outcome of the integrated framework. Rather than predicting only whether an existing compound possesses a desired property, generative models can propose new molecules conditioned on activity, toxicity, solubility, spectral compatibility, or target-binding constraints. Experimental analytical data can serve as validation evidence, optimization feedback, or even a conditioning modality. For instance, a model could propose structures that are predicted to bind a target, remain below a defined toxicity threshold, possess suitable physicochemical properties, and produce analytical signals consistent with an experimentally observed unknown. Multimodal foundation models may therefore connect molecular generation with analytical verification, creating a closed loop between computational design, synthesis, measurement, and model refinement.

Despite its promise, convergence does not automatically guarantee reliable decisions. Analytical signals may contain instrument-specific artifacts, while molecular databases may include duplicate structures, inconsistent protonation states, incorrect stereochemistry, or uncertain biological labels. Multimodal datasets frequently contain missing measurements, unequal numbers of samples across modalities, and substantial differences in scale and noise. Data leakage can occur when spectra from the same sample, related molecular scaffolds, experimental replicates, or measurements from the same batch are divided between training and test sets. Models must therefore be evaluated using external samples, scaffold-aware splits, instrument- or laboratory-based validation, uncertainty estimation, and clearly defined applicability domains.

Overall, AI enables analytical signals and molecular representations to function as interconnected components of a unified chemical-information ecosystem. FTIR, Raman, NIR, UV–visible, fluorescence, NMR, LC–MS, GC–MS, HRMS, electrochemical, and hyperspectral platforms provide experimentally grounded views of composition, structure, concentration, and spatial distribution. Descriptors, fingerprints, SMILES strings, molecular graphs, three-dimensional conformers, protein structures, and learned embeddings provide complementary computational descriptions of molecular identity and behavior. Their convergence allows models to move beyond isolated classification or regression tasks toward integrated systems that identify substances, quantify analytes, predict toxicological and biological effects, optimize pharmaceutical formulations, and design new molecules. The long-term impact of this approach will depend not only on increasingly powerful architectures but also on rigorous data curation, chemically meaningful representation, transparent model interpretation, external validation, and the preservation of a clear link between computational outputs and experimental evidence.

Artificial intelligence is increasingly being incorporated into analytical chemistry, molecular modeling, environmental monitoring, and pharmaceutical research; however, its applications are frequently discussed within separate disciplinary boundaries. Reviews of AI-assisted spectral interpretation commonly emphasize automated processing of infrared, Raman, nuclear magnetic resonance, ultraviolet–visible, and mass-spectrometric data, whereas molecular machine-learning studies focus primarily on chemical descriptors, fingerprints, molecular graphs, biological activities, and structure–property relationships (Beck *et al.,*2024; Xue *et al.,*2023). Environmental studies often examine contaminant detection, non-target screening, chemical fingerprinting, source attribution, and toxicity prioritization, while pharmaceutical studies emphasize drug discovery, formulation development, manufacturing, quality control, and regulatory decision-making (Arturi & Hollender, 2023; Huanbutta *et al.,*2024). Although these domains employ many of the same data structures, algorithms, validation principles, and predictive objectives, they are rarely evaluated within a unified analytical and molecular framework.

Accordingly, this review focuses on AI models that use chemically or biologically meaningful data to generate predictions relevant to environmental and pharmaceutical sciences. The term AI is used broadly to include classical machine learning, deep learning, graph-based learning, transformer architectures, self-supervised learning, multimodal learning, generative modeling, and related data-driven approaches. Conventional chemometric and statistical methods are also considered where they provide essential methodological foundations, benchmarks, preprocessing strategies, or interpretable alternatives to more complex AI systems. This inclusion is important because model sophistication alone does not guarantee superior scientific performance; predictive success depends on the suitability of the representation, quality and quantity of the data, validation strategy, and compatibility between the model and the intended application (Wu *et al.,*2018; Yang *et al.,*2019).

The analytical data considered in this review include spectroscopic, chromatographic, mass-spectrometric, electrochemical, sensor-derived, imaging, and process-monitoring measurements. These encompass, but are not limited to, FTIR, Raman, near-infrared, UV–visible, fluorescence, NMR, LC–MS, GC–MS, high-resolution mass spectrometry, electrochemical-sensor responses, hyperspectral imaging, dissolution profiles, and pharmaceutical process analytical technology data. Such measurements

may be used as raw signals, extracted features, latent variables, peak tables, spectral embeddings, or fused multimodal representations. AI-assisted interpretation of these data can support signal correction, peak detection, spectral deconvolution, sample classification, analyte quantification, unknown-compound identification, structural annotation, and process-state prediction (Beck *et al.,*2024; Xue *et al.,*2023).

The molecular information included in the review comprises physicochemical descriptors, structural keys, molecular fingerprints, SMILES strings, molecular graphs, three-dimensional conformations, protein sequences and structures, molecular-interaction features, biological assay results, omics profiles, pharmacokinetic parameters, and toxicological endpoints. These representations provide complementary descriptions of chemical identity, topology, geometry, biological function, environmental behavior, and pharmaceutical performance. Their integration with analytical signals creates a continuum between experimental measurement and computational prediction, enabling models to connect what is physically detected with what is structurally, biologically, or functionally inferred. Molecular representation and dataset construction are therefore treated as core components of model development rather than as preliminary technical details (Vamathevan *et al.,*2019; Yang *et al.,*2019).

Within the environmental domain, the review covers AI applications involving the detection, identification, quantification, classification, and prioritization of environmental contaminants. Relevant chemical groups include pesticides, pharmaceuticals and personal-care products, per- and polyfluoroalkyl substances, endocrine-disrupting chemicals, industrial compounds, persistent organic pollutants, heavy metals, microplastics, nanoplastics, metabolites, and environmental transformation products. Particular attention is given to machine-learning-assisted suspect and non-target screening because high-resolution analytical platforms can detect thousands of features that cannot be efficiently interpreted through manual methods alone. Machine learning can assist with dimensionality reduction, feature prioritization, chemical fingerprinting, toxicity classification, and pollution-source attribution, thereby converting complex analytical datasets into information that may support environmental surveillance and risk management (Arturi & Hollender, 2023; Dávila-Santiago *et al.,*2022).

The environmental scope further includes models for predicting chemical fate and effects, including persistence, degradation, adsorption, partitioning, mobility, bioaccumulation, ecotoxicity, endocrine activity, and mixture-related hazards. Analytical measurements may be combined with molecular descriptors, structural fingerprints, experimental bioassays, and environmental metadata to move from simple contaminant detection toward integrated exposure and hazard assessment. This transition is scientifically important because detecting an unknown or emerging contaminant does not by itself establish its environmental significance. Data-driven models must therefore help determine which features are most likely to represent persistent, bioaccumulative, toxic, or source-specific substances while clearly communicating uncertainty and identification confidence (Arturi & Hollender, 2023).

Within the pharmaceutical domain, the review encompasses AI applications across drug discovery, analytical method development, formulation, manufacturing, quality assurance, and post-market monitoring. At the molecular level, relevant applications include target identification, virtual screening, quantitative structure–activity relationships, binding-affinity estimation, ADME and toxicity prediction, drug repurposing, lead optimization, and generative molecular design. At the analytical and product levels, AI can support chromatographic method optimization, raw-material identification, impurity profiling, polymorph classification, dissolution prediction, stability assessment, formulation design, process monitoring, batch classification, and real-time quality control (Huanbutta *et al.,*2024; Vamathevan *et al.,*2019).

Formulation development is included because it illustrates how molecular, analytical, and process variables can be integrated within a single predictive system. Drug and excipient descriptors may be combined with formulation composition, manufacturing conditions, particle properties, spectroscopic measurements, and release profiles to predict product behavior. For example, data-driven models have been used to relate drug and polymer properties to release patterns from long-acting injectable systems, demonstrating how AI can guide the selection of promising formulation combinations before extensive experimental testing (Bannigan *et al.,*2023). Such cases move the field beyond retrospective data classification toward experimentally actionable model-guided design.

The first major objective of this review is to compare the principal AI methodologies used across analytical, environmental, and pharmaceutical research. This comparison includes conventional regression and classification algorithms, ensemble methods, artificial neural networks, convolutional networks, recurrent models, graph neural networks, transformers, autoencoders, foundation models, multimodal architectures, and generative AI. Rather than ranking methods solely according to reported accuracy, the review evaluates their suitability for different data types, sample sizes, prediction endpoints, and deployment settings. The analysis also considers computational requirements, interpretability, dependence on labeled data, transferability, and the capacity to incorporate chemical or mechanistic knowledge.

The second objective is to evaluate complete data-modeling workflows, including sample and data acquisition, preprocessing, feature extraction, molecular standardization, representation selection, model training, hyperparameter optimization, validation, interpretation, and deployment. The review emphasizes that AI performance cannot be separated from the quality of the workflow that produces it. Spectral normalization, chromatographic alignment, missing-value treatment, batch correction, molecular protonation, stereochemical representation, class balance, and data-splitting procedures can substantially influence model outcomes. Consequently, models reporting high internal accuracy without independent testing, leakage control, or chemically meaningful data division are not considered equivalent to systems validated across laboratories, instruments, batches, time periods, or previously unseen molecular scaffolds (Wu *et al.,*2018; Yang *et al.,*2019).

The third objective is to examine practical case studies that connect computational performance with environmental or pharmaceutical decisions. Environmental examples include contaminant-source identification, non-target feature prioritization, PFAS fingerprinting, pollutant classification, and hazard screening. Pharmaceutical examples include drug-candidate prioritization, release-profile prediction, impurity recognition, manufacturing-process control, and product-quality assessment. Priority is given to studies demonstrating external validation, prospective testing, experimental confirmation, or integration into operational analytical workflows. This approach distinguishes proof-of-concept models from systems with demonstrated translational or decision-support value.

The fourth objective is to identify the principal validation, interpretability, reproducibility, and regulatory challenges that limit real-world adoption. Black-box models may produce accurate predictions while offering little explanation of the spectral regions, molecular fragments, biological features, or process variables responsible for those outputs. This lack of transparency can reduce scientific trust and complicate the assessment of whether a prediction reflects a meaningful chemical relationship or an unintended statistical correlation. Explainable AI methods can partially address this problem, but explanations must themselves be evaluated for stability, chemical plausibility, and consistency with experimental evidence (Jiménez-Luna *et al.,*2020).

Uncertainty assessment is similarly essential. Models used for contaminant prioritization, toxicity prediction, drug selection, or pharmaceutical quality decisions must indicate when predictions are unreliable or fall outside the domain represented by the training data. Existing uncertainty-estimation methods do not perform uniformly across datasets and model architectures, which means that uncertainty values should not be accepted without calibration and empirical evaluation (Hirschfeld *et al.,*2020). The review therefore examines prediction intervals, ensemble uncertainty, confidence calibration, applicability-domain analysis, out-of-distribution detection, and prospective error monitoring as necessary components of trustworthy AI.

Regulatory readiness is evaluated in relation to model purpose, risk, traceability, reproducibility, data provenance, human oversight, and lifecycle management. Pharmaceutical AI raises particular concerns because model outputs may influence decisions involving product quality, safety, efficacy, manufacturing, or candidate advancement. Published assessments indicate that AI can improve information management and decision-making throughout drug development, but regulatory guidance and real-world evidence remain less mature than algorithmic development (Nene *et al.,*2024). Comparable issues arise in environmental applications when model predictions are used to prioritize chemicals, identify pollution sources, or support regulatory monitoring without sufficient confirmatory evidence.

The fifth objective is to propose research priorities that can advance the field from isolated demonstrations toward integrated, reproducible, and decision-relevant systems. These priorities include the creation of curated multimodal datasets, development of shared benchmarks, improved reporting standards, external and interlaboratory validation, model calibration, interpretable molecular and spectral learning, and integration of analytical measurements with biological and toxicological evidence. Future opportunities also include chemical foundation models, autonomous laboratories, digital twins, physics-informed AI, active learning, federated learning, and closed-loop systems in which computational predictions guide experiments and newly generated measurements update the model.

The principal novelty of this review lies in its treatment of environmental and pharmaceutical applications as interconnected rather than independent fields. Both domains require the analysis of complex chemical mixtures, interpretation of high-dimensional instrumental signals, representation of molecular structure, prediction of biological effects, and translation of computational outputs into practical decisions. Environmental monitoring seeks to determine what chemicals are present, where they originate, how they behave, and whether they present ecological or human-health hazards. Pharmaceutical research asks related questions concerning molecular identity, product composition, biological activity, toxicity, stability, manufacturing behavior, and therapeutic suitability. Their shared methodological foundation creates opportunities for cross-domain learning: approaches developed for pharmaceutical impurity identification may inform environmental unknown screening, while

environmental mixture analysis and hazard prioritization may strengthen pharmaceutical safety assessment.

This integrated perspective also enables a more critical comparison of how the same AI methodology behaves under different data and regulatory conditions. A graph neural network used for drug-toxicity prediction and a model used for environmental hazard prioritization may employ similar molecular representations but differ substantially in endpoint certainty, chemical-space coverage, acceptable error, and evidentiary requirements. Likewise, a spectral classifier used for pharmaceutical quality control may operate under standardized manufacturing conditions, whereas an environmental sensor must contend with variable matrices, weather, geography, and contaminant mixtures. Examining these contrasts can reveal which methodological practices are transferable and which require application-specific adaptation.

To maintain conceptual coherence, the review excludes AI applications that do not directly involve chemical, analytical, molecular, biological, toxicological, environmental-composition, or pharmaceutical-product data. General medical imaging applications—such as radiological image diagnosis—are excluded unless imaging is explicitly connected to molecular characterization, pharmaceutical response, drug delivery, or chemically resolved analysis. Hospital-management systems, appointment scheduling, insurance analytics, administrative automation, and general electronic health-record prediction are outside the review's scope because they do not address analytical or molecular data modeling.

Environmental remote sensing based only on geographic, meteorological, or conventional visual imagery is also excluded when it lacks a chemical, spectroscopic, molecular, or contaminant-related component. However, hyperspectral or chemically informative remote-sensing studies may be included when the measured signal is used to identify, quantify, or classify environmental substances. Similarly, general climate forecasting, land-use classification, and ecosystem-image recognition are not considered unless they are directly integrated with chemical exposure, pollution, toxicology, or molecular measurements.

The review additionally excludes purely commercial descriptions of AI systems that provide insufficient methodological information, studies that report no meaningful validation, and applications in which the term "AI" is used without a clearly defined algorithm or predictive task. Studies focused exclusively on generic text generation, administrative natural-language processing, or nonchemical robotics are not included. This boundary ensures that the evidence base remains centered on models whose inputs, outputs, and scientific interpretation are directly relevant to analytical measurement, molecular behavior, environmental assessment, or pharmaceutical development.

Overall, the scientific significance of this review lies in establishing a unified framework for understanding how AI converts heterogeneous chemical and biological data into environmental and pharmaceutical knowledge. By comparing methods, workflows, case studies, validation practices, and regulatory requirements across both domains, the review seeks to identify not only where AI performs well, but also why certain models fail to generalize or translate into practice. The intended outcome is a critical roadmap for developing AI systems that are analytically robust, chemically meaningful, biologically relevant, interpretable, uncertainty-aware, and suitable for responsible scientific and regulatory decision-making.

METHODOLOGY

Literature Search Strategy and Information Sources

A structured, multidisciplinary literature-search strategy was developed to identify peer-reviewed research on artificial intelligence–driven analytical and molecular data modeling in environmental and pharmaceutical sciences. The search procedure was designed according to the principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement and its literature-search extension, PRISMA-S. These frameworks emphasize transparent reporting of information sources, search dates, complete search strings, database-specific adaptations, applied limits, duplicate removal, and supplementary searching procedures (Page *et al.,*2021; Rethlefsen *et al.,*2021). Because the present topic spans analytical chemistry, environmental science, computational molecular science, pharmaceutical technology, toxicology, and engineering, a single database was considered insufficient to provide comprehensive disciplinary coverage.

The primary evidence window was defined as January 1, 2018, to June 14, 2026. This period was selected to capture the rapid expansion of machine learning, deep learning, graph neural networks, transformers, multimodal learning, generative AI, and molecular foundation models within analytical and molecular sciences. Earlier publications were retained selectively when they provided essential historical or methodological foundations for chemometrics, quantitative structure–activity relationship modeling, quantitative structure–property relationship modeling, applicability-domain assessment, or model-validation principles. These foundational publications were identified mainly through backward citation tracking rather than being included in quantitative assessments of recent publication trends. Separating the contemporary evidence window from historically important literature helps maintain current relevance while preserving the conceptual development of the field (Snyder, 2019).

The principal bibliographic databases were Web of Science Core Collection, Scopus, and PubMed. Web of Science and Scopus were used to retrieve multidisciplinary studies spanning chemistry, environmental science, pharmacology, toxicology, materials science, computer science, and chemical engineering. PubMed was included to strengthen coverage of pharmacology, toxicology, pharmaceutical development, molecular biology, ADME studies, and biologically oriented AI applications. The use of multiple databases was intended to reduce database-specific selection bias because academic search systems differ substantially in journal coverage, subject representation, indexing policies, search functionality, and citation-linking capabilities (Gusenbauer & Haddaway, 2020).

IEEE Xplore was searched for engineering-focused publications involving electrochemical sensors, biosensors, electronic noses and tongues, signal processing, pattern recognition, edge computing, portable analytical devices, hyperspectral systems, and automated process monitoring. This source was particularly relevant because some technically important AI architectures and sensor-processing methods are initially reported in engineering journals or peer-reviewed conference proceedings rather than chemistry-focused publications. Nevertheless, conference papers were considered only when they provided sufficient methodological detail and represented a substantive technical contribution that was not adequately described in a corresponding journal article. Targeted supplementary searches were undertaken through ScienceDirect, SpringerLink, ACS Publications, and Wiley Online Library. These resources were treated as publisher platforms rather than complete substitutes for multidisciplinary bibliographic databases. They were used to locate recently published articles, verify publication metadata, retrieve papers identified through citation searching, and search journals with high relevance to analytical chemistry, environmental monitoring, chemometrics, pharmaceutical sciences, molecular informatics, toxicology, and process engineering. Distinguishing between bibliographic databases and publisher-specific platforms is important because the latter primarily retrieve content hosted by an individual publisher and may therefore provide incomplete coverage of the wider evidence base (Gusenbauer & Haddaway, 2020).

Google Scholar was used as a supplementary discovery and citation-tracking resource rather than as the principal evidence source. It was applied to identify highly cited foundational studies, recently available papers not yet consistently indexed across major databases, related articles, and forward citations to key publications. It was also used for known-item searching when a paper title, author, or DOI had already been identified. However, Google Scholar results were not relied upon as the sole basis for study identification because its ranking procedures, indexing boundaries, result limits, and search reproducibility are less transparent than those of structured bibliographic databases (Gusenbauer & Haddaway, 2020). Records located through Google Scholar were therefore verified through the relevant journal, publisher, PubMed, Crossref-linked record, or DOI registration before inclusion.

The search vocabulary was organized into three main concept blocks: AI methodology, analytical or molecular data, and application domain. The first block included terms related to artificial intelligence and computational modeling, such as “artificial intelligence,” “machine learning,” “deep learning,” “chemometrics,” “neural network,” “convolutional neural network,” “graph neural network,” “transformer,” “foundation model,” “self-supervised learning,” “multimodal learning,” “generative AI,” and “explainable AI.” These terms were selected to capture both established machine-learning methods and emerging architectures used for spectral interpretation, molecular-property prediction, chemical-language modeling, and automated decision support (Vamathevan *et al.,*2019; Xue *et al.,*2023).

The second concept block represented analytical platforms, molecular representations, and modeling tasks. It included “analytical chemistry,” “molecular modeling,” “molecular representation,” “QSAR,” “QSPR,” “spectroscopy,” “FTIR,” “infrared spectroscopy,” “Raman spectroscopy,” “near-infrared spectroscopy,” “UV–visible spectroscopy,” “fluorescence,” “NMR,” “chromatography,” “HPLC,” “LC–MS,” “GC–MS,” “mass spectrometry,” “high-resolution mass spectrometry,” “electrochemical sensor,” “biosensor,” “hyperspectral imaging,” “molecular descriptor,” “molecular fingerprint,” “SMILES,” “molecular graph,” “protein structure,” and “molecular embedding.” This broad terminology was necessary because AI models may operate on raw instrumental signals, extracted analytical features, predefined descriptors, molecular strings, graph representations, three-dimensional structures, or learned latent embeddings (Beck *et al.,*2024; Yang *et al.,*2019).

The third concept block was divided into environmental and pharmaceutical application terms. Environmental searches used expressions such as “environmental monitoring,” “pollutant identification,” “contaminant detection,” “environmental analysis,” “non-target screening,” “suspect screening,” “environmental toxicology,” “ecotoxicity,” “chemical fate,” “bioaccumulation,” “PFAS,” “pesticide,” “microplastic,” and “transformation product.” These terms were intended to capture AI applications involving contaminant identification, chemical fingerprinting, source attribution, hazard prioritization, exposure assessment, environmental fate, and ecological-risk prediction (Arturi & Hollender, 2023).

Pharmaceutical searches included "pharmaceutical analysis," "drug discovery," "drug formulation," "formulation optimization," "quality control," "impurity profiling," "stability prediction," "process analytical technology," "continuous manufacturing," "dissolution," "drug delivery," "pharmacokinetics," "ADMET prediction," "drug–target interaction," and "molecular design." The terminology covered both molecular-level applications, such as virtual screening and toxicity prediction, and product-level applications, such as formulation development, manufacturing-process control, and pharmaceutical quality assurance (Huanbutta *et al.,*2024; Vamathevan *et al.,*2019).

The three concept blocks were combined using Boolean operators. A representative general search structure was:

("artificial intelligence" OR "machine learning" OR "deep learning" OR chemometric OR "graph neural network" OR transformer* OR "foundation model*" OR "generative AI") AND ("analytical chemistry" OR "molecular model*" OR QSAR OR QSPR OR spectroscop* OR chromatograph* OR "mass spectrometr*" OR sensor* OR "molecular representation*") AND (environment* OR pollut* OR contaminant* OR pharmaceutical* OR drug* OR formulation OR ADMET) **

The asterisk was used as a truncation operator where supported to retrieve variations such as spectroscopy, spectroscopic, and spectrometric. Quotation marks were used for exact multiword phrases, while parentheses preserved the intended Boolean logic. Search strategies were initially developed in a structured format and then translated for each database, an approach recommended for improving both search efficiency and reproducibility (Bramer *et al.,*2018). Because requiring every article to contain terms from both the environmental and pharmaceutical domains could exclude relevant evidence, separate domain-specific searches were performed. The environmental search combined the AI and analytical–molecular blocks with environmental terms, whereas the pharmaceutical search combined the same methodological blocks with pharmaceutical terms. A third cross-domain search used terms such as "environmental pharmaceutical," "pharmaceutical contaminant," "drug residue," "environmental ADMET," "molecular toxicity prediction," and "analytical molecular data fusion" to identify studies situated directly at the interface of the two fields. This modular approach increased retrieval sensitivity while preserving the integrated conceptual focus of the review.

Search syntax was adapted to the indexing and field-code conventions of each platform. In Web of Science, topic-field searching was applied to titles, abstracts, author keywords, and Keywords Plus. Scopus searches were structured primarily within the title, abstract, and keyword fields. PubMed searches combined free-text terms with relevant Medical Subject Headings where appropriate, particularly for artificial intelligence, machine learning, environmental monitoring, pharmaceutical preparations, drug discovery, toxicology, and pharmacokinetics. IEEE Xplore searches prioritized article titles, abstracts, indexing terms, and metadata. Database-specific translation was necessary because identical search strings may behave differently across platforms owing to variations in controlled vocabulary, phrase searching, proximity operators, truncation rules, and automatic term mapping (Bramer *et al.,*2018; Rethlefsen *et al.,*2021).

No geographical restriction was applied because advances in environmental monitoring, pharmaceutical modeling, and analytical AI are internationally distributed. Language filters were not applied during the initial database search to avoid prematurely removing potentially relevant evidence. During screening, priority was given to publications for which the title, abstract, and full methodological content could be reliably assessed in English. Searches primarily targeted peer-reviewed original articles and high-quality reviews; relevant perspectives, benchmark studies, and methodological papers were retained when they contributed directly to data representation, model validation, analytical interpretation, or regulatory evaluation. The electronic search was supplemented by backward and forward citation tracking. Reference lists of key reviews, benchmark papers, and highly relevant original studies were examined to identify publications not retrieved through the initial keyword combinations. Forward citation searching was performed through Web of Science, Scopus, and Google Scholar to locate later studies that extended, validated, criticized, or applied influential models. Citation tracking is particularly important in rapidly evolving interdisciplinary fields because terminology may change over time and important studies may not use the same keywords as more recent publications (Rethlefsen *et al.,*2021).

Search alerts and targeted update searches were also considered necessary because AI research develops quickly and articles may appear online before assignment to a journal issue. The final search should therefore be rerun immediately before manuscript submission, using the original search strings and an updated publication-date limit. Any newly identified studies should undergo the same eligibility assessment as records found in the original search. This procedure reduces the risk that the review will omit important studies published between the initial search and final submission (Page *et al.,*2021).

All retrieved records should be exported in a standardized bibliographic format, such as RIS or BibTeX, and consolidated in a reference-management system. Duplicate records should be identified using DOI, PubMed identifier, title, author, year, and journal

metadata, followed by manual inspection of uncertain matches. The search log should preserve the name of each database and platform, the date searched, the complete search expression, applied filters, number of retrieved records, and number of duplicates removed. Maintaining this audit trail is essential for reproducibility and is explicitly encouraged by PRISMA-S (Rethlefsen *et al.,*2021). The study-identification process should be summarized through a PRISMA flow diagram showing the numbers of records identified, duplicates removed, titles and abstracts screened, full texts assessed, studies excluded with reasons, and publications retained for synthesis. Importantly, the flow diagram and record counts must be completed using the actual outputs of the database searches rather than estimated values. This structured approach will ensure that the review provides a transparent, reproducible, and sufficiently comprehensive account of the evidence supporting AI-driven analytical and molecular data modeling in environmental and pharmaceutical applications (Page *et al.,*2021; Snyder, 2019).

Eligibility Criteria, Screening, and Study Classification

Eligibility criteria were predefined to ensure that the evidence included in this review was scientifically relevant to AI-driven analytical and molecular data modeling and sufficiently detailed for critical methodological evaluation. The selection process was designed to capture research at the intersection of artificial intelligence, analytical chemistry, molecular modeling, environmental assessment, and pharmaceutical science. Study identification, screening, eligibility assessment, and documentation of exclusion reasons were organized in accordance with the transparent reporting principles of PRISMA 2020 and PRISMA-S (Page *et al.,*2021; Rethlefsen *et al.,*2021).

Studies were considered eligible when they examined an explicitly defined artificial intelligence, machine-learning, deep-learning, chemometric, or hybrid computational approach applied to chemical, analytical, molecular, biological, environmental, toxicological, or pharmaceutical data. Eligible AI approaches included supervised and unsupervised machine learning, ensemble learning, artificial neural networks, convolutional and recurrent neural networks, graph neural networks, transformers, autoencoders, self-supervised learning, transfer learning, multimodal learning, explainable AI, generative modeling, and molecular foundation models. Conventional chemometric techniques were also eligible when they served as direct comparators, preprocessing components, feature-extraction methods, or components of an integrated AI workflow (Xue *et al.,*2023). The principal publication types eligible for inclusion were peer-reviewed original research articles reporting model development, validation, application, or comparison. Original studies were required to provide sufficient information about the input data, analytical or molecular representations, model architecture, training procedure, evaluation method, and target outcome. Studies reporting only that an AI tool was used, without describing the algorithm, data, prediction task, or validation procedure, were not considered suitable for methodological synthesis (Beck *et al.,*2024; Yang *et al.,*2019).

Systematic reviews, scoping reviews, and high-quality critical reviews were included when they synthesized evidence directly relevant to the review objectives, identified methodological trends, summarized validation practices, or provided a structured overview of environmental or pharmaceutical applications. These secondary sources were used to establish context, identify additional primary studies, and compare conclusions across subfields. However, review articles were not treated as independent performance evidence when their underlying primary studies were already included, thereby reducing the risk of double counting the same model results. Benchmark studies were eligible when they provided standardized datasets, molecular representations, predefined data partitions, evaluation metrics, open-source implementations, or direct comparisons among several algorithms. Such studies are especially important in molecular machine learning because model performance can vary substantially according to dataset composition, featurization, splitting strategy, endpoint type, and class imbalance. MoleculeNet, for example, demonstrated the importance of standardized datasets, task-appropriate metrics, and chemically meaningful data partitioning when comparing molecular-learning algorithms (Wu *et al.,*2018).

Validated computational frameworks, software pipelines, and open-source platforms were also eligible when they presented a clearly defined scientific workflow and included empirical evaluation on relevant analytical or molecular datasets. Eligible frameworks could address spectral preprocessing, chromatographic alignment, mass-spectral annotation, molecular featurization, model training, candidate ranking, toxicity prediction, formulation design, or uncertainty estimation. Software descriptions without performance evaluation, reproducible methods, or demonstration on chemically relevant data were excluded. Authoritative regulatory, governmental, and intergovernmental documents were eligible when they addressed the validation, reporting, governance, credibility, or regulatory use of AI, QSAR, analytical models, pharmaceutical manufacturing models, toxicity predictions, or environmental risk-assessment tools. This category included formal guidance documents, reflection papers, technical reports, and consensus frameworks issued by recognized regulatory agencies or international scientific organizations. These documents were evaluated separately from peer-reviewed studies because their purpose is to define evidentiary or governance expectations rather than report experimental model performance.

Peer-reviewed conference proceedings were considered only when they presented complete technical methods, results, and validation information not available in a corresponding journal publication. Conference abstracts, posters, presentation summaries, editorials, news items, commercial webpages, and promotional descriptions were excluded because they generally lacked sufficient methodological detail for independent appraisal. Preprints were not included in the primary evidence synthesis unless a peer-reviewed version became available before completion of the final search. Studies were eligible when AI was applied to analytical-signal processing or interpretation. Relevant tasks included noise reduction, baseline correction, peak detection, signal alignment, spectral deconvolution, feature extraction, dimensionality reduction, anomaly detection, calibration transfer, pattern recognition, and quantitative analysis. Eligible analytical data could originate from FTIR, Raman, near-infrared, UV–visible, fluorescence, NMR, chromatography, mass spectrometry, electrochemical sensing, hyperspectral imaging, process analytical technology, or related chemically informative platforms (Beck *et al.,*2024; Xue *et al.,*2023).

Chemical-identification studies were included when AI was used to classify known substances, prioritize candidate structures, match spectra, predict molecular formulae, model fragmentation, estimate chromatographic retention, annotate metabolites or degradation products, or elucidate unknown compounds. Studies integrating several analytical modalities, such as chromatography, high-resolution mass spectrometry, infrared spectroscopy, or NMR, were eligible when their objective was to improve chemical identification or structural confidence.

Molecular-property prediction studies were included when they used descriptors, fingerprints, SMILES strings, molecular graphs, three-dimensional conformers, protein representations, or learned embeddings to predict a clearly defined chemical, physical, biological, environmental, or pharmaceutical endpoint. Relevant endpoints included solubility, lipophilicity, partitioning, permeability, binding affinity, bioactivity, degradation, persistence, adsorption, bioaccumulation, metabolic behavior, and pharmacokinetic properties. Because different molecular representations and splitting strategies can produce markedly different performance estimates, studies were required to describe both the molecular inputs and the procedure used to divide data into training and evaluation sets (Wu *et al.,*2018; Yang *et al.,*2019).

Toxicity-modeling studies were eligible when they predicted defined endpoints such as acute toxicity, cytotoxicity, genotoxicity, mutagenicity, hepatotoxicity, cardiotoxicity, developmental toxicity, endocrine activity, ecotoxicity, or adverse biological responses. Studies combining molecular descriptors with bioassay, omics, exposure, or analytical measurements were also included. Toxicity studies were required to identify the biological endpoint, test system, label source, unit of analysis, and evaluation metric because toxicity datasets often differ in assay design, species, exposure duration, class distribution, and endpoint definition (Cavasotto & Scardino, 2022).

Environmental studies were eligible when they used AI for contaminant detection, pollutant identification, non-target or suspect screening, source attribution, concentration estimation, exposure assessment, environmental-fate prediction, ecotoxicity modeling, treatment optimization, or chemical-risk prioritization. Eligible studies could address water, wastewater, soil, sediment, air, food, biota, or environmental mixtures. Research using high-resolution mass spectrometry for data-driven feature prioritization was included when the model contributed to the identification or hazard ranking of environmentally relevant chemical features (Arturi & Hollender, 2023).

Pharmaceutical studies were eligible when AI was applied to raw-material identification, analytical-method development, chromatographic optimization, impurity profiling, degradation-product analysis, polymorph classification, formulation development, dissolution prediction, drug-release modeling, process monitoring, manufacturing control, batch classification, stability prediction, quality assurance, or real-time release testing. Models supporting drug discovery, target identification, virtual screening, ADME prediction, drug–target interaction analysis, lead optimization, repurposing, or generative molecular design were also included (Vamathevan *et al.,*2019).

Formulation and manufacturing studies were required to connect model inputs with an experimentally or operationally relevant product outcome. Eligible inputs could include drug and excipient descriptors, formulation composition, particle characteristics, processing variables, analytical measurements, dissolution profiles, or stability data. For example, studies linking molecular and formulation variables to experimental drug-release behavior were considered practically relevant because their predictions could guide the selection of new formulation combinations rather than merely classify historical data (Bannigan *et al.,*2023).

Studies were excluded when their research questions lacked direct chemical, analytical, molecular, biological, environmental, toxicological, or pharmaceutical relevance. General AI applications in hospital administration, appointment scheduling, insurance processing, electronic health-record management, or healthcare-resource allocation were outside the scope of this review. General medical-imaging studies were excluded unless the imaging data were explicitly integrated with molecular, chemical,

pharmaceutical, toxicological, or treatment-response information.

Environmental remote-sensing studies based solely on geographic, meteorological, visual, or land-use data were excluded when they did not contain a chemical, spectroscopic, contaminant, molecular, or toxicological component. Hyperspectral studies remained eligible when spectral signatures were used to detect, quantify, or classify environmentally or pharmaceutically relevant materials. General climate prediction, ecological image recognition, and habitat mapping without chemical-data integration were not included. Studies were excluded if they failed to define the model input, output, target endpoint, dataset source, or evaluation procedure. Publications reporting only a final accuracy value without specifying the train–validation–test structure, cross-validation method, sample numbers, class distribution, or performance metric were considered insufficient for critical comparison. Studies were also excluded when the reported model could not be distinguished from a rule-based calculation, conventional database search, or unspecified commercial software system.

Models evaluated exclusively on their training data were excluded from the main comparative synthesis because training performance does not provide evidence of generalization. Studies without a held-out test set, cross-validation procedure, temporal evaluation, scaffold-based split, independent dataset, or other defensible validation approach were retained only when they provided exceptional methodological or historical relevance and were clearly identified as preliminary. Molecular benchmarks have shown that random splitting can overestimate performance when structurally related compounds occur in both training and test sets; therefore, the chemical meaning of the validation design was treated as a central eligibility consideration (Wu *et al.,*2018; Yang *et al.,*2019). Studies were excluded when substantial data leakage was evident and could not be separated from the reported results. Potential leakage mechanisms included preprocessing before data partitioning, feature selection using the complete dataset, distribution of replicate spectra across training and test subsets, inclusion of measurements from the same sample in several partitions, or splitting closely related molecular structures without controlling scaffold similarity. Studies with unclear leakage risk could be included but were flagged during quality assessment and interpreted cautiously.

Publications lacking adequate methodological detail were excluded when missing information prevented assessment of the model's reproducibility or validity. Examples included failure to describe signal preprocessing, molecular standardization, missing-data management, feature engineering, hyperparameter selection, software implementation, or evaluation metrics. Lack of publicly accessible data or code did not automatically result in exclusion, but it was recorded as a reproducibility limitation. Studies were excluded when outcomes were poorly defined, biologically ambiguous, or disconnected from the stated application. For example, a model predicting an unspecified "toxicity score" without identifying the assay or biological endpoint was not considered equivalent to a model predicting a defined toxicological outcome. Similarly, pharmaceutical models were required to link predictions to measurable attributes such as concentration, dissolution, stability, release, purity, biological activity, or manufacturing performance.

Duplicate publications describing the same dataset and model were consolidated. When several articles reported overlapping work, the most complete and recent peer-reviewed version was treated as the primary record, while earlier reports were used only to clarify model development or historical progression. Multiple models reported within a single article were extracted separately when they differed materially in algorithm, representation, outcome, or validation design. All retrieved records were to be imported into reference-management or systematic-review software and deduplicated using DOI, title, author, year, journal, and database identifiers. Automated duplicate detection was to be followed by manual review because differences in punctuation, abbreviated journal names, online-first dates, and conference-to-journal publication histories can prevent exact matching. The number of records removed at the deduplication stage should be documented for inclusion in the PRISMA flow diagram (Page *et al.,*2021; Rethlefsen *et al.,*2021).

Screening was planned in two sequential stages. During the first stage, titles and abstracts would be independently assessed against the predefined scope and eligibility criteria. Records would be retained for full-text evaluation whenever relevance could not be confidently determined from the title and abstract alone. A deliberately sensitive approach would be used at this stage to minimize the premature exclusion of interdisciplinary papers that used unfamiliar terminology. During the second stage, the complete texts of potentially eligible publications would be evaluated. Full-text assessment would examine publication type, data source, analytical or molecular relevance, AI method, outcome definition, methodological completeness, validation design, and application domain. Each excluded full-text article would be assigned a single principal exclusion reason, such as insufficient chemical relevance, absence of validation, unclear outcome, inadequate methods, duplicate publication, ineligible article type, or unavailable full text. Ideally, title–abstract and full-text screening should be conducted independently by at least two reviewers. Disagreements should be resolved through discussion, with a third reviewer consulted when consensus cannot be reached. Before formal screening, reviewers should pilot the criteria on a shared sample of records to identify

ambiguous terms and improve consistency. Any modification of the eligibility criteria after screening begins should be documented, justified, and applied consistently to all previously assessed records.

The final selection process should be presented in a PRISMA flow diagram reporting the numbers of records identified, duplicates removed, records screened, full texts assessed, articles excluded with reasons, and studies included in the qualitative or quantitative synthesis. No numerical values should be inserted until the searches and screening stages have actually been completed (Page *et al.,*2021). Each eligible study was to be classified using a predefined multidimensional framework rather than assigned to only one broad subject category. This approach allows studies using similar algorithms to be compared across environmental and pharmaceutical applications and enables differences in data type, representation, validation, and practical maturity to be examined systematically.

First, studies were classified according to data type. Categories included raw spectra, processed spectra, chromatographic traces, peak tables, mass-spectral data, sensor signals, hyperspectral images, molecular structures, physicochemical descriptors, molecular fingerprints, biological assays, omics data, protein data, formulation variables, process data, environmental metadata, and multimodal combinations. Studies using more than one data category were coded as multimodal and the individual modalities were recorded separately.

Second, studies were classified by analytical platform. Categories included FTIR, Raman, near-infrared, UV–visible, fluorescence, NMR, HPLC, LC–MS, GC–MS, HRMS, electrochemical sensors, biosensors, electronic noses or tongues, hyperspectral imaging, process analytical technology, and combinations of platforms. The instrument type, manufacturer or model where reported, sampling mode, laboratory or field setting, and number of independent instruments were also recorded because instrument-specific models may show limited transferability (Beck *et al.,*2024; Xue *et al.,*2023).

Third, studies were classified according to molecular representation. Categories comprised physicochemical descriptors, topological descriptors, structural keys, circular or extended-connectivity fingerprints, pharmacophore fingerprints, SMILES strings, molecular graphs, three-dimensional conformers, protein sequences, protein structures, molecular surfaces, docking-derived features, learned embeddings, and hybrid representations. The review also recorded whether representations were predefined, learned from labeled data, pretrained through self-supervised learning, or derived from a foundation model (Yang *et al.,*2019).

Fourth, studies were grouped by AI algorithm. The principal categories included linear or latent-variable chemometrics, support vector machines, nearest-neighbour methods, decision trees, random forests, gradient boosting, Gaussian processes, shallow neural networks, convolutional networks, recurrent networks, autoencoders, graph neural networks, transformers, ensemble models, multimodal networks, generative models, and hybrid physics-informed or knowledge-guided systems. Studies comparing several algorithms were retained as comparative studies, and the model-selection procedure was recorded.

Fifth, each study was classified by application domain and prediction task. Environmental categories included contaminant detection, chemical identification, concentration prediction, non-target screening, source attribution, fate modeling, ecotoxicity, exposure assessment, treatment optimization, and risk prioritization. Pharmaceutical categories included analytical quality control, impurity identification, formulation optimization, dissolution or release prediction, process monitoring, manufacturing, ADME or toxicity prediction, drug–target interaction, virtual screening, and molecular generation. Cross-domain studies were coded separately when they addressed pharmaceutical contaminants in the environment or used common analytical–molecular workflows in both fields.

Sixth, studies were categorized according to validation approach. Validation classes included resubstitution or training-only evaluation, internal hold-out testing, k-fold cross-validation, repeated or nested cross-validation, scaffold-based splitting, temporal splitting, batch-based splitting, instrument-based splitting, laboratory- or site-based testing, independent external datasets, prospective validation, and experimental confirmation. The presence of hyperparameter tuning, class balancing, uncertainty estimation, calibration assessment, applicability-domain analysis, and comparison with conventional or expert methods was also recorded (Hirschfeld *et al.,*2020; Wu *et al.,*2018).

Seventh, studies were classified by level of practical implementation using a review-specific five-level scale. Level 0 represented conceptual, simulation-based, or method-development studies without application to real experimental samples. Level 1 included retrospective proof-of-concept studies evaluated on a single dataset, instrument, laboratory, or institution. Level 2 required independent external validation across a new dataset, chemical series, instrument, laboratory, site, geographical region, or manufacturing batch.

Level 3 represented prospective pilot implementation in which the AI system was tested within an active laboratory, field, manufacturing, or drug-development workflow, with outputs evaluated against

confirmatory experiments or expert decisions. Level 4 represented routine or regulation-relevant implementation, including sustained operational deployment, continuous performance monitoring, defined human oversight, documented model updating, and evidence that the model influenced real environmental or pharmaceutical decisions. This classification distinguished high-performing retrospective models from systems that demonstrated genuine operational utility.

Finally, studies were classified according to reporting completeness and reproducibility. Recorded elements included availability of raw or processed data, access to code, software and package versions, hyperparameters, preprocessing details, random seeds, molecular-standardization procedures, train–test identifiers, external-validation data, and sufficient information to reproduce the workflow. Model performance was interpreted together with these features because a marginally lower-performing transparent model may offer greater scientific and practical value than a poorly documented model reporting very high internal accuracy.

This multidimensional classification framework was designed to support comparisons across heterogeneous studies without assuming that performance metrics obtained from different datasets, endpoints, or validation designs are directly interchangeable. The resulting evidence matrix will enable the review to identify which combinations of data, representation, AI methodology, validation strategy, and implementation level are most strongly supported in environmental and pharmaceutical applications.

Data Extraction, Quality Assessment, and Evidence Synthesis

A standardized and pilot-tested extraction form will be used to collect information consistently from all eligible studies. Data extraction will preferably be conducted by two reviewers, with disagreements resolved through discussion or consultation with a third reviewer. The extraction database will be maintained in a structured, machine-readable format to support verification, tabulation, visualization, and future reuse of the evidence (Haddaway *et al.,*2021; Page *et al.,*2021).

For each study, bibliographic information, publication type, application domain, study objective, dataset source, dataset size, number of independent samples or molecules, class distribution, and experimental setting will be recorded. Where applicable, details of sample collection, preparation, analytical standards, instrument type, acquisition conditions, biological assay, environmental matrix, pharmaceutical formulation, manufacturing batch, and reference-label generation will also be extracted. These factors are important because model performance may reflect dataset composition and experimental design rather than the inherent superiority of an AI algorithm (Deng *et al.,*2023).

Information on data preprocessing will include missing-value treatment, baseline correction, denoising, normalization, scaling, peak alignment, spectral smoothing, batch correction, outlier handling, molecular standardization, and duplicate removal. Feature-engineering procedures—including peak extraction, dimensionality reduction, variable selection, physicochemical descriptors, molecular fingerprints, SMILES tokenization, molecular graphs, three-dimensional conformers, and learned embeddings—will be documented. Representation type must be considered alongside model architecture because the same algorithm may perform differently when trained using descriptors, fingerprints, strings, or graphs (Wu *et al.,*2018; Yang *et al.,*2019).

Model-related extraction will cover algorithm type, architecture, software environment, model parameters, hyperparameter-search method, optimization objective, loss function, regularization, class weighting, resampling, transfer learning, and ensemble construction. The training strategy will be recorded as supervised, unsupervised, self-supervised, semi-supervised, transfer-learning, multimodal, or generative. The use of conventional chemometric or machine-learning models as comparators will also be extracted because complex deep-learning approaches should be evaluated against appropriately tuned baseline methods rather than judged only by their absolute performance (Deng *et al.,*2023).

Validation procedures will be classified as training-only evaluation, hold-out testing, k-fold or nested cross-validation, scaffold splitting, temporal splitting, batch-based splitting, instrument- or laboratory-based validation, independent external testing, or prospective experimental confirmation. The extraction form will record whether preprocessing, feature selection, oversampling, and hyperparameter optimization were confined to the training data. Performing these operations before data partitioning can introduce leakage and produce overly optimistic estimates of generalization (Wu *et al.,*2018).

Performance metrics will be extracted according to task type. Classification studies may report sensitivity, specificity, precision, recall, balanced accuracy, F1-score, Matthews correlation coefficient, area under the receiver-operating-characteristic curve, or area under the precision–recall curve. Regression studies may report mean absolute error, root-mean-square error, coefficient of determination, correlation coefficients, or calibration statistics. Results will be recorded separately for internal and external test sets, together with confidence intervals, standard deviations, repeated-run variability, or statistical comparisons where available (Deng *et al.,*2023). Interpretability will be assessed by

recording the use of feature importance, regression coefficients, loading plots, SHAP values, attention maps, saliency analysis, counterfactual explanations, spectral-region attribution, or molecular-substructure attribution. Uncertainty assessment will include prediction intervals, deep ensembles, Monte Carlo dropout, Bayesian approaches, conformal prediction, calibration analysis, or other confidence-estimation methods. Because uncertainty estimates differ considerably in reliability, studies will receive greater methodological weight when uncertainty is empirically calibrated against unseen data rather than merely generated by the model (Hirschfeld *et al.,*2020).

Applicability-domain assessment will be extracted for molecular and analytical models. Relevant information will include structural similarity, leverage, distance to the training distribution, chemical-space coverage, out-of-distribution detection, activity cliffs, and criteria for identifying unreliable predictions. Models will be considered more scientifically credible when they define the conditions under which predictions are valid and demonstrate declining reliability outside the represented domain (Wang *et al.,*2021). Data and code availability will be documented by recording access to raw or processed datasets, molecular identifiers, sample metadata, preprocessing scripts, model code, trained weights, hyperparameters, software versions, random seeds, and train–test assignments. The absence of public data or code will not automatically exclude a study, but it will be treated as a limitation affecting transparency, verification, and reproducibility. Sufficient reporting is particularly important when different studies claim improvements using nominally identical benchmark datasets but employ different data partitions or preprocessing decisions (Deng *et al.,*2023; Haddaway *et al.,*2021).

Methodological quality will be evaluated through a review-specific domain-based framework. The principal domains will include data provenance and curation, sample size, class balance, chemical or biological diversity, preprocessing transparency, leakage control, validation rigor, external testing, applicability-domain assessment, uncertainty analysis, reproducibility, and comparison with conventional methods. Each domain will be rated as presenting low, moderate, high, or unclear concern, supported by a written justification rather than an unweighted total score. Class imbalance will be examined using the distribution of positive and negative samples and the suitability of reported metrics. Studies relying mainly on overall accuracy for strongly imbalanced datasets will be interpreted cautiously because high accuracy may result from predicting the majority class. Greater confidence will be assigned to studies reporting class-sensitive metrics and applying resampling, class weighting, or threshold optimization within the training data only (Bae *et al.,*2021).

Chemical diversity will be evaluated through scaffold distribution, structural similarity, chemical-space coverage, endpoint range, and representation of relevant environmental or pharmaceutical compounds. Independent validation involving new molecular scaffolds, laboratories, instruments, locations, batches, or time periods will be considered stronger evidence of generalizability than repeated random splitting of a single homogeneous dataset (Deng *et al.,*2023; Wu *et al.,*2018). Evidence will be synthesized narratively because the included studies are expected to differ substantially in analytical platforms, datasets, endpoints, AI algorithms, validation designs, and performance metrics. Studies will be grouped by application domain, data type, molecular representation, algorithm, and implementation level. Comparative tables will summarize dataset characteristics, modeling workflows, validation strategies, principal results, interpretability, uncertainty, and reproducibility features.

Evidence maps will be used to visualize the distribution of studies across environmental and pharmaceutical applications, analytical platforms, molecular representations, algorithms, and levels of practical validation. Application-specific case-study summaries will highlight representative examples of contaminant identification, toxicity prediction, pharmaceutical quality control, formulation optimization, process monitoring, and molecular design. This approach will identify both evidence clusters and under-researched areas while retaining the context necessary to interpret individual studies (Haddaway *et al.,*2021). Reported accuracy will not be used as the sole indicator of model quality. Conclusions will consider whether performance was obtained using representative data, appropriate baselines, leakage-resistant validation, external testing, interpretable outputs, calibrated uncertainty, and a defined applicability domain. Quantitative pooling will be undertaken only when studies are sufficiently comparable in dataset, endpoint, validation design, and reported metrics; otherwise, structured narrative comparison will provide the primary basis for evidence synthesis.

Analytical and Molecular Data Acquisition, Preprocessing, and Representation

Sources and Acquisition of Analytical and Molecular Data

The reliability of an artificial-intelligence model is fundamentally determined by the quality, diversity, and traceability of the data from which it learns. In analytical and molecular sciences, data may originate from laboratory instruments, portable sensors, imaging platforms, molecular databases, biological assays, or toxicological experiments. These sources differ substantially in scale, dimensionality, resolution, noise structure, and biological or chemical meaning. Consequently, data acquisition should preserve not only the primary measurements but also sample-preparation details, instrumental settings, calibration records,

quality-control results, batch identifiers, and other metadata required to distinguish genuine chemical variation from technical variation (Beck *et al.,*2024; Xue *et al.,*2023). Fourier-transform infrared spectroscopy, Raman spectroscopy, and near-infrared spectroscopy are major sources of high-dimensional analytical data. FTIR records wavelength- or wavenumber-dependent infrared absorption associated with molecular vibrations, providing information about functional groups, chemical bonds, and complete molecular fingerprints. The resulting datasets commonly consist of absorbance or transmittance values measured across hundreds or thousands of spectral variables. FTIR data may be acquired through transmission, diffuse reflectance, or attenuated-total-reflectance configurations, each of which produces signals with different penetration depths, matrix sensitivities, and scattering characteristics (Xue *et al.,*2023).

FTIR is widely used for environmental identification of polymers, microplastics, organic contaminants, and biological materials, as well as for pharmaceutical raw-material authentication, polymorph analysis, formulation evaluation, and quality control. Reliable data acquisition requires documentation of the spectral range, resolution, number of scans, background measurement, sample thickness, crystal material, and atmospheric corrections. Water vapor, carbon dioxide, sample contact, and variations in particle size can alter the recorded spectra and may subsequently be learned by an AI model as misleading classification features (Xue *et al.,*2023).

Raman spectroscopy measures the inelastic scattering of incident light and provides vibrational information that is complementary to infrared absorption. Raman datasets typically contain intensity values as a function of Raman shift and may be acquired using conventional Raman instruments, Raman microscopy, surface-enhanced Raman spectroscopy, or spatially resolved mapping. The technique is valuable for aqueous environmental samples, microplastic identification, pharmaceutical polymorphism, counterfeit-drug detection, and chemical imaging because water generally produces relatively weak Raman interference (Xue *et al.,*2023). The quality of Raman data depends on excitation wavelength, laser power, integration time, spectral resolution, optical alignment, substrate characteristics, and the potential occurrence of fluorescence or sample heating. Surface-enhanced Raman spectroscopy may provide extremely sensitive measurements, but signal reproducibility can be affected by nanoparticle morphology, surface chemistry, and the spatial distribution of enhancement sites. These acquisition conditions must be recorded because AI models can otherwise learn laboratory- or substrate-specific patterns instead of analyte-specific molecular information (Xue *et al.,*2023). Near-infrared spectroscopy records overtone and combination vibrations and is commonly used for rapid and nondestructive analysis of powders, tablets, liquids, agricultural materials, and environmental samples. NIR data can be collected through transmission, transflection, diffuse reflectance, or hyperspectral configurations. Although individual NIR bands are broader and less structurally specific than many mid-infrared or Raman bands, the complete multivariate pattern can support prediction of moisture, composition, concentration, blend uniformity, particle characteristics, and pharmaceutical quality attributes (Xue *et al.,*2023).

UV–visible spectroscopy produces absorbance or transmittance measurements associated primarily with electronic transitions. It is frequently used for concentration determination, reaction monitoring, colorimetric sensing, pharmaceutical assays, nanoparticle characterization, and environmental water analysis. AI models may use complete UV–visible spectra rather than a single analytical wavelength, enabling the resolution of overlapping analytes and nonlinear matrix-dependent responses. Acquisition metadata should include optical path length, solvent, pH, temperature, wavelength interval, scan speed, reference solution, and instrument bandwidth because these conditions influence spectral shape and intensity (Xue *et al.,*2023). Fluorescence measurements may include excitation spectra, emission spectra, fluorescence lifetimes, polarization, or complete excitation–emission matrices. These data are highly sensitive and can be generated by fluorometers, fluorescence microscopes, sensor arrays, and biosensing platforms. In environmental research, fluorescence is used to characterize dissolved organic matter, pollutants, microbial activity, and water-quality changes, whereas pharmaceutical applications include drug quantification, interaction studies, protein binding, and formulation analysis (Xue *et al.,*2023).

Fluorescence acquisition is particularly sensitive to photobleaching, quenching, inner-filter effects, pH, solvent polarity, temperature, and molecular aggregation. Excitation–emission matrices also generate three-dimensional datasets containing excitation wavelength, emission wavelength, and intensity. Such data are well suited to chemometric and deep-learning analysis, but only when acquisition conditions are standardized and the model is supplied with appropriate blank, reference, and calibration measurements (Xue *et al.,*2023).

Nuclear magnetic resonance spectroscopy provides detailed information about molecular connectivity, local chemical environments, conformational behavior, intermolecular interactions, and molecular dynamics. The primary raw output is a time-domain free-induction decay signal, which is transformed into a frequency-domain spectrum. One-dimensional proton, carbon, phosphorus, or fluorine NMR can be complemented by two-dimensional and multidimensional experiments that reveal through-bond

or through-space relationships among nuclei (Shukla *et al.,*2023). For small molecules, NMR datasets may include chemical shifts, peak intensities, multiplet patterns, coupling constants, relaxation parameters, diffusion coefficients, and cross-peak coordinates. For proteins and other biomolecules, NMR provides residue-specific information concerning structure, binding, conformational exchange, flexibility, and transient molecular states. These data can support AI-assisted peak detection, spectral reconstruction, chemical-shift prediction, compound identification, mixture analysis, and molecular-structure determination (Shukla *et al.,*2023).

Acquisition variables include magnetic-field strength, probe design, nucleus, pulse sequence, temperature, solvent, sample concentration, pH, number of scans, spectral width, relaxation delay, and sampling strategy. Non-uniform sampling and multidimensional experiments can generate sparse or incomplete datasets that are increasingly reconstructed through deep-learning methods. However, models trained on simulated or highly standardized NMR data may not generalize to experimental spectra affected by peak overlap, imperfect phase correction, instrument drift, or heterogeneous sample composition (Shukla *et al.,*2023).

Liquid chromatography and gas chromatography generate time-resolved separation profiles in which individual compounds are represented by retention times, peak shapes, intensities, and integrated areas. HPLC and LC systems are suitable for polar, nonvolatile, and thermally unstable compounds, while GC is frequently used for volatile and semivolatile substances. When chromatography is coupled with mass spectrometry, each chromatographic feature may additionally contain a mass-to-charge ratio, isotope pattern, adduct information, ion intensity, and fragmentation spectrum (Beck *et al.,*2024).LC–MS data are central to environmental contaminant screening, pharmaceutical impurity analysis, metabolomics, proteomics, degradation studies, and exposure assessment. GC–MS data are widely used for volatile organic compounds, pesticides, hydrocarbons, residual solvents, fragrances, and environmental pollutants. Acquisition can be targeted, focusing on predefined analytes, or untargeted, attempting to record all detectable chemical features within an instrumental range (Beck *et al.,*2024).

High-resolution mass spectrometry improves mass accuracy and resolving power, supporting molecular-formula assignment, isotope-pattern interpretation, suspect screening, and non-target analysis. HRMS datasets may contain tens of thousands of detected features across multiple samples, particularly when full-scan and data-dependent or data-independent tandem-mass-spectrometric acquisition modes are used. These data are especially valuable for detecting emerging environmental contaminants, metabolites, transformation products, pharmaceutical impurities, and degradation compounds for which authentic reference standards may not be available (Beck *et al.,*2024).Critical acquisition variables include chromatographic column, stationary phase, mobile-phase composition, gradient, flow rate, injection volume, ionization method, polarity, collision energy, mass range, scan speed, resolving power, and data-acquisition mode. Instrument manufacturers often store raw data in proprietary formats; conversion into open formats can improve interoperability but may alter metadata or data precision. Quality-control samples, blanks, pooled samples, internal standards, and system-suitability measurements should therefore accompany the analytical batch to support subsequent drift correction and model validation (Beck *et al.,*2024).

Electrochemical sensors generate electrical signals in response to chemical or biological interactions at an electrode surface. Depending on the device, the recorded inputs may include current, potential, charge, impedance, conductance, capacitance, frequency response, or time-dependent voltammetric profiles. Common acquisition modes include amperometry, potentiometry, cyclic voltammetry, differential-pulse voltammetry, square-wave voltammetry, and electrochemical-impedance spectroscopy (Puthongkham *et al.,*2021). Sensor arrays and electronic tongues can generate multichannel response patterns in which individual sensing elements are only partially selective. AI can learn the combined response fingerprint to classify analytes, estimate concentrations, distinguish complex mixtures, and compensate for interferents. Nevertheless, electrode material, surface modification, electrolyte, pH, temperature, scan rate, accumulation time, and reference-electrode stability strongly influence the acquired data and must be standardized or explicitly incorporated into the model (Puthongkham *et al.,*2021).

Biosensors incorporate a biological recognition component, such as an enzyme, antibody, nucleic acid, aptamer, receptor, cell, or microorganism, with a physical transducer. Their outputs may be electrochemical, optical, piezoelectric, thermal, or field-effect based. Biosensor data can support environmental pathogen detection, contaminant monitoring, biomarker measurement, pharmaceutical analysis, and point-of-care testing. However, biological recognition elements introduce additional variability through immobilization efficiency, degradation, nonspecific binding, matrix interference, and storage conditions (Puthongkham *et al.,*2021).

Sensor-generated AI datasets should therefore include calibration curves, negative and positive controls, replicates, response times, limits of detection, selectivity tests, interference studies, and stability measurements. Longitudinal data are particularly important because sensor drift can cause a model trained on newly fabricated devices to perform poorly after

repeated use or storage. Distinguishing measurements obtained from truly independent sensor units from repeated readings of the same device is also essential for unbiased model evaluation. Hyperspectral imaging integrates spectroscopy with spatial imaging by recording a spectrum for every pixel. The resulting data cube normally contains two spatial dimensions and one spectral dimension, allowing the chemical or physical composition of a sample to be mapped across its surface. Depending on the instrument, data may be collected in the visible, near-infrared, short-wave-infrared, mid-infrared, or combined spectral regions (Bhargava *et al.,*2024).

Environmental applications include vegetation-stress detection, contaminant mapping, waste sorting, water assessment, plastic identification, and soil characterization. Pharmaceutical applications include distribution mapping of active ingredients and excipients, tablet-coating analysis, blend-uniformity assessment, counterfeit detection, and characterization of solid dosage forms. Hyperspectral datasets may contain millions of pixel-level observations and hundreds of correlated spectral bands, creating substantial computational and storage requirements (Bhargava *et al.,*2024). Acquisition quality depends on illumination, camera geometry, spatial and spectral resolution, exposure time, sensor calibration, background correction, sample movement, and environmental conditions. White and dark references are normally required to convert raw detector responses into relative reflectance or radiance. Models developed from laboratory images may perform poorly under field or industrial conditions because of changes in illumination, viewing angle, temperature, background material, and sample presentation. Molecular databases provide structured chemical information that complements experimentally acquired analytical signals. PubChem organizes chemical substances, standardized compound structures, biological assays, proteins, genes, pathways, cell lines, taxonomic information, and bioactivity records. Its chemical structures and associated biological data are widely used to develop models for molecular-property prediction, toxicity classification, virtual screening, compound identification, and similarity searching (Kim *et al.,*2023).

ChEMBL provides manually curated bioactivity information for drug-like molecules, including chemical structures, targets, assays, potency values, approved drugs, and clinical candidates. Data are extracted from medicinal-chemistry literature and supplemented through direct dataset deposition. These records support QSAR development, target prediction, drug repurposing, selectivity assessment, ADMET modeling, and AI-driven drug discovery (Mendez *et al.,*2019).

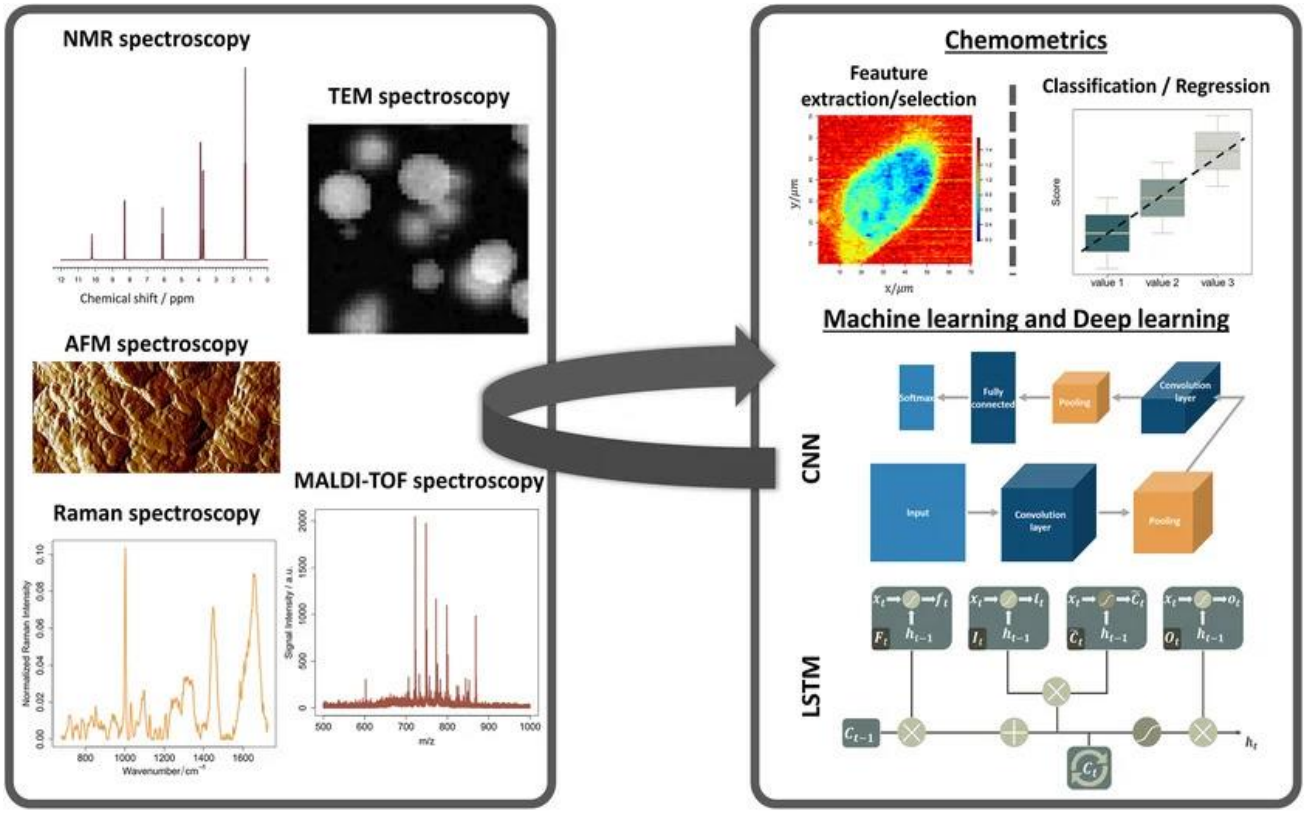

Figure 1: Integration of Multimodal Analytical Data with Chemometrics and Artificial Intelligence for Chemical Information Extraction

Public molecular datasets can contain SMILES strings, InChI identifiers, molecular graphs, physicochemical properties, assay outcomes, toxicity labels, quantum-mechanical calculations, and protein–ligand measurements. MoleculeNet consolidated several such datasets into standardized molecular-machine-learning benchmarks covering quantum, physical, biophysical, and physiological endpoints. It also demonstrated that dataset size, class imbalance, molecular featurization, and data-splitting strategy strongly influence apparent model performance (Wu *et al.,*2018). Database records should not be assumed to be immediately model-ready. Different repositories may represent salts, mixtures, stereochemistry, tautomers, protonation states, assay units, or inactive compounds differently. Duplicate structures, inconsistent identifiers, conflicting activity values, and incomplete assay descriptions can introduce label noise. Molecular database acquisition must therefore be accompanied by source tracking, structure standardization, unit harmonization, duplicate management, and documentation of the database version and access date (Kim *et al.,*2023; Mendez *et al.,*2019).

The figure presents major analytical data sources including NMR, mass spectrometry, spectroscopy, microscopy, and chromatography and their integration with chemometric, machine-learning, and deep-learning methods. This framework supports chemical identification, classification, quantification, and interpretation in environmental and pharmaceutical research. *Reproduced from Houhou and Bocklitz (2021) under the CC BY-NC 4.0 licence.*

Data Preprocessing, Quality Control, and Feature Extraction

Raw analytical and molecular data are rarely suitable for direct use in artificial-intelligence models. Instrumental measurements frequently contain background signals, random noise, retention-time shifts, missing observations, batch effects, overlapping peaks, and variations caused by sample preparation or instrument configuration. Data preprocessing is therefore required to separate chemically meaningful information from technical variation and transform heterogeneous measurements into consistent numerical features. As illustrated in Figure 2, a typical chromatography–mass-spectrometry workflow progresses from noise smoothing and baseline correction to peak detection, alignment, normalization, spectral-library matching, visualization, and downstream statistical or machine-learning analysis (Kiseleva *et al.,*2022).

The preprocessing workflow should be selected according to the analytical platform, sample matrix, intended prediction task, and dominant sources of variability. A transformation that improves one dataset may remove useful information or introduce artifacts into another. Consequently, preprocessing decisions should be treated as model-development parameters, documented transparently, optimized using training data, and evaluated through independent validation rather than applied as fixed universal procedures (Butler *et al.,*2018; Deng *et al.,*2023). Baseline drift is common in FTIR, Raman, fluorescence, NMR, chromatographic, and mass-spectrometric measurements. It may originate from fluorescence backgrounds, light scattering, detector instability, mobile-phase changes, column bleeding, solvent signals, sample thickness, or changing environmental conditions. Common baseline-correction methods include polynomial fitting, rubber-band correction, asymmetric least-squares procedures, morphological filtering, derivatives, and local background subtraction. The selected algorithm must preserve genuine peak intensities and shapes because excessive correction may remove broad but chemically informative bands or generate artificial negative signals (Butler *et al.,*2018; Kiseleva *et al.,*2022).

Random noise can conceal low-intensity analytical features and reduce the reproducibility of peak detection. Moving-average filters, Gaussian smoothing, wavelet denoising, median filters, and Savitzky–Golay smoothing are commonly used to improve the signal-to-noise ratio. However, strong smoothing can broaden narrow peaks, merge neighboring signals, suppress trace-level compounds, and distort quantitative relationships. The filter type, window length, polynomial order, and number of iterations should therefore be optimized against reference signals or quality-control samples rather than selected only on the visual smoothness of the processed spectrum (Butler *et al.,*2018).

Deep-learning approaches provide an alternative to conventional filtering. Brandt et al. (2021) demonstrated that an autoencoder could reconstruct low-quality FTIR and Raman spectra affected by noise, baseline bending, fluorescence, scattering, and band distortion. Such models may process complex artifacts in a single computational step, but they require representative training data and can generate misleading reconstructions when applied to noise patterns or instruments outside their training domain. Reconstructed spectra should therefore be compared with independently acquired high-quality measurements before being accepted for chemical interpretation (Brandt *et al.,*2021).

Normalization aims to reduce differences in overall signal magnitude caused by sample amount, dilution, optical path length, ionization efficiency, detector response, or injection variability. Common methods include total-area normalization, vector normalization, probabilistic-quotient normalization, standard-normal-variate transformation, multiplicative scatter correction, internal-standard normalization, mean centering, autoscaling, and Pareto scaling. The choice of method affects both the distribution of variables and the relative importance assigned to abundant and low-intensity features (González-Domínguez *et al.,*2024;

Kiseleva *et al.,*2022). Normalization cannot compensate for poor experimental design or severe technical failure. Total-area normalization assumes that the overall measured signal is reasonably comparable among samples, whereas internal-standard normalization depends on the stability and suitability of the selected reference compound. Autoscaling gives each feature equal variance but can amplify unstable low-intensity variables, while Pareto scaling preserves more of the original variance structure. Several normalization strategies should therefore be evaluated using quality-control reproducibility, biological separation, and external predictive performance rather than choosing the procedure that produces the clearest visual clustering (González-Domínguez *et al.,*2024).

Peak detection converts continuous chromatographic or spectral signals into discrete, machine-readable features. In chromatography–mass spectrometry, each feature may be defined by retention time, mass-to-charge ratio, intensity, peak area, peak width, isotope distribution, and fragmentation information. Peak-detection algorithms distinguish genuine analytical signals from local noise using intensity thresholds, signal-to-noise ratios, expected peak shapes, chromatographic continuity, or learned pattern-recognition models (Guo & Huan, 2023; Kiseleva *et al.,*2022).

Different peak-picking algorithms can generate substantially different feature tables from the same raw dataset. Guo and Huan (2023) showed that commonly used LC–MS algorithms differ in how they respond to peak width, sharpness, intensity, mass deviation, and scan density. These discrepancies can influence the number of retained features, compound annotation, biomarker selection, and downstream model performance. Peak-picking software, parameter values, minimum signal thresholds, and manual-curation rules must therefore be reported as part of the analytical methodology. Machine learning is increasingly being used to improve peak curation. Gloaguen et al. (2022) developed a deep-learning-assisted approach that distinguishes high-quality chromatographic peaks from false or poorly integrated features, reducing dependence on labor-intensive manual inspection. Nevertheless, automated peak curation should be validated using expert-labelled chromatograms, authentic standards, and independent datasets because the model may learn instrument- or pipeline-specific peak characteristics rather than general chromatographic quality criteria (Gloaguen *et al.,*2022).

Peak alignment matches corresponding signals across samples when their positions vary because of retention-time drift, spectral shifts, temperature changes, pH, column aging, or calibration differences. Alignment can be performed using internal standards, landmark peaks, retention indices, correlation-based warping, dynamic time warping, or nonlinear transformation. Inadequate alignment may represent one compound as several separate variables, whereas overly permissive alignment may merge distinct compounds. Alignment tolerances should therefore reflect the instrument's mass accuracy, chromatographic resolution, and expected retention variability.

Peak deconvolution is particularly important when several analytes co-elute or produce overlapping spectral signals. Deconvolution methods separate composite peaks using differences in retention profiles, ion traces, spectral shapes, fragmentation patterns, or multivariate decomposition. For GC–MS, deconvolution can recover individual mass spectra from overlapping chromatographic peaks before comparison with spectral libraries. In spectroscopy, multivariate curve resolution and related methods can estimate the contributions of overlapping chemical components. The quality of deconvolution directly affects compound identification and quantitative feature extraction (Kiseleva *et al.,*2022). Following peak detection, alignment, and deconvolution, chromatographic data are commonly converted into a sample-by-feature matrix. Each column represents a detected chemical feature and each row represents an independent sample, while matrix values correspond to peak area, height, relative abundance, or normalized intensity. Additional variables may include retention indices, isotope ratios, adduct groups, collision cross-sections, and tandem-mass-spectral similarity scores.

Feature annotation links detected signals to probable chemical identities through comparison with experimental standards, retention indices, accurate masses, isotope patterns, spectral libraries, or in-silico predictions. Figure 2 depicts this transition from cleaned GC–MS data to spectral-library matching and downstream analysis. Because library similarity does not independently prove chemical identity, annotation confidence should distinguish confirmed compounds from probable structures, molecular classes, and unknown features (Kiseleva *et al.,*2022). The selected peak-processing pipeline can materially alter the final chemical interpretation. Schulze et al. (2023) demonstrated that peak-picking choices influence environmental non-target analysis, including the features retained for subsequent prioritization. Such findings emphasize that an AI model operates on a computationally constructed feature table rather than an entirely objective representation of the original sample. Raw-data inspection and sensitivity analyses using alternative preprocessing parameters are therefore important for evaluating whether conclusions remain stable (Schulze *et al.,*2023).

Missing observations are common in chromatographic, mass-spectrometric, sensor, and biological datasets. A missing value may reflect genuine absence, concentration below the detection limit, failed peak integration, ion suppression, incomplete alignment, sample loss, or instrument malfunction. These

mechanisms have different statistical meanings; therefore, missing values should not automatically be replaced by zero or by a single low value. Approaches include feature removal, minimum-value substitution, half-minimum replacement, k-nearest-neighbour imputation, random-forest imputation, Bayesian methods, principal-component-based procedures, and mechanism-aware models. Dekermanjian *et al.,* (2022) proposed a two-step approach that first evaluates the likely missingness mechanism and then applies an appropriate imputation strategy. This is preferable to treating all missing values as random because abundance-dependent missingness is common in analytical datasets (Dekermanjian *et al.,*2022).

Imputation must be conducted after the dataset is divided into training and testing subsets. Applying imputation to the complete dataset allows information from the test set to influence the training data and creates data leakage. The imputation model should be fitted using training observations only and subsequently applied without refitting to validation or external samples. The proportion and pattern of missing values should also be reported so readers can distinguish a well-measured dataset from one largely reconstructed computationally. Batch effects arise from differences in sample-preparation date, reagent lot, operator, chromatographic column, plate, instrument, acquisition sequence, laboratory environment, or storage duration. They can create systematic variation stronger than the environmental or pharmaceutical signal of interest. Appropriate experimental design should therefore include randomized acquisition order, blanks, internal standards, pooled quality-control samples, replicate reference materials, and system-suitability tests (González-Domínguez *et al.,*2024).

Quality-control samples enable assessment of signal stability, retention-time drift, mass accuracy, peak reproducibility, background contamination, and analytical precision. Features with excessive coefficients of variation in pooled quality-control samples may be removed or flagged as unreliable. Blanks assist in identifying contaminants introduced by solvents, columns, extraction materials, or laboratory handling. The position and frequency of quality-control injections should be sufficient to reveal both gradual drift and sudden instrument instability (González-Domínguez *et al.,*2024). Batch-correction methods include mean or median centering, scaling, quality-control-based robust locally estimated scatterplot smoothing, ComBat, EigenMS, removal of unwanted variation, wavelet-independent-component approaches, and random-forest-based correction. The malbacR framework consolidates multiple batch-correction techniques and permits comparison of their effects on small-molecule omics data (Leach *et al.,*2023). Correction should remove technical variation while preserving true chemical or biological differences; a visibly well-mixed batch distribution is not sufficient evidence that the correction is scientifically valid.

Batch and biological class must not be completely confounded. When all control samples are analyzed in one batch and all exposed or treated samples in another, computational correction cannot reliably distinguish technical and biological effects. Preventive experimental design is therefore more reliable than attempting to repair severe confounding after data collection. Outliers may represent sample contamination, preparation failure, instrument malfunction, incorrect labels, unusual but genuine chemistry, or observations outside the model's intended domain. Detection methods include univariate range checks, robust z-scores, Hotelling's T^2, Q-residuals, principal-component score plots, hierarchical clustering, robust covariance estimation, isolation forests, and local-outlier methods. Outliers should be investigated rather than removed automatically because rare environmental contaminants or unusual pharmaceutical batches may be scientifically important.

Removal decisions should be based on predefined analytical criteria, documented reasons, and examination of raw data. A point should not be deleted merely because it reduces model accuracy or weakens group separation. Where uncertainty remains, sensitivity analysis can compare results with and without the observation, allowing the influence of the suspected outlier to be reported transparently (González-Domínguez *et al.,*2024). Spectral augmentation expands a training dataset by introducing controlled and chemically plausible variations into existing measurements. Common transformations include adding realistic noise, altering baselines, shifting peaks, scaling intensities, masking spectral regions, interpolating between spectra, or mixing signals at known proportions. Generative adversarial networks, autoencoders, variational models, and diffusion-based approaches can also produce synthetic spectra for underrepresented classes (Flanagan *et al.,*2025). Augmentation can improve robustness to minor acquisition changes and class imbalance, but unrealistic synthetic variation may teach the model relationships that do not occur experimentally. Augmentation parameters should be derived from measured instrument variability, applied only to training data, and evaluated on untouched experimental samples. Synthetic spectra should supplement rather than replace independent measurements, particularly when the model will support environmental-risk or pharmaceutical-quality decisions (Flanagan *et al.,*2025). Molecular data require preprocessing parallel to that applied to instrumental measurements. Chemical structures obtained from databases may contain salts, counterions, mixtures, duplicate records, inconsistent aromaticity, alternative tautomers, uncertain stereochemistry, or different protonation states. Before descriptors, fingerprints, SMILES strings, molecular graphs, or three-dimensional

conformers are generated, structures should be parsed, validated, standardized, and linked to persistent identifiers (David *et al.,*2020).

Typical molecular-standardization operations include salt removal or consistent salt retention, charge normalization, aromaticity assignment, tautomer handling, protonation-state selection, stereochemical verification, duplicate detection, and unit harmonization for experimental endpoints. Inconsistent standardization may cause the same compound to appear in both training and test sets under different representations or may incorrectly merge biologically distinct chemical forms. Such errors can inflate performance and weaken the chemical validity of structure–property models (David *et al.,*2020; Deng *et al.,*2023). Feature extraction subsequently converts standardized molecules into numerical inputs, including physicochemical descriptors, structural fingerprints, fragment counts, SMILES tokens, graph-based atomic features, spatial coordinates, or learned embeddings. The representation should be selected according to the endpoint and dataset size. Comparative evidence shows that complex learned representations do not consistently outperform carefully constructed descriptors or fingerprints, especially in small, noisy, or chemically narrow datasets (Deng *et al.,*2023).

Instrument variability is a major source of distribution shift in analytical AI. Differences in detector response, spectral resolution, wavelength calibration, laser power, column properties, ionization efficiency, mass accuracy, acquisition software, and vendor-specific processing can alter the measured data even when the chemical sample is unchanged. A model developed using one instrument may therefore learn device-specific signatures and fail when transferred to another laboratory or platform. Mitigation strategies include common reference materials, instrument standardization, calibration transfer, domain adaptation, signal harmonization, transfer learning, and inclusion of multi-instrument data during model development. Instrument identity should also be retained as metadata so that performance can be stratified by device and laboratory. Models intended for broad deployment should be evaluated using external instruments rather than only repeated measurements from the device used for training (Brandt *et al.,*2021; Leach *et al.,*2023).

Preprocessing and feature extraction ultimately determine which aspects of the original measurement remain available to an AI model. Baseline correction, smoothing, normalization, peak detection, alignment, missing-value treatment, batch correction, augmentation, and molecular standardization can improve reliability, but each step can also remove information or introduce bias. All data-dependent transformations must therefore be fitted within the training set and applied unchanged to validation and external-test samples. As represented in Figure 2, the objective is not simply to produce visually clean spectra or chromatograms, but to create a traceable and chemically defensible feature matrix for library matching, visualization, statistical analysis, and predictive modeling. A robust pipeline should preserve raw files, document every processing parameter, incorporate appropriate quality controls, test alternative preprocessing choices, and demonstrate that model performance remains stable across independent samples, batches, instruments, and chemical spaces.

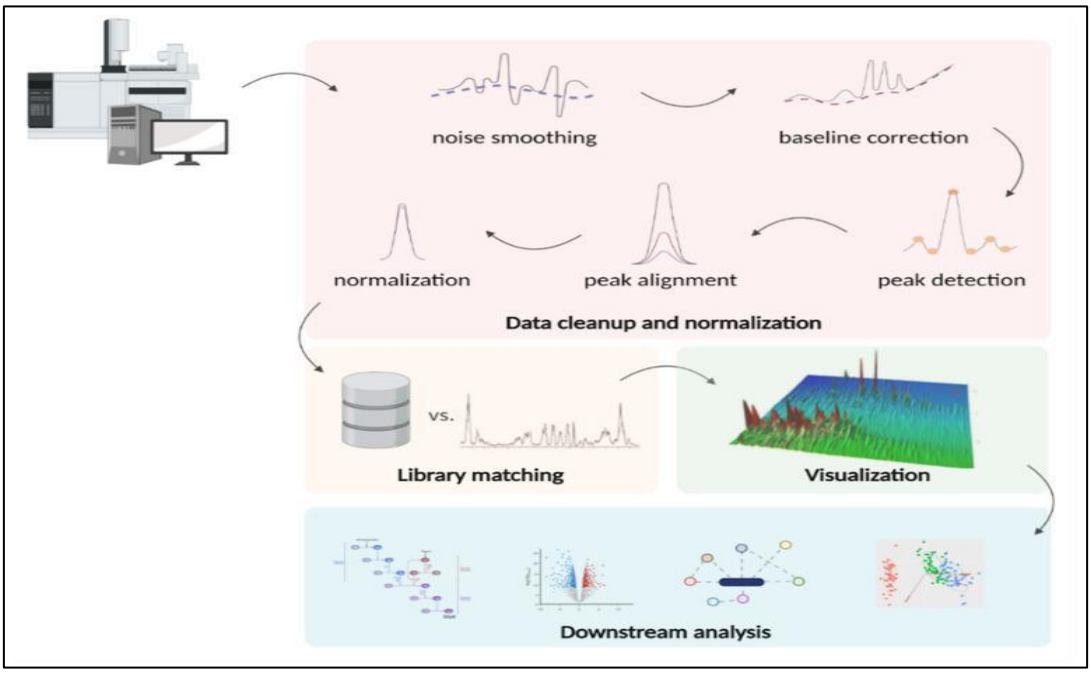

Figure 2: Preprocessing, Feature Extraction, and Downstream Analysis of Analytical Data

The figure illustrates the transformation of raw analytical data through baseline correction, noise removal, smoothing, normalization, and peak alignment. Peak deconvolution, chromatographic feature extraction, missing-value treatment, batch correction, and outlier detection improve data quality and consistency. Spectral augmentation, molecular standardization, and control of instrument variability enhance model robustness and transferability. These steps generate reliable, standardized, and informative features for downstream machine-learning and molecular-modeling applications.

Molecular Representations and Multimodal Data Integration

Molecular representation is the interface through which chemical structure, experimental evidence, and biological context become accessible to machine-learning algorithms. No single representation preserves every molecular property; each emphasizes a different information level, from bulk physicochemical behavior to atom-level topology, three-dimensional geometry, protein context, or measured spectra. Representation choice influences accuracy and interpretability, while multimodal integration combines complementary views within a unified predictive framework (David *et al.,*2020; Zhou *et al.,*2025).

Physicochemical descriptors provide compact numerical summaries such as molecular weight, lipophilicity, polar surface area, hydrogen-bond counts, formal charge, rotatable bonds, and electronic or topological indices. Their advantages are efficiency and direct chemical interpretation. However, predefined descriptors may omit structural patterns important to a particular endpoint. Molecular fingerprints instead encode the presence or frequency of fragments and local atomic environments as binary or count vectors. Extended-connectivity fingerprints remain strong baselines for QSAR, toxicity prediction, and virtual screening, although sparsity, hashing collisions, and weak representation of stereochemistry or global geometry can limit performance (David *et al.,*2020; Yang *et al.,*2019). String representations express molecular graphs as sequences suitable for recurrent networks and transformers. SMILES is compact and can encode atoms, bonds, rings, branching, charge, and stereochemistry. Its sequential form supports large-scale chemical language modeling; MoLFormer, for example, learned representations from 1.1 billion unlabeled molecular strings and transferred them to multiple property-prediction tasks (Ross *et al.,*2022). Nevertheless, one molecule may have several valid SMILES forms, and arbitrary token combinations can generate invalid structures. SELFIES addresses this limitation through a grammar that constrains generated strings to valid molecular graphs, making it attractive for molecular generation and optimization, although its sequences can be longer and less intuitive than canonical SMILES (Krenn *et al.,*2020).

Molecular graphs represent atoms as nodes and bonds as edges, with attributes describing elements, hybridization, charge, aromaticity, and bond order. Graph neural networks learn task-specific features by propagating information between neighboring atoms, reducing reliance on manually designed variables. Their architecture naturally supports structure–property and structure–activity modeling. However, two-dimensional graphs may represent conformers and stereoisomers inadequately when properties depend on spatial arrangement. Three-dimensional representations add coordinates, distances, bond angles, torsions, molecular surfaces, or conformational ensembles. Geometry-enhanced models can capture spatial effects relevant to binding, solubility, reactivity, and optical behavior, but depend on accurate conformer generation, protonation state, stereochemistry, and treatment of flexibility (Fang *et al.,*2022). Protein representations extend molecular modeling to biological targets. Protein sequences can be tokenized as amino-acid strings, whereas structures may be encoded through residue-contact maps, atomistic graphs, binding-pocket surfaces, or three-dimensional coordinates. Protein language models learn contextual embeddings that capture evolutionary, structural, and functional regularities. Rives et al. (2021) trained an unsupervised model on 250 million protein sequences and found that learned features reflected biological structure and function. Such embeddings can be combined with ligand fingerprints, graphs, or conformers to predict drug–target interactions, binding affinity, selectivity, and off-target activity.

Spectral embeddings provide experimentally grounded representations. Instead of reducing NMR, infrared, Raman, or mass spectra to manually selected peaks, encoders transform profiles into continuous latent vectors. These vectors can capture distributed signal patterns and structural relationships not readily summarized by conventional peak tables. Spec2Mol maps MS/MS spectra into embeddings that are decoded into candidate SMILES structures, thereby connecting fragmentation evidence with molecular representation learning (Litsa *et al.,*2023). Spectral embeddings are valuable when integrated with structural data, but they require consistent acquisition, preprocessing, missing-data handling, and cross-instrument validation. Self-supervised learning reduces dependence on costly labels by pretraining models through tasks constructed from the data, including atom or token masking, fragment reconstruction, contrastive learning, and geometric prediction. MolCLR uses atom masking, bond deletion, and subgraph removal to create contrastive views from 10 million unlabeled molecules (Wang *et al.,*2022). Geometry-aware self-supervision can additionally learn bond lengths, angles, and spatial relationships (Fang *et al.,*2022). These methods yield transferable embeddings, but their value depends on pretraining quality; large datasets do not eliminate duplicate structures, chemical bias, or noisy labels.

Multimodal fusion combines complementary representations at different stages. In early fusion, raw variables or modality-specific features are merged before or during representation learning. This permits immediate cross-modal interaction but requires compatible scaling, complete modalities, and control of dimensional imbalance. As shown in Figure 3, the multimodal fusion with relational learning framework combines multimodal chemical information during pretraining to initialize a single graph neural network. Early fusion can be effective when the modalities are strongly related to the downstream endpoint, but predefined weighting may allow a large or noisy modality to dominate (Zhou *et al.,*2025). Intermediate fusion first processes each modality through a specialized encoder and then merges the resulting embeddings by concatenation, attention, cross-modal interaction, or shared neural layers. This preserves modality-specific structure while allowing complementary information to interact at an abstract level. In Figure 3, separate initialized graph networks encode modalities before their representations are merged for prediction. Intermediate fusion is suited to combining graphs, fingerprints, SMILES, images, protein features, and spectra, but requires sample alignment and can be difficult when one modality is missing or substantially noisier than the others (Zhou *et al.,*2025).

Late fusion, or decision-level fusion, trains separate models for each modality and combines their outputs through averaging, voting, stacking, or learned weighting. It accommodates heterogeneous inputs and missing modalities more easily while preserving each source's contribution. Figure 3 represents this strategy through modality-specific graph networks and predictors whose outputs form a final decision. Its limitation is that cross-modal interactions are modeled only after predictions have been produced. Fusion-stage selection should therefore be task dependent: early fusion favors tightly related inputs, intermediate fusion captures complementary latent interactions, and late fusion is preferable when modalities differ in quality, availability, or predictive strength (Zhou *et al.,*2025). Overall, descriptors and fingerprints provide efficient and interpretable baselines; strings support chemical language modeling; graphs encode connectivity; conformers and protein representations introduce spatial and biological context; spectral embeddings anchor predictions in experimental evidence; and self-supervised models transfer knowledge from large unlabeled collections. Their integration can produce richer models, but benefits should be demonstrated through ablation studies, external validation, missing-modality testing, and chemically meaningful interpretation rather than inferred from architectural complexity alone.

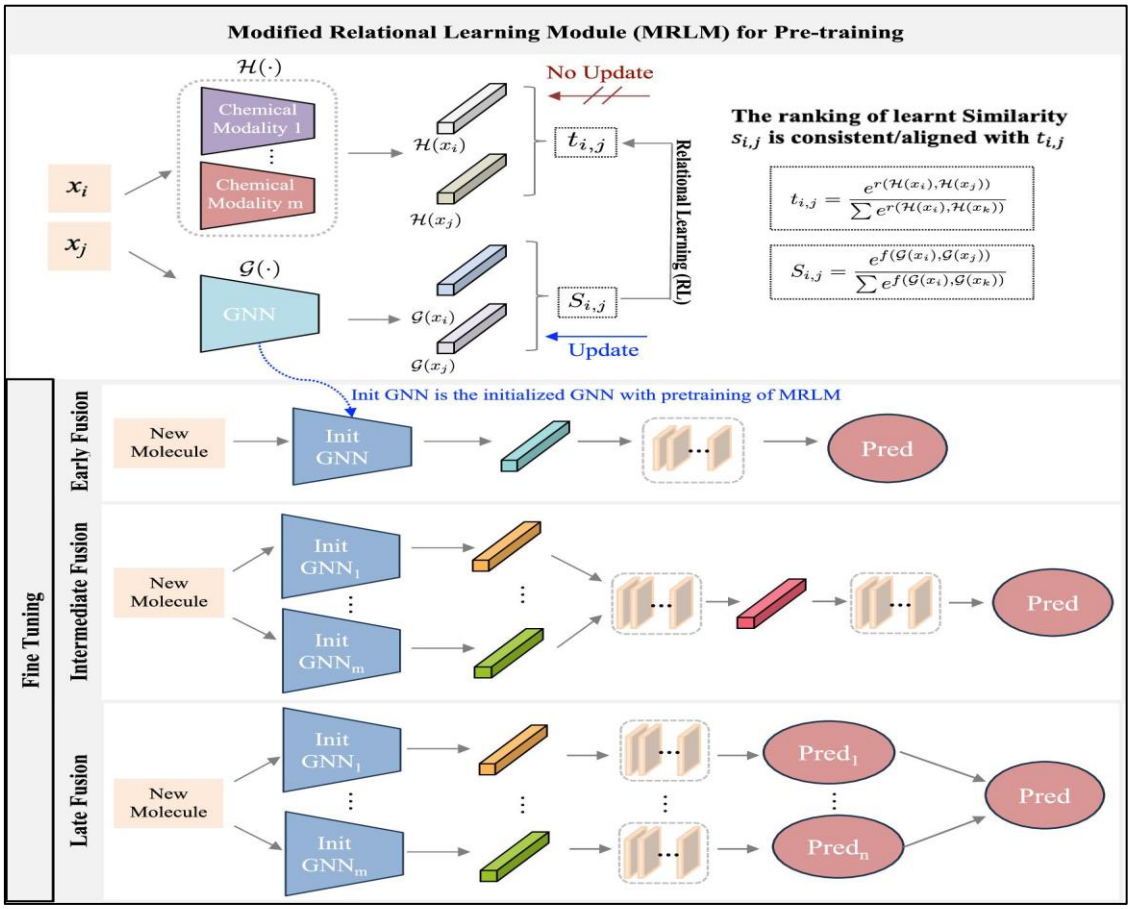

Figure 3: Multimodal Molecular Representation Learning and Fusion for Predictive Chemical Modeling

This section compares conventional and learned molecular representations, including physicochemical descriptors, fingerprints, SMILES/SELFIES strings, molecular graphs, three-dimensional conformers, protein features, spectral embeddings, and self-supervised representations. It further examines how complementary modalities are combined through early-, intermediate-, and late-stage fusion to generate unified representations for molecular-property prediction and related analytical applications. Zhou et al. (2025).

Artificial Intelligence Modeling, Validation, and Interpretation

Chemometric, Machine-Learning, and Deep-Learning Approaches

Artificial-intelligence modeling in analytical science spans a continuum from conventional latent-variable chemometrics to nonlinear machine-learning and deep-learning architectures. As illustrated in Figure 4, traditional workflows commonly separate preprocessing, handcrafted feature extraction, and model fitting, whereas deep-learning pipelines can learn representations directly from spectra, chromatograms, images, or sequential signals. These approaches should not be arranged as a simple hierarchy in which newer methods are assumed to be superior; suitability depends on sample size, dimensionality, noise, interpretability requirements, and independent validation (Brereton *et al.,*2018; Houhou & Bocklitz, 2021).

Principal component analysis is an unsupervised dimensionality-reduction method that transforms correlated variables into orthogonal principal components ordered by explained variance. In spectroscopy, chromatography, sensor analysis, and omics, PCA is used to visualize clustering, identify outliers, evaluate batch structure, and explore dominant sources of variation. Its strengths include efficiency and interpretable scores and loadings. However, PCA maximizes total variance rather than class separation or predictive relevance, so low-variance chemical signals associated with a target endpoint may be overlooked (Brereton *et al.,*2018). Partial least-squares regression addresses supervised quantitative prediction by constructing latent variables that maximize covariance between a predictor matrix and a continuous response. It remains effective for highly collinear spectral data and relatively small calibration datasets. PLS-discriminant analysis adapts this framework to categorical outcomes, while linear discriminant analysis searches for projections maximizing between-class separation relative to within-class variation. Although efficient and interpretable, PLS-DA and LDA can overfit when classes are imbalanced, sample numbers are small, or feature selection precedes validation. Permutation testing, nested cross-validation, and external testing are therefore essential (Ruiz-Perez *et al.,*2020).

Multivariate curve resolution differs from conventional classification and regression because it decomposes mixed measurements into chemically meaningful concentration and spectral profiles. MCR, particularly MCR–alternating least squares, is useful for resolving overlapping peaks, evolving reactions, chromatographic elution profiles, hyperspectral images, and multicomponent process data. Non-negativity, closure, selectivity, and known-profile constraints can improve chemical interpretability, but rotational ambiguity means that several mathematically acceptable solutions may exist. MCR is therefore strongest when numerical decomposition is combined with experimental knowledge and realistic constraints (de Juan & Tauler, 2021). Support vector machines extend analytical modeling into nonlinear spaces through kernel functions. Support vector classification constructs decision boundaries, while support vector regression predicts continuous properties using an error-tolerant margin. SVMs often perform well with high-dimensional, small-to-medium datasets because model complexity depends mainly on support vectors. However, performance is sensitive to kernel choice, regularization, and scaling, and nonlinear SVM predictions are less directly interpretable than latent-variable models (Houhou & Bocklitz, 2021; Xue *et al.,*2023).

Random forests combine many decorrelated decision trees trained on bootstrapped samples and randomized feature subsets. They handle nonlinear interactions, mixed data types, and complex decision boundaries with limited preprocessing. Their ensemble structure reduces the instability of individual trees and provides feature-importance measures. Nevertheless, correlated spectral variables can distribute importance across neighboring wavelengths, impurity-based importance may favor variables with many split points, and random forests may extrapolate poorly beyond the response range represented during training (Houhou & Bocklitz, 2021). Gradient boosting constructs trees sequentially, with each learner correcting residual errors from the current ensemble. XGBoost strengthens this strategy through regularization, shrinkage, row and column subsampling, and efficient handling of sparse data. These models frequently perform strongly on structured analytical and molecular datasets and capture interactions inaccessible to linear chemometrics. Their flexibility introduces many hyperparameters, however, and aggressive tuning on limited data can produce optimistic results. Nested validation and independent testing are needed to distinguish generalization from tuning-related overfitting (Li *et al.,*2025).

Artificial neural networks model nonlinear relationships through interconnected layers of weighted units. Multilayer perceptrons can perform classification or regression after preprocessing or feature extraction. They offer greater flexibility than linear methods but generally require more observations, regularization, and systematic architecture selection. With limited analytical datasets, shallow networks may be more reliable than highly parameterized architectures, especially when

conventional descriptors or selected spectral variables already capture the dominant information (Xue *et al.,*2023). Convolutional neural networks learn local patterns through shared filters and are well suited to one-dimensional spectra, two-dimensional images, and three-dimensional hyperspectral cubes. Early layers may detect peaks, slopes, or local band combinations, while deeper layers combine these into class- or property-relevant representations. CNNs reduce dependence on handcrafted features and can tolerate moderate local distortions. Their learned filters are not automatically chemically interpretable, and performance may deteriorate when instruments, matrices, resolutions, or acquisition conditions differ from the training data (He *et al.,*2021; Li *et al.,*2025).

Recurrent neural networks are designed for ordered information by passing hidden states across successive input positions. They can model temporal sensor responses, chromatographic sequences, spectral series, and process trajectories. Standard RNNs may struggle with long-range dependencies because of vanishing or exploding gradients; long short-term memory and gated recurrent unit architectures address this through gates controlling information retention and updating. These models are valuable when sequence order is scientifically meaningful but may be unnecessary for static spectra when simpler CNNs or tree-based models perform comparably (He *et al.,*2021).

Autoencoders contain an encoder that compresses data into a latent representation and a decoder that reconstructs the input. They support nonlinear dimensionality reduction, denoising, anomaly detection, feature learning, and augmentation. In FTIR and Raman analysis, autoencoder-based reconstruction has suppressed instrumental noise and baseline artifacts while retaining useful spectral information (Brandt *et al.,*2021). Reconstruction quality should not be judged visually alone, because a network may remove weak genuine peaks or generate plausible but artificial structures.

The strongest strategy is therefore not to select the most complex algorithm automatically, but to compare models under a common, leakage-resistant validation framework. PCA, PLS, PLS-DA, LDA, and MCR offer transparency and remain powerful for structured, collinear, or small datasets. SVMs, random forests, gradient boosting, and XGBoost add nonlinear flexibility for engineered features, whereas ANN, CNN, RNN, and autoencoder architectures enable automated representation learning from larger inputs. Fair comparison requires identical data partitions, preprocessing within training folds, tuning separated from testing, and evaluation on independent samples, instruments, batches, or chemical domains (Brereton *et al.,*2018; Ruiz-Perez *et al.,*2020).

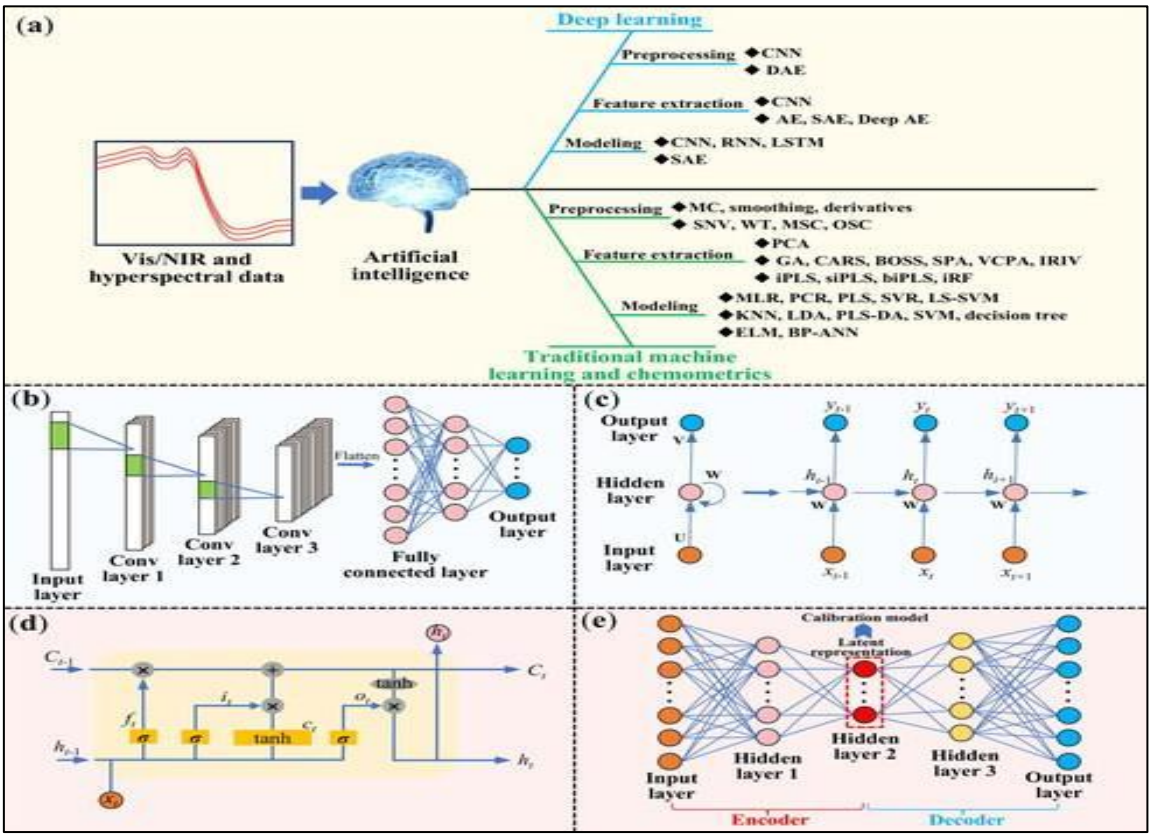

Figure 4: Comparative Workflows of Chemometric, Machine-Learning, and Deep-Learning Methods for Spectral Data Analysis

The figure compares traditional chemometric and machine-learning workflows, involving preprocessing, handcrafted feature extraction, and regression or classification, with deep-learning approaches capable of automated representation learning. It summarizes commonly applied methods, including PCA, PLS, PLS-DA, LDA, SVM, ANN, CNN, RNN, LSTM, and autoencoders, for qualitative and quantitative analytical modeling.

Graph Networks, Transformers, and Emerging AI Paradigms

The rapid expansion of chemical and biological datasets has shifted artificial intelligence from task-specific predictors toward architectures capable of learning transferable molecular representations. As illustrated in Figure 5, graph-based, language-model, high-dimensional, multimodal, and contrastive-learning approaches encode molecular topology, three-dimensional structure, protein context, phenotypic information, and complementary data modalities. These advances support molecular-property prediction, scaffold exploration, toxicity assessment, drug discovery, and chemical design (Wang *et al.,*2025).

Graph neural networks represent molecules as graphs in which atoms are nodes and chemical bonds are edges. Message-passing neural networks iteratively update each atomic representation by aggregating information from neighboring atoms and bonds, after which node-level information is pooled into a molecule-level embedding. This structure-aware formulation is well suited to QSAR, ADMET, environmental-fate, and bioactivity prediction. However, graph models may suffer from over-smoothing, limited treatment of long-range interactions, and inadequate representation of conformational variability unless spatial information is incorporated explicitly (Wang *et al.,*2022; Wang *et al.,*2025). Transformers model molecular strings, reactions, proteins, or graph-derived tokens as contextual sequences. Their multi-head attention mechanisms allow individual tokens to assign different weights to other relevant tokens, enabling the detection of long-range dependencies without recurrent processing. The Molecular Transformer established attention-based chemical reaction prediction using reactant and reagent strings while also providing confidence estimates for its outputs (Schwaller *et al.,*2019). Attention maps may indicate influential molecular regions, although attention weights should not automatically be interpreted as mechanistic explanations.

Protein-language models extend transformer-based learning to amino-acid sequences. Through masked-token prediction and other self-supervised objectives, these systems learn embeddings related to evolutionary conservation, secondary structure, residue contacts, and biological function. Such representations can be integrated with ligand graphs, fingerprints, or three-dimensional conformers for target prediction, protein engineering, binding-affinity estimation, and drug–target interaction modeling. Rives et al. (2021) demonstrated that unsupervised learning from 250 million protein sequences generated representations containing multiscale structural and functional information.

Self-supervised learning similarly exploits large collections of unlabeled molecules through masked-token prediction, graph reconstruction, contrastive learning, or geometric objectives. MolCLR generates alternative molecular views through atom masking, bond deletion, and subgraph removal and trains graph encoders to align representations derived from the same molecule (Wang *et al.,*2022). Transfer learning then adapts a pretrained model to smaller labeled datasets, reducing data requirements. For example, transfer learning successfully specialized a general Molecular Transformer for carbohydrate reactions, where regioselectivity and stereoselectivity present distinctive modeling challenges (Pesciullesi *et al.,*2020). Active learning creates an iterative relationship between prediction and experimentation. A model estimates candidate properties and uncertainty, an acquisition function selects the most informative or promising compounds, and newly obtained computational or experimental results are returned for model updating. This approach can reduce screening costs while directing resources toward uncertain or valuable regions of chemical space. In molecular virtual screening, pool-based active learning recovered many highly ranked ligands while evaluating only a small fraction of extremely large chemical libraries (Graff *et al.,*2021).

Federated learning addresses situations in which pharmaceutical, toxicological, clinical, or proprietary chemical datasets cannot be transferred to a centralized repository. Participating organizations train a shared model by exchanging model updates rather than raw records. FL-QSAR demonstrated the feasibility of collaborative structure–activity modeling while maintaining local control of institutional data (Chen *et al.,*2020). Nevertheless, federated systems remain affected by heterogeneous datasets, inconsistent endpoint definitions, communication requirements, possible information leakage through model updates, and unequal contributions among participating organizations. Physics-informed AI incorporates conservation laws, differential equations, molecular symmetries, energy constraints, or established mechanistic relationships into model architectures or objective functions. Such constraints can reduce dependence on labeled data and discourage physically impossible predictions, making these approaches relevant to molecular simulation, reaction kinetics, transport processes, and environmental-fate modeling. Physics-informed learning is particularly useful for inverse and data-scarce problems, although incomplete or incorrect physical assumptions can introduce

systematic bias into predictions (Karniadakis *et al.,*2021).

Generative AI moves beyond evaluating existing compounds by proposing new molecular structures. Variational autoencoders, generative adversarial networks, autoregressive transformers, and reinforcement-learning models can explore chemical space under constraints related to activity, toxicity, solubility, stability, and synthesizability. Diffusion models generate molecules through progressive denoising and can be conditioned on protein-binding pockets or desired molecular properties. Equivariant diffusion preserves rotational and translational relationships and has enabled structure-based ligand generation, scaffold modification, molecular inpainting, and property optimization (Schneuing *et al.,*2024). Chemical foundation models integrate several of these developments through large-scale pretraining followed by task-specific adaptation. MolE combines molecular graphs, transformer-based disentangled attention, self-supervised learning, and multitask biological training across approximately 842 million molecular graphs (Méndez-Lucio *et al.,*2024). Such models may provide transferable representations for property prediction, toxicity evaluation, and molecular design. Their scientific value, however, depends on training-data quality, chemical-space coverage, calibration, interpretability, external validation, and prospective experimental performance rather than model scale alone.

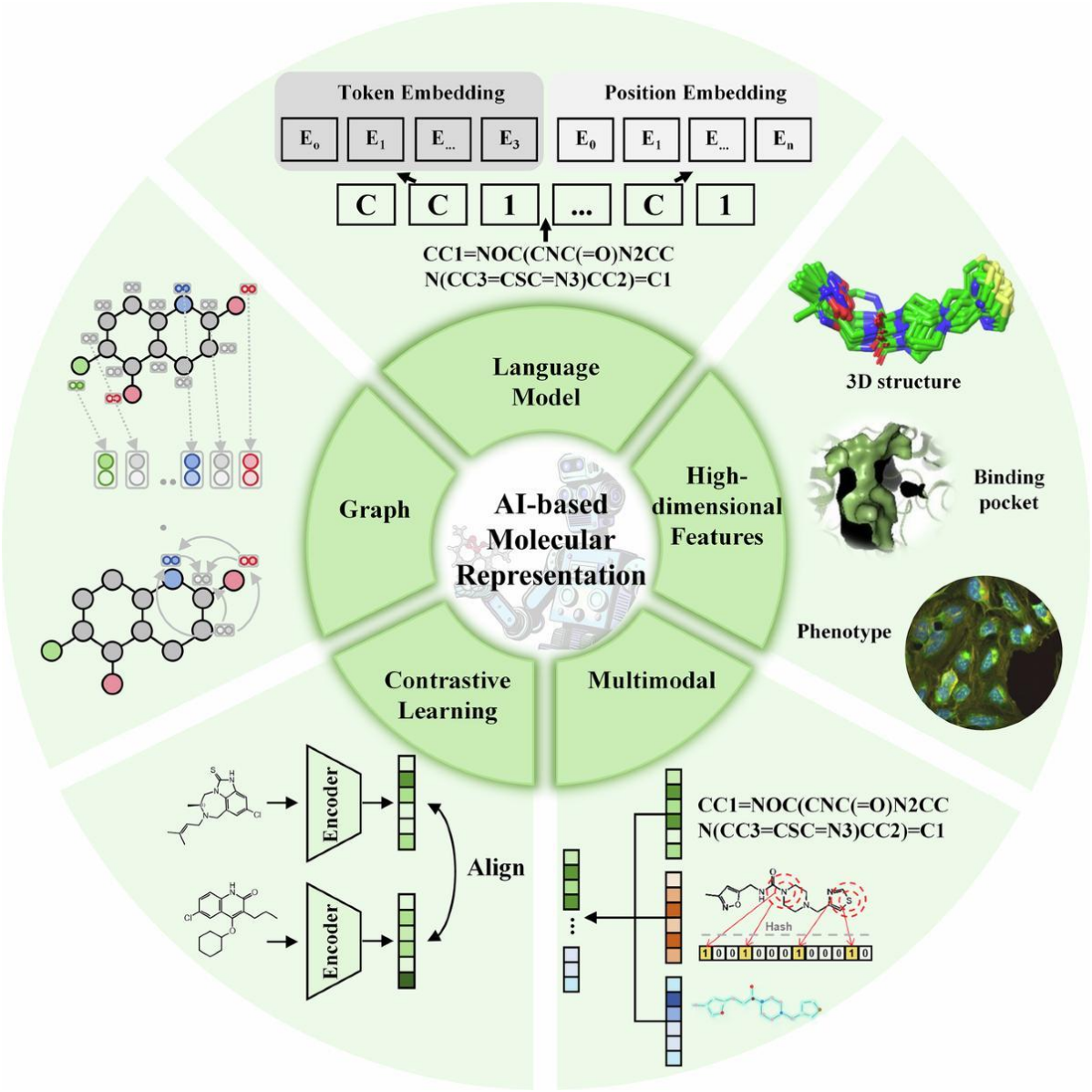

Figure 5: AI-Based Molecular Representation and Emerging Learning Paradigms for Chemical Intelligence.

The figure summarizes five major AI-based molecular learning paradigms: language models, graph neural networks, high-dimensional molecular features, multimodal integration, and contrastive learning. These approaches combine molecular sequences, topology, 3D structures, binding pockets, cellular phenotypes, and

complementary data modalities to support molecular-property prediction, drug discovery, and chemical design.

Model Development, Performance Evaluation, and Explainability

Reliable chemical prediction requires more than selecting a high-performing algorithm. Model development must integrate representative data partitioning, leakage-resistant training, appropriatevalidation, uncertainty assessment, applicability-domain analysis, and chemically meaningful interpretation. As summarized in Figure 6, the workflow should progress from data collection and molecular feature engineering to model training, validation, performance evaluation, confidence estimation, and explainability before practical deployment (Zhang *et al.,*2025). Dataset splitting is a critical determinant of reported performance. Random splitting may place structurally similar molecules, replicate spectra, or measurements from the same experimental batch in both training and test sets, producing optimistic estimates of generalization. Molecular studies should therefore consider scaffold splitting, in which compounds sharing core chemical frameworks are separated, while analytical studies should partition independent samples rather than repeated measurements. Scaffold-based evaluation is generally more demanding because it tests prediction on structurally novel chemical families (Deng *et al.,*2023; Wu *et al.,*2018).

Time-based splitting is particularly relevant when models are intended to predict compounds, formulations, or measurements generated after model development. Earlier records are used for training and later records for testing, thereby approximating prospective deployment. Laboratory-, site-, batch-, and instrument-based partitions similarly evaluate robustness against changes in operators, devices, sample preparation, and acquisition conditions. These designs are more informative than random cross-validation when the final model will be transferred across analytical platforms or research settings (Heid *et al.,*2023; Xue *et al.,*2023). Cross-validation estimates performance variability and supports model selection when external data are limited. However, hyperparameter optimization must be separated from final performance estimation. Nested cross-validation provides an inner loop for tuning parameters and an outer loop for unbiased evaluation. The untouched external test set should not influence preprocessing, descriptor selection, imputation, class balancing, model selection, or threshold optimization. Using test-set information during any of these operations constitutes data leakage and can substantially inflate performance (Deng *et al.,*2023).

External testing provides the strongest retrospective evidence of generalizability when it includes new chemical scaffolds, laboratories, instruments, geographical locations, formulation batches, or acquisition periods. Performance should also be reported across multiple random seeds or repeated partitions because a single split may produce unusually favorable or unfavorable results. Large-scale benchmarking has demonstrated that model rankings and metric values can change considerably with dataset composition, splitting strategy, and chemical-space coverage (Deng *et al.,*2023). Performance metrics must match the scientific task and dataset characteristics. Classification studies commonly report sensitivity, specificity, precision, recall, F1-score, balanced accuracy, Matthews correlation coefficient, receiver-operating-characteristic area under the curve, and precision–recall area under the curve. Precision–recall metrics are particularly informative for rare toxic, active, or contaminated samples because overall accuracy and ROC-AUC may obscure poor minority-class detection. Confusion matrices should accompany summary metrics to reveal the distribution of false-positive and false-negative predictions (Zhang *et al.,*2025). Regression models are commonly evaluated using mean absolute error, root-mean-square error, coefficient of determination, and correlation coefficients. MAE provides an interpretable average error, whereas RMSE penalizes large deviations more strongly. Performance should be compared with experimental variability and simple baseline models rather than interpreted in isolation. Confidence intervals, standard deviations, and repeated-run distributions should also be reported to distinguish meaningful improvement from random variation. Uncertainty quantification determines whether a prediction should be trusted. Relevant approaches include Bayesian models, Gaussian processes, deep ensembles, Monte Carlo dropout, conformal prediction, and distance-based confidence measures. Uncertainty can arise from experimental noise, limited training data, model parameters, dataset splitting, or predictions outside the training distribution. Consequently, uncertainty estimates should be calibrated on independent data and evaluated for their ability to identify high-error predictions rather than presented only as numerical confidence scores (Heid *et al.,*2023).

Applicability-domain analysis defines the chemical or analytical space in which model predictions are considered reliable. Domain assessment may use molecular similarity, leverage, descriptor distance, density estimation, nearest-neighbour analysis, or ensemble disagreement. A compound that lies far from the training set should be flagged even when the model generates a precise numerical output. Defined applicability domains improve the responsible use of QSAR models in environmental screening, toxicology, and pharmaceutical decision-making (Wang *et al.,*2021). Explainability methods help determine whether models rely on chemically meaningful information. SHAP assigns positive or negative contributions to individual descriptors, fingerprints, wavelengths, or experimental variables, enabling global and sample-specific

interpretation. Saliency maps identify influential regions of spectra, images, molecular graphs, or protein sequences, while attention and gradient-based methods can highlight atoms, bonds, or substructures associated with a prediction. Explanations must nevertheless be tested for stability because different methods may provide inconsistent attributions for the same model (Wellawatte *et al.,*2023).

Molecular-substructure interpretation translates feature importance into medicinal or environmental chemistry insights. MolSHAP, for example, estimates the contributions of R-groups within compounds sharing a common molecular core, supporting interpretation of structure–activity relationships and compound optimization (Tian *et al.,*2024). However, attribution does not independently establish causality; highlighted fragments should be compared with known mechanisms, experimental evidence, activity cliffs, and prospective validation. Ultimately, trustworthy model evaluation should integrate predictive performance, uncertainty, applicability, and explanation. A slightly less accurate model with stable external performance, calibrated uncertainty, and chemically defensible interpretations may be more useful than a highly accurate but opaque model evaluated only through random internal splitting.

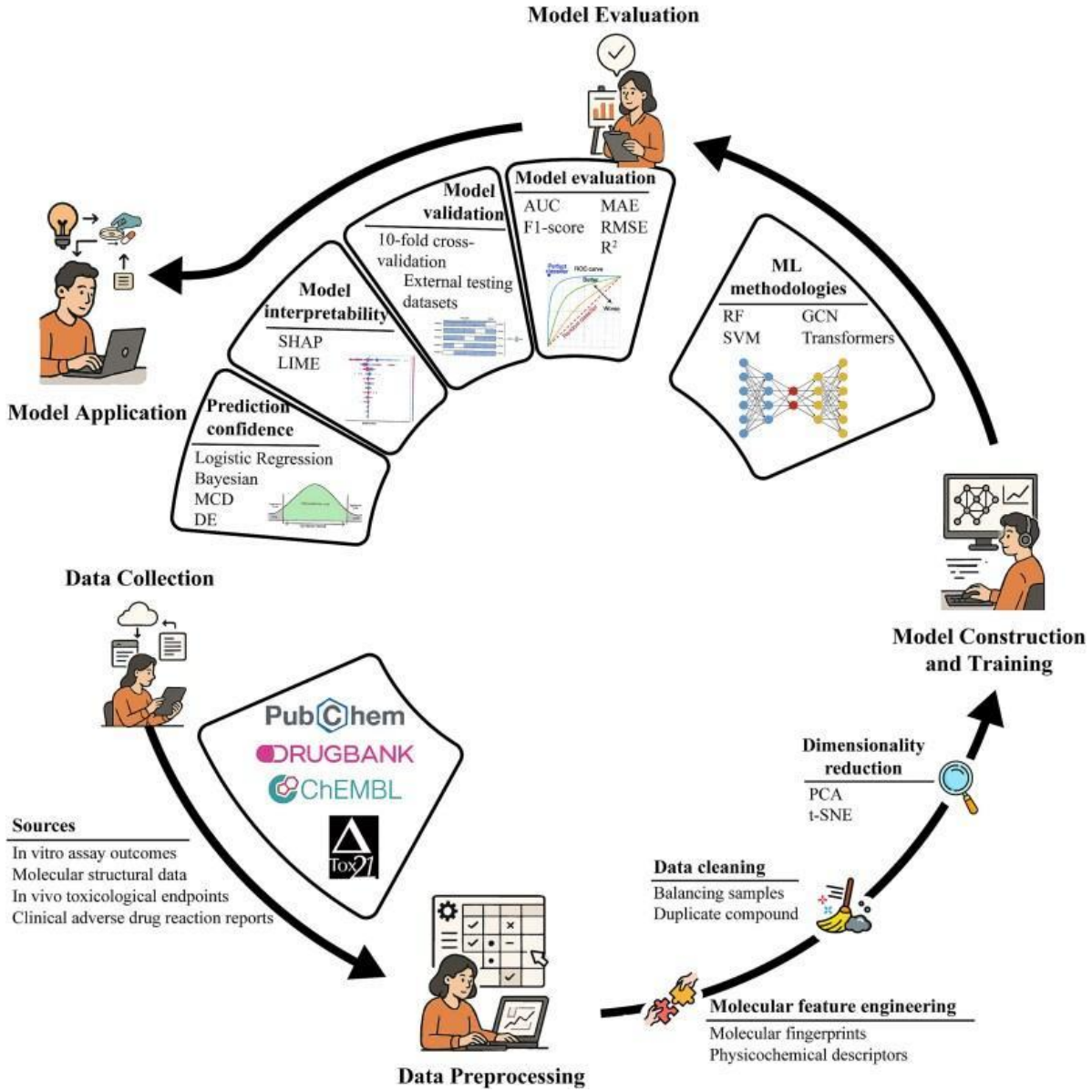

Figure 6: Integrated Workflow for Chemical AI Model Development, Validation, Uncertainty, and Explainability.

The figure illustrates the progression from chemical-data collection and preprocessing to feature engineering, model training, and performance evaluation using classification and regression metrics. It incorporates cross-validation, external testing, SHAP- and LIME-based interpretation, and prediction-confidence assessment to support transparent and reliable chemical predictions.

Environmental Applications

AI-Assisted Detection and Quantification of Environmental Contaminants

Artificial intelligence is increasingly transforming environmental analysis by converting complex instrumental and sensor responses into rapid contaminant identification and quantitative estimates. As illustrated in Figure 7, AI can be integrated with microfluidic devices, paper-based sensors, smartphones, nanoparticles, aptamers, polymers, proteins, and lab-on-a-chip systems to detect multiple pollutant classes. These platforms are particularly valuable when environmental samples contain overlapping chemical signatures, low contaminant concentrations, and strong matrix interference. For pesticides, AI-assisted surface-enhanced Raman spectroscopy enables classification and quantification from complex spectral fingerprints rather than relying on a single diagnostic peak. Flexible silver-nanoparticle SERS substrates combined with machine-learning classifiers have demonstrated sensitive discrimination of pesticide residues in water, illustrating the potential for portable screening outside centralized laboratories (Sahin *et al.,*2022). Similar workflows can be extended to soil extracts, food surfaces, agricultural runoff, and biological samples, provided that training datasets represent realistic concentrations, interferents, and environmental conditions.

Pharmaceuticals, personal-care products, endocrine disruptors, persistent organic pollutants, and industrial chemicals are commonly investigated using LC–HRMS or GC–HRMS. These platforms can generate thousands of chromatographic and mass-spectral features from water, wastewater, sediment, soil, food, and tissue samples. Machine learning supports feature filtering, retention-time prediction, fragmentation interpretation, candidate ranking, source classification, and hazard-based prioritization, thereby directing identification efforts toward environmentally relevant signals rather than only the most abundant compounds (Arturi & Hollender, 2023). This capability is particularly important for transformation products and previously unrecognized contaminants. Parent pesticides, pharmaceuticals, plastic additives, and industrial chemicals may undergo photolysis, oxidation, hydrolysis, microbial degradation, or metabolic transformation, producing compounds that are absent from conventional target lists. AI-assisted non-target screening can link accurate mass, isotopic patterns, retention behavior, tandem mass spectra, predicted structures, and toxicological information to prioritize unknown features. However, machine-learning ranking does not constitute definitive identification; confirmation still requires authentic standards or multiple independent lines of analytical evidence.

PFAS analysis exemplifies the value of combining advanced sensing with machine learning. Conventional quantitative methods depend heavily on reference standards, which are unavailable for many PFAS structures. Zuo et al. (2026) integrated nanopore-based single-molecule electrochemical sensing with multidimensional feature classification and demonstrated standard-free identification and quantification of perfluoroalkyl carboxylic acids in complex environmental matrices. The approach linked ionic-current blockades with molecular volume and used complementary signal features to distinguish structurally similar compounds and isomers (Zuo *et al.,*2026). Heavy-metal detection benefits from sensor arrays because ions such as lead, mercury, cadmium, arsenic, and chromium can produce partially overlapping optical or electrochemical responses. Mandal et al. (2022) combined a fluorescent carbon-nanoparticle array with supervised learning and data augmentation to classify toxic metal ions in laboratory, river-water, and sewage-water samples. AI can improve selectivity by analyzing the complete multichannel response pattern, although reliable quantification requires matrix-matched calibration, interference testing, sensor-stability assessment, and independent validation (Mandal *et al.,*2022).

Volatile organic compounds and toxic gases can be measured through electronic noses containing cross-sensitive sensor arrays. Instead of requiring each sensor to respond selectively to one compound, machine learning recognizes the collective response pattern. Anisimov et al. (2021) developed an organic field-effect-transistor electronic nose capable of discriminating toxic gases under humid conditions using multivariate analysis. Such systems may support air-quality surveillance, industrial-emission monitoring, food assessment, and biological-sample analysis, but temperature, humidity, sensor aging, and long-term drift must be included in model development (Anisimov *et al.,*2021). Microplastic detection increasingly combines FTIR or Raman imaging with machine learning to replace slow manual spectral interpretation. Random-forest analysis of μFTIR images has enabled automated identification of more than 20 polymer types in complex environmental matrices, including wastewater, sediment, soil, compost, and salt samples (Hufnagl *et al.,*2022). Deep-learning and image-based classifiers can also distinguish blended polymers from raw FTIR spectra and spectral images, supporting higher-throughput particle counting and polymer classification (Liu *et al.,*2023).

Nanoplastic detection remains more difficult because particle dimensions approach or fall below the spatial resolution of many routine spectroscopic methods. AI may assist by extracting weak features from Raman, hyperspectral, scattering, or sensor signals, but models require experimentally validated nanoplastic datasets rather than relying exclusively on pristine polymers or simulated spectra. Weathering, additives, biofilms, particle aggregation, and matrix-derived organic matter can alter measured signatures and reduce transferability between laboratory and environmental samples. Quantification across all pollutant classes

requires more than successful classification. Models should report detection and quantification limits, calibration range, recovery, precision, selectivity, false-positive and false-negative rates, uncertainty, and performance in independent real samples. Training and test data must be separated at the level of sampling site, batch, instrument, or environmental specimen to prevent replicate measurements from inflating model accuracy. Matrix-specific validation is essential because a model trained in pure water may fail in sediment extracts, food, serum, wastewater, or biological tissues.

Overall, AI connects microfluidic and sensor technologies with spectroscopy, chromatography, mass spectrometry, imaging, and molecular information to accelerate environmental contaminant analysis. Its greatest contribution lies in resolving overlapping signals, prioritizing unknown features, compensating for multivariate interference, and enabling portable or automated measurements. Nevertheless, AI predictions should complement not replace analytical quality control, certified standards, confirmatory measurements, and transparent uncertainty assessment.

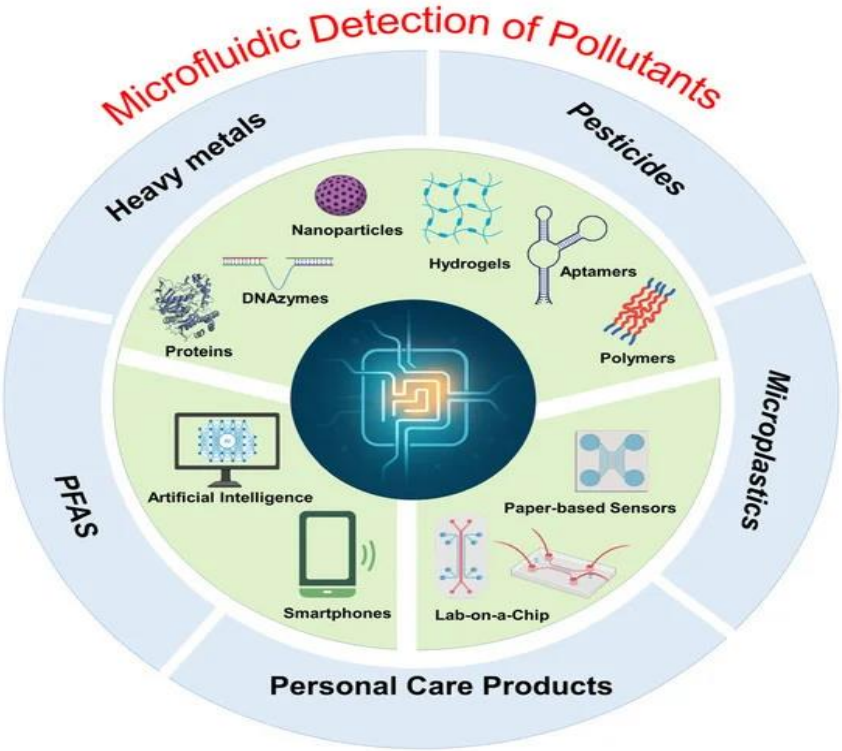

Figure 7: AI-Integrated Microfluidic and Biosensing Platforms for Multipollutant Environmental Detection

The figure illustrates the integration of microfluidics, artificial intelligence, smartphones, paper-based sensors, and lab-on-a-chip systems for rapid environmental monitoring. Nanoparticles, hydrogels, aptamers, polymers, DNAzymes, and proteins enable selective detection of heavy metals, pesticides, PFAS, microplastics, and personal-care contaminants. These interconnected platforms support portable, sensitive, and data-driven pollutant identification and quantification across complex environmental samples.

Suspect and Non-Target Screening of Emerging Pollutants

Suspect screening and non-target screening expand environmental monitoring beyond conventional target-compound lists. Suspect screening searches high-resolution mass spectrometry data for chemicals expected to occur on the basis of production, use, physicochemical properties, environmental pathways, or prior monitoring evidence. Non-target screening begins without a predefined list and attempts to detect, prioritize, and annotate previously unrecognized chemical signals. As illustrated in Figure 8, modern workflows integrate liquid or gas chromatography, ion mobility, high-resolution mass spectrometry, computational chemistry, and machine learning to transform unresolved instrumental signals into prioritized emerging pollutants (Hollender *et al.,*2017; Helmus *et al.,*2021).

The analytical process begins with feature detection, during which software identifies chromatographic peaks using mass-to-charge ratio, retention time, signal intensity, and peak shape. Detected features are aligned across samples to correct small instrumental shifts and enable comparisons among sites, treatments, or sampling periods. Feature grouping is then required because a single chemical may generate several signals through isotopes, adducts, multimers, neutral losses, or in-source fragments. Software selection and parameter settings can substantially influence the number of features retained; therefore, peak-picking

thresholds, alignment tolerances, replicate reproducibility, and quality-control criteria should be reported transparently (Hohrenk *et al.,*2020; Fisher *et al.,*2022). Blank filtering is essential because solvents, extraction materials, laboratory plastics, chromatographic systems, and ion sources can introduce signals that resemble environmental contaminants. Features are commonly evaluated according to their abundance ratio between samples and procedural blanks, frequency of detection, absolute signal intensity, and reproducibility across analytical replicates. A universal blank-subtraction threshold is inappropriate because aggressive filtering may remove genuine low-concentration pollutants, whereas weak filtering increases false-positive identification. Batch-specific blanks, field blanks, extraction blanks, pooled quality-control samples, and replicate injections should therefore be incorporated into the workflow. Open-source platforms such as patRoon support integrated feature processing, blank filtering, componentization, formula assignment, and suspect screening within reproducible environmental workflows (Helmus *et al.,*2021).

Isotope-pattern recognition provides important evidence for elemental composition. Algorithms identify monoisotopic peaks and related isotopologues by evaluating isotope spacing, relative abundance, mass accuracy, and charge state. Distinctive isotope signatures can indicate the presence of chlorine, bromine, sulfur, or other elements, substantially narrowing the possible molecular formulas. Molecular-formula assignment then combines accurate precursor mass, isotope fit, chemical valence rules, elemental restrictions, and MS/MS fragment information. Computational tools such as SIRIUS construct fragmentation trees and assess whether precursor and product-ion formulas are chemically consistent, improving formula ranking beyond accurate mass alone (Dührkop *et al.,*2019). Nevertheless, formula assignment does not establish a unique structure because numerous constitutional and stereochemical isomers can share the same elemental composition. Retention-time prediction supplies an orthogonal criterion for candidate evaluation. Quantitative structure–retention relationships and machine-learning models estimate chromatographic behaviour from molecular descriptors, fingerprints, or learned structural representations. Candidates with predicted retention times that differ substantially from the experimental value can be deprioritized. However, retention depends on stationary phase, mobile-phase composition, gradient, temperature, pH, and instrument conditions; therefore, prediction models should be calibrated for the analytical system and evaluated within a clearly defined applicability domain. Retention-time evidence should support, rather than replace, accurate-mass and fragmentation evidence.

Ion mobility adds collision cross section as another molecular descriptor related to an ion's gas-phase size, shape, charge, and conformation. Experimentally measured CCS values can be compared with database entries or machine-learning predictions to distinguish candidates with similar masses and fragmentation patterns. Celma et al. (2022) demonstrated simultaneous prediction of retention time and CCS values for protonated, deprotonated, and sodium-adducted emerging contaminants using multiple adaptive regression splines. Integrating retention time and CCS with MS^1 and MS^2 data can reduce candidate lists, although prediction uncertainty, ion form, instrument type, and interlaboratory CCS variability must be considered (Celma *et al.,*2022). Spectral-library matching compares experimental MS/MS spectra with reference spectra using fragment masses, intensity distributions, precursor information, and similarity scores. A strong match with a spectrum generated from an authentic standard provides substantial identification evidence, but environmental library coverage remains incomplete, particularly for transformation products, industrial chemicals, metabolites, and newly introduced compounds. When reference spectra are unavailable, in-silico fragmentation predicts plausible product ions for candidate structures. MetFrag combines computational fragmentation with compound databases and additional evidence, while SIRIUS and molecular-fingerprint approaches infer structural characteristics directly from tandem spectra (Ruttkies *et al.,*2016; Dührkop *et al.,*2019).

Candidate ranking should integrate all available evidence rather than relying on a single score. Relevant inputs include precursor-mass error, isotope similarity, molecular-formula plausibility, MS/MS matching, in-silico fragment coverage, retention-time agreement, CCS agreement, suspect-list membership, production volume, occurrence frequency, and environmental plausibility. Machine-learning models can combine these heterogeneous variables and recognize nonlinear relationships, but ranking performance depends on representative training data and independent validation. Models trained mainly on pharmaceuticals or well-characterized metabolites may perform poorly for polymers, surfactants, highly halogenated compounds, or industrial transformation products outside their chemical domain.When no convincing database candidate is available, de novo structure elucidation attempts to infer partial or complete structures directly from spectral evidence. Fragmentation trees, diagnostic ions, neutral losses, molecular fingerprints, chemical-class predictors, and generative neural networks can propose previously unregistered structures. Such predictions are valuable for identifying new transformation pathways and prioritizing unknown pollutants, but generated structures remain hypotheses until supported by complementary evidence. Identification confidence should therefore be communicated using established levels, ranging from confirmed structures supported by authentic standards to probable structures, tentative candidates, unequivocal formulas, or unexplained exact-mass features (Schymanski *et al.,*2014).

Semi-quantification is required when authentic standards are unavailable. Conventional approaches use structurally related surrogate standards, while newer methods predict electrospray ionization efficiency or instrumental response from molecular properties. Liigand et al. (2020) used random-forest models to predict electrospray responses, reporting mean response-prediction errors of approximately 2.2-fold in positive mode and 2.0-fold in negative mode; application to cereal samples produced average concentration errors of approximately 5.4-fold. MS2Quant alternatively estimates concentrations from MS/MS spectra and can be applied to identified and unidentified chemicals (Liigand *et al.,*2020; Sepman *et al.,*2023). Such results should be described as semi-quantitative and accompanied by prediction intervals, matrix-effect assessments, surrogate selection criteria, and uncertainty estimates. Overall, AI-enabled suspect and non-target screening connects feature extraction, chemical annotation, structure prediction, prioritization, and approximate quantification within a unified workflow. Its greatest strength lies in reducing thousands of unresolved features to a manageable set of environmentally relevant candidates. Nevertheless, reliable application requires rigorous blanks, reproducibility testing, transparent software parameters, independent validation, confidence-level reporting, and confirmatory measurements. Machine learning can accelerate interpretation, but authentic standards and orthogonal analytical evidence remain necessary for definitive identification and accurate regulatory quantification (Fisher *et al.,*2022).

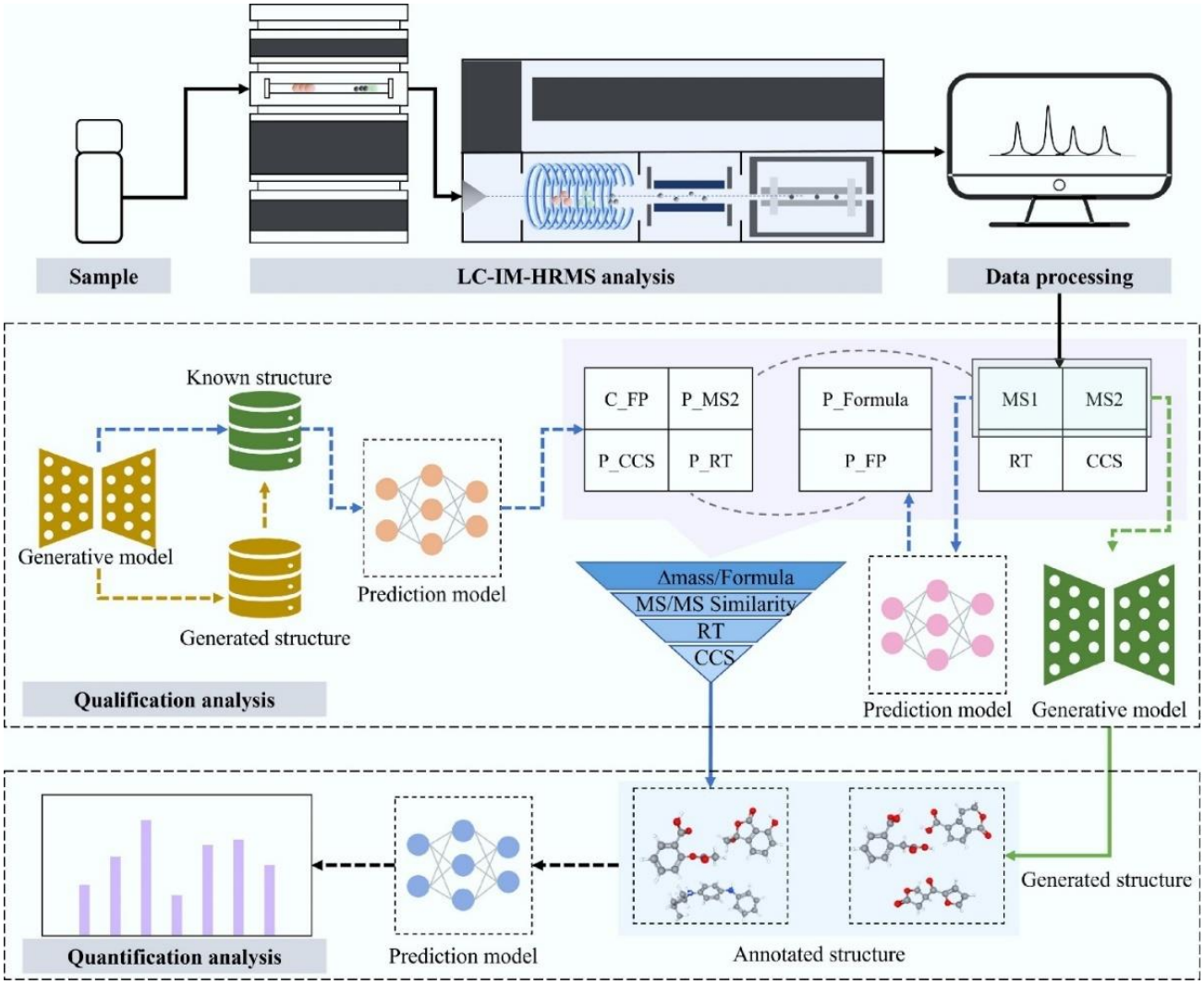

Figure 8: AI-Enabled Suspect and Non-Target Screening: From Unresolved HRMS Features to Prioritized and Semi-Quantified Emerging Pollutants

The figure illustrates an integrated LC– or GC–HRMS workflow in which raw environmental data undergo feature detection, alignment, blank filtering, isotope- and adduct-pattern recognition, and molecular-formula assignment. Suspect and unknown features are subsequently evaluated through spectral-library matching, in-silico fragmentation, retention-time and collision-cross-section prediction, AI-assisted candidate ranking, and de novo structure elucidation. When authentic standards are unavailable, predicted response

factors, ionization efficiencies, or MS/MS-based models can provide semi-quantitative concentration estimates, with identification confidence and quantitative uncertainty reported transparently.

Environmental Fate, Ecotoxicity, and Risk Prioritization

Artificial intelligence can integrate molecular structure, environmental conditions, monitoring observations, and biological responses into a unified framework for assessing emerging pollutants. As shown in Figure 9, raw chemical and biological data are first normalized and transformed into comparable features, after which informative variables are selected to reduce dimensionality and overfitting. Regression predicts continuous endpoints, whereas classification assigns chemicals to persistence, toxicity, or risk categories. Cross-validation, independent testing, and explainable methods such as SHAP, partial-dependence plots, and individual conditional-expectation curves are essential for identifying influential predictors and interpreting model outputs (Cui *et al.,*2026; Wu *et al.,*2022). Environmental-fate models use fingerprints, functional groups, ionization, polarity, molecular size, hydrogen bonding, and hydrophobicity together with pH, temperature, sunlight, microbial activity, redox state, and organic carbon. These inputs support prediction of hydrolysis, photolysis, biodegradation, degradation half-life, and persistence, as well as solubility, vapour pressure, volatilization, adsorption, and air–water, octanol–water, soil–water, or sediment–water partitioning. Hybrid mechanistic–ML approaches are preferable because these processes are interdependent: strong sorption may reduce mobility while limiting bioavailability and biodegradation (Uluseker *et al.,*2025).

Transport predictions combine chemical properties with rainfall, flow, land use, groundwater chemistry, dissolved organic carbon, and hydrogeology to estimate plume migration and contaminant transfer among air, water, soil, and sediment. Interpretable models can reveal the environmental drivers governing forecasts rather than returning unexplained concentrations, as demonstrated for pentachlorophenol transport in groundwater (Rad *et al.,*2024). Bioaccumulation models similarly relate lipophilicity, ionization, metabolism, organism traits, and exposure conditions to bioconcentration or trophic-transfer potential, supporting early prioritization of persistent and mobile chemicals.

AI also predicts wastewater-treatment removal and remediation efficiency. Molecular descriptors and operating variables can be related to removal by activated sludge, sorption, membranes, chlorination, ozonation, ultraviolet irradiation, and advanced oxidation. Choi et al. (2025) used physicochemical descriptors to model removal patterns for 149 pharmaceuticals and personal-care products across treatment trains. Related models optimize adsorbent selection, membrane properties, oxidant dose, reaction time, energy consumption, and process combinations, although facility-specific conditions and external validation remain necessary (Qiu *et al.,*2025). For ecotoxicity, supervised learning, graph models, and deep neural networks predict acute mortality, immobilization, and growth inhibition in fish, algae, and aquatic invertebrates (Takata *et al.,*2020). Chronic endpoints including reproduction, development, behaviour, population fitness, and multigenerational effects are more context-dependent. Nevertheless, models trained on structural descriptors, high-throughput bioassays, toxicogenomics, molecular initiating events, and adverse-outcome pathways can screen endocrine disruption and other mechanistic hazards. Genotoxicity can be predicted using fingerprints, descriptors, and structural alerts, but applicability domains and confirmatory assays remain essential (Fan *et al.,*2018; Wu *et al.,*2022).

Mixture toxicity requires models capable of learning concentration-dependent, synergistic, antagonistic, and shared-mode-of-action effects. Chen et al. (2022) combined 5,720 laboratory tests involving 100 contaminants with observations from more than 900 field sites, illustrating how AI can extend risk estimation beyond experimentally tested mixtures. Source attribution can then use chemical fingerprints and spatial–temporal metadata to distinguish industrial, agricultural, domestic, and wastewater inputs (Dávila-Santiago *et al.,*2022). Final prioritization should integrate persistence, mobility, bioaccumulation, exposure, acute and chronic toxicity, mixture effects, treatment resistance, uncertainty, and ecological vulnerability. Explainability should test chemical plausibility rather than decorate predictions: feature importance can identify fate or toxicity drivers, while causal estimands and counterfactual simulations can evaluate how reducing a contaminant, changing treatment conditions, or interrupting a source pathway may alter ecological outcomes. These analyses should report uncertainty and avoid extrapolation beyond the training domain. Independent-site validation, applicability-domain checks, causal interpretation, and counterfactual analysis are required before models guide monitoring or remediation decisions (Cui *et al.,*2026).

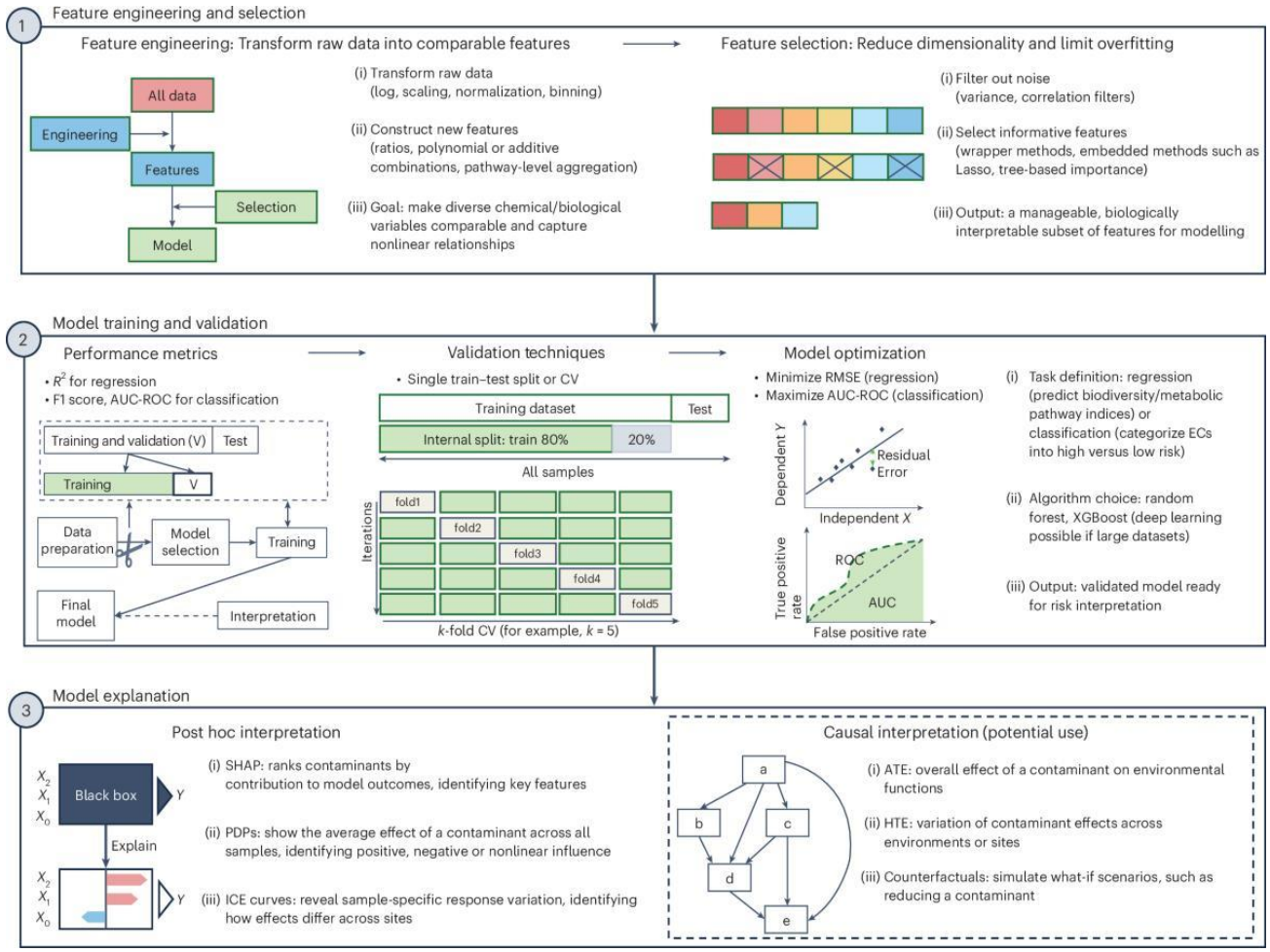

Figure 9: Machine-Learning Integration of Chemical, Environmental, and Biological Data for Ecological-Risk Assessment of Emerging Contaminants.

The figure illustrates a machine-learning framework for converting high-dimensional chemical, environmental, and biological datasets into interpretable ecological-risk predictions. Following feature engineering and selection, predictive models are trained, independently validated, optimized, and interpreted using explainable-AI and causal-analysis methods. The resulting evidence supports the identification of influential contaminants, biological-response pathways, exposure drivers, and priority environmental risks requiring monitoring or management.

Pharmaceutical Applications

Pharmaceutical Analysis, Quality Control, and Impurity Profiling

Artificial intelligence is transforming pharmaceutical quality control by converting spectroscopic, chromatographic, mass-spectrometric, and manufacturing signals into decisions about identity, potency, uniformity, solid state, and stability. Near-infrared, Raman, mid-infrared, fluorescence, and terahertz spectra contain overlapping information on API concentration, moisture, particle size, excipient composition, and crystal form. After preprocessing, partial least-squares regression, support-vector machines, random forests, and neural networks can perform rapid classification or quantitative prediction without destroying the sample. This capability is valuable for low-dose products, where excipient bands obscure the API signal. Harms et al. (2019) showed that in-line NIR and Raman spectroscopy could monitor a low-drug-load formulation during continuous manufacturing, supporting blend-potency and content-uniformity assessment. Similar models can classify batches and detect process drift. Their reliability, however, depends on representative calibration sets, instrument-transfer testing, drift monitoring, and independent validation across material lots and operating conditions.

For incoming materials, AI compares spectral fingerprints of APIs, excipients, botanicals, and packaging components with authenticated references. Fourier-transform NIR coupled with kernel extreme learning machines differentiated closely related *Rhodiola* raw materials, illustrating how nonlinear learning can resolve subtle compositional differences (Li & Su, 2018). Anomaly-detection models can flag substitution, dilution, moisture damage, contamination, or supplier changes, although compendial identity tests

remain necessary for regulatory confirmation. Counterfeit and substandard medicines often differ from genuine products in API content, excipients, coatings, or manufacturing consistency. Portable NIR or Raman devices combined with machine learning can screen intact dosage forms. Awotunde et al. (2022) demonstrated that models trained on laboratory mixtures of acetaminophen and diluents distinguished genuine, substandard, and falsified formulations. Libraries must represent legitimate product variability to prevent authentic medicines from being misclassified.

AI-assisted dissolution testing links process data with drug-release behaviour. Models may combine NIR spectra, compression force, particle-size distribution, porosity, hardness, and coating characteristics to predict dissolution profiles. Galata et al. (2021) developed machine-learning surrogate models for real-time release testing of sustained-release tablets. This approach can reduce destructive testing, but predictions must remain anchored to validated compendial methods and be challenged with variations in formulation, processing, and storage. Polymorphs, hydrates, solvates, cocrystals, and amorphous fractions influence solubility, manufacturability, and shelf stability. Low-frequency Raman spectroscopy is sensitive to lattice vibrations and crystal packing. Inoue et al. (2019) demonstrated transmission low-frequency Raman quantification of crystalline polymorphs in tablets. Machine learning can enhance classification and detect transformations during processing or storage, although powder X-ray diffraction and thermal analysis remain important for orthogonal confirmation.

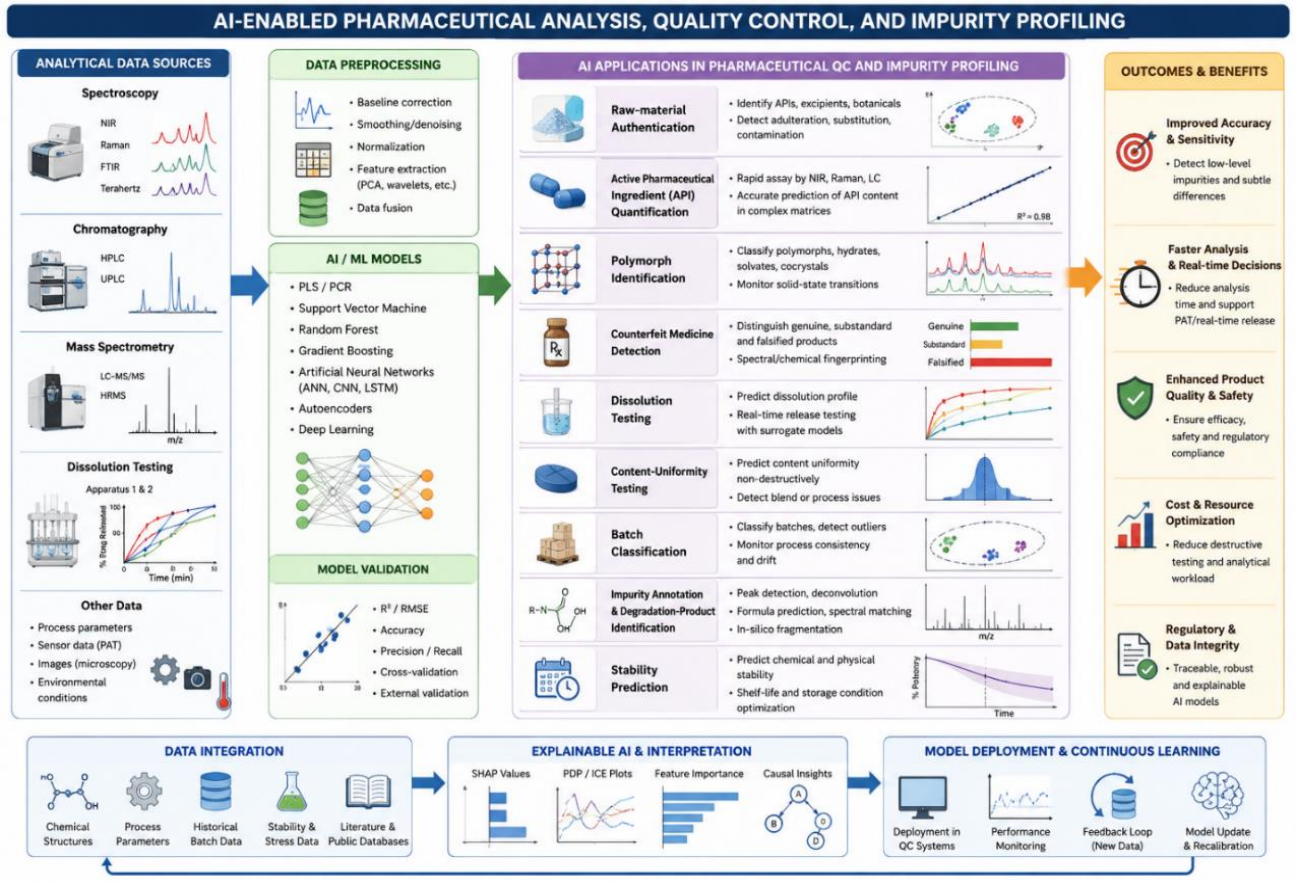

Figure 10: Integrated AI Workflow for Pharmaceutical Analysis, Quality Control, and Impurity Profiling

Chromatography and mass spectrometry remain indispensable for assay, related-substance testing, impurity annotation, and degradation-product identification. AI supports peak detection, deconvolution, isotope recognition, retention prediction, spectral matching, and candidate ranking. Bonciarelli et al. (2021) combined LC–HRMS with MassChemSite and WebChembase to automate identification of lansoprazole degradants formed under multiple stress conditions. Definitive assignments still require authentic standards, mass balance, and complementary structural evidence. Predictive systems can anticipate degradation before long-term studies are complete. Reaction rules, molecular descriptors, reactivity calculations, and machine learning can prioritize hydrolysis, oxidation, photolysis, isomerization, and drug–excipient reactions. A multi-company Zeneth benchmark compared predictions with experimental profiles for 25 drug substances and found improved prediction from curated reaction knowledge (Hemingway *et al.*,2024). These tools can guide forced-degradation design, method development, packaging selection, and safety

assessment. Stability prediction also addresses crystallization, phase separation, aggregation, and dissolution changes. Han et al. (2019) trained algorithms using 646 solid-dispersion stability records represented by molecular and formulation descriptors. Models can prioritize drug–polymer combinations and reduce experimental burden, but shelf-life assignment still requires validated stability-indicating methods and appropriate storage studies. AI is therefore most useful as an evidence-integration layer that accelerates analysis while preserving traceability, uncertainty reporting, applicability-domain assessment, and regulatory confirmation. Prospective validation across instruments, sites, products, and time is essential before these models support routine release or regulatory decisions.

The figure illustrates an integrated AI-assisted pharmaceutical quality-control workflow in which data from spectroscopy, chromatography, mass spectrometry, dissolution testing, process sensors, and stability studies are preprocessed and analyzed using chemometric and machine-learning models. The framework supports raw-material authentication, API quantification, polymorph identification, counterfeit-medicine detection, dissolution and content-uniformity prediction, batch classification, impurity and degradation-product annotation, and stability assessment. Model validation, explainable AI, continuous performance monitoring, and recalibration improve analytical reliability, traceability, product quality, and regulatory decision-making.

Formulation Development, Drug Delivery, and Manufacturing

Artificial intelligence is shifting pharmaceutical formulation from trial-and-error experimentation toward data-driven design. Molecular descriptors, solid-state properties, excipient characteristics, processing conditions, and experimental outcomes can be integrated to predict critical quality attributes and identify promising formulations before extensive laboratory testing. Supervised learning predicts formulation performance, whereas Bayesian optimization, active learning, and generative models can recommend subsequent experiments and refine the design space (Bannigan *et al.,*2021; Bao *et al.,*2023). During preformulation, models estimate aqueous solubility, pH-dependent dissolution, permeability, and formulation-dependent bioavailability from molecular structure, ionization, crystal properties, and composition. These predictions guide salt selection, amorphous dispersions, lipid systems, nanosuspensions, and precipitation-inhibiting excipients. Compatibility models can combine chemical descriptors, thermal analysis, spectroscopy, and stability records to identify drug–excipient combinations vulnerable to hydrolysis, oxidation, adsorption, or phase separation. Nevertheless, predictions require experimental confirmation within relevant chemical, storage, and gastrointestinal conditions (Bannigan *et al.,*2021; Bao *et al.,*2023).

For tablets and controlled-release systems, AI relates polymer grade, drug-to-polymer ratio, porosity, compression force, coating thickness, and particle-size distribution to hardness, disintegration, dissolution, and release kinetics. Such models capture nonlinear interactions that conventional response-surface methods may miss. In polymeric long-acting injectables, machine learning has predicted experimental release profiles and guided selection of new drug–polymer combinations, demonstrating its value for sustained-delivery design (Bannigan *et al.,*2023). Emulsions, liposomes, nanoparticles, and hydrogels create larger design spaces because formulation and processing jointly determine particle size, polydispersity, surface charge, drug loading, encapsulation efficiency, stability, and release. In microfluidic liposome production, validated models predicted liposome formation, size, and the flow-rate ratio required to achieve a desired product (Eugster *et al.,*2024). Hydrogel models can relate polymer identity, crosslinking density, swelling, degradation, mechanics, and stimulus responsiveness to release behaviour, although standardized datasets remain limited (Neguț & Bita, 2023).

Optimization should address competing objectives rather than maximizing one endpoint. Greater drug loading may increase particle size or burst release, while improved encapsulation may reduce permeability or manufacturability. Multi-objective models can balance potency, dissolution, particle attributes, release duration, bioavailability, stability, safety, and process robustness. Combining in vitro data with physiologically based or data-driven models may strengthen translation to systemic exposure, but external validation remains essential. AI can also forecast aggregation, crystallization, emulsion separation, hydrogel degradation, or loss of release control from composition, moisture, temperature, packaging, particle properties, and time-dependent measurements. Digital twins extend this approach by maintaining a dynamic virtual representation of a product and process. An AI-enabled digital twin for milling and spray drying of solidified nanosuspensions integrated experimental data, neural networks, and physicochemical modelling to optimize outcomes relevant to solubility enhancement (Davidopoulou & Ouranidis, 2022).

In manufacturing, process analytical technology supplies real-time NIR, Raman, imaging, particle-size, temperature, pressure, torque, and flow measurements. Machine-learning soft sensors convert these streams into estimates of blend uniformity, moisture, API concentration, granule properties, tablet quality, or dissolution. Galata et al. (2021) combined NIR spectra, compression force, and particle-size distribution to predict tablet dissolution in real time, illustrating how PAT can support earlier detection of process drift. Continuous manufacturing connects unit operations through material flow, residence-time distributions, automated control, and continuous quality monitoring. A

hybrid digital twin of a continuous direct-compression line combined first-principles, discrete-element, residence-time, scale, and data-driven models for product and process design (Moreno-Benito *et al.,*2022). When validated PAT measurements and predictive models adequately assure critical quality attributes, real-time release testing can reduce reliance on delayed offline tests. Model lifecycle governance, data integrity, uncertainty reporting, calibration maintenance, and change control nevertheless remain indispensable (Markl *et al.,*2020).

Overall, AI can unite formulation design, drug-delivery performance, and manufacturing control within a continuous learning framework. Its principal value lies in prioritizing experiments, revealing nonlinear relationships, balancing competing objectives, and connecting process variables with clinically relevant performance. Reliable deployment requires representative datasets, prospective validation, mechanistic plausibility, explainability, and continued expert oversight throughout product development.

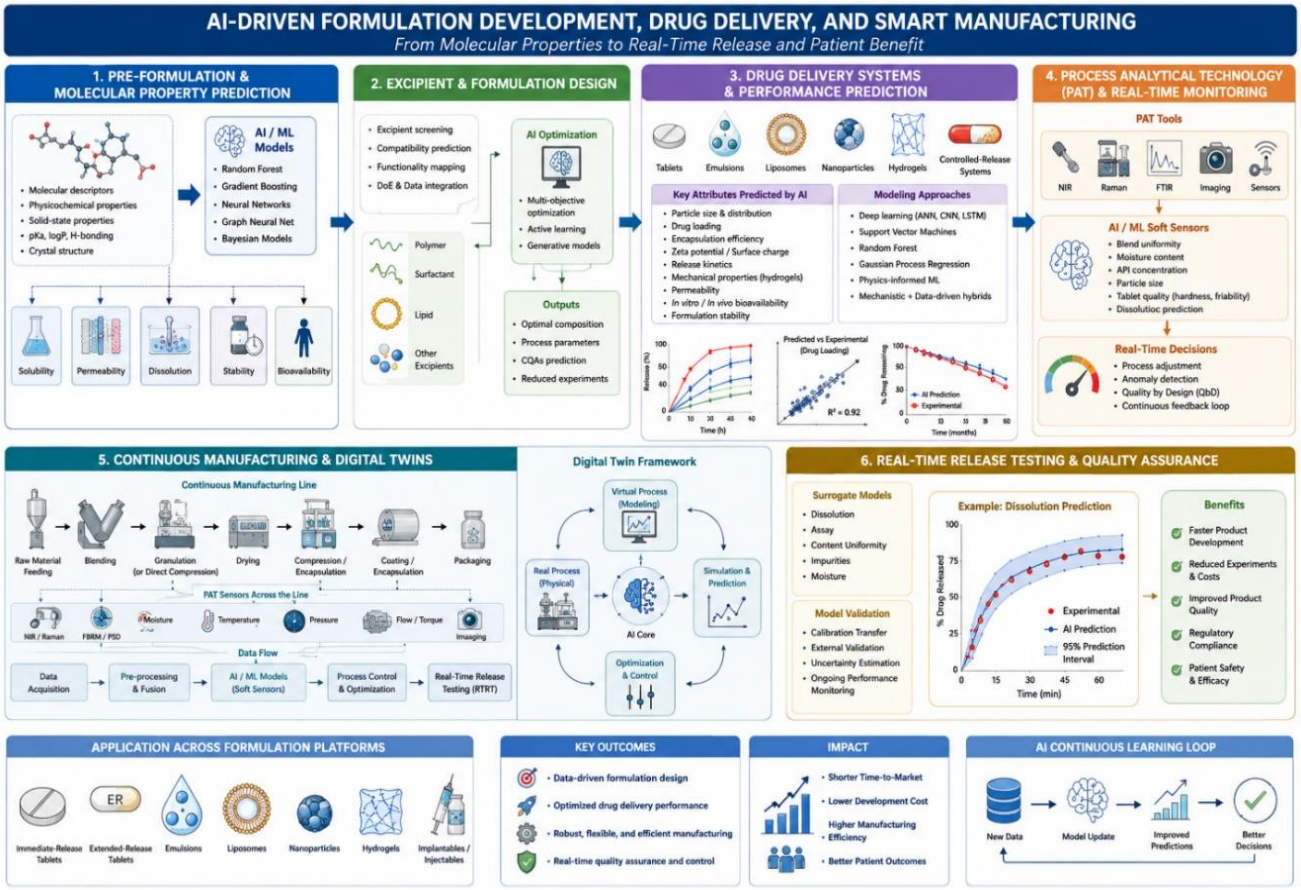

Figure 11: AI-Driven Integration of Formulation Design, Drug-Delivery Optimization, and Smart Pharmaceutical Manufacturing.

The figure illustrates an integrated AI framework that connects preformulation data, molecular and physicochemical properties, excipient selection, and process variables with the optimization of tablets, emulsions, liposomes, nanoparticles, hydrogels, and controlled-release systems. Machine-learning models predict solubility, dissolution, permeability, bioavailability, particle size, drug loading, encapsulation efficiency, release kinetics, and formulation stability. Process analytical technology, digital twins, continuous manufacturing, and real-time release testing enable continuous monitoring, process control, quality assurance, and model updating, supporting faster development, reduced experimentation, improved product consistency, and enhanced patient safety.

Drug Discovery, Molecular Design, and ADMET Prediction

Artificial intelligence is increasingly integrated throughout drug discovery, from identifying disease-relevant targets to optimizing candidate molecules for efficacy, selectivity, pharmacokinetics, and safety. Machine-learning systems can combine genomic variation, transcriptomics, proteomics, protein–protein interactions, disease pathways, phenotypic screening, scientific literature, and clinical observations to prioritize targets with stronger biological and therapeutic evidence. These methods are particularly valuable when relationships among genes, proteins, pathways, and disease phenotypes are nonlinear or distributed across heterogeneous datasets. Nevertheless, predicted targets

must still be evaluated for causal relevance, tissue specificity, tractability, safety, and the availability of appropriate experimental models (Vamathevan *et al.,*2019).

Once a target has been selected, virtual screening reduces extremely large chemical libraries to experimentally manageable candidate sets. Ligand-based models prioritize molecules according to similarity, pharmacophores, or learned activity patterns, whereas structure-based approaches evaluate molecular docking, protein–ligand complementarity, binding-site interactions, and predicted binding energies. AI can accelerate these workflows by replacing expensive calculations with surrogate models, rescoring docking poses, identifying active compounds missed by conventional scoring functions, and selecting chemically diverse candidates. However, performance depends on protein-structure quality, protonation and tautomer states, conformational sampling, and whether the screening library lies within the model's applicability domain (Sadybekov & Katritch, 2023). Quantitative structure–activity relationship models connect molecular characteristics with experimentally measured potency or biological response. Traditional QSAR relies on physicochemical descriptors, structural fragments, and fingerprints, while contemporary models learn directly from SMILES strings, molecular graphs, three-dimensional conformations, or multimodal combinations. Classification models distinguish active from inactive compounds, whereas regression models predict continuous endpoints such as IC_{50}, EC_{50}, K_i, or Kd. Modern graph neural networks represent atoms as nodes and chemical bonds as edges, enabling models to learn local chemical environments and molecular-level relationships without depending entirely on manually engineered descriptors (Nguyen *et al.,*2021; Vamathevan *et al.,*2019). Drug–target interaction and binding-affinity prediction extend activity modelling by jointly representing both the compound and its potential protein target. DeepDTA demonstrated that convolutional neural networks could learn binding-affinity relationships directly from SMILES strings and amino-acid sequences, without requiring experimentally determined three-dimensional complexes (Öztürk *et al.,*2018). GraphDTA subsequently represented compounds as molecular graphs and combined graph neural networks with protein-sequence models, improving the representation of atomic connectivity for affinity estimation (Nguyen *et al.,*2021). Such models can prioritize experimentally testable drug–target pairs, but random data splitting may exaggerate performance when similar compounds or proteins occur in both training and test sets.

Protein-language models provide another route to target characterization and molecular interaction prediction. Trained on very large collections of amino-acid sequences, these models learn evolutionary, structural, and functional patterns without requiring labels for every sequence. Their embeddings can support protein-function annotation, mutation-effect prediction, binding-site analysis, target comparison, and drug–target modelling. ESMFold demonstrated that a large protein-language model could infer full atomic-level structures directly from primary sequences, expanding access to structural hypotheses for proteins lacking experimental structures (Lin *et al.,*2023). These predictions are valuable for screening and hypothesis generation, although low-confidence or flexible regions require careful interpretation and experimental confirmation. Generative molecular design shifts AI from evaluating existing molecules to proposing new chemical structures. Recurrent neural networks, variational autoencoders, generative adversarial networks, reinforcement-learning systems, transformers, and diffusion models can generate molecules conditioned on potency, selectivity, novelty, synthetic accessibility, or ADMET constraints. Grisoni et al. (2021) combined generative AI with microfluidic on-chip synthesis to design and experimentally evaluate novel liver X receptor agonists, illustrating how computational generation can be connected with automated synthesis. However, optimizing a single predicted endpoint may generate chemically unstable, synthetically inaccessible, or toxic compounds; therefore, generative design should use multi-objective constraints and uncertainty-aware experimental feedback.

During lead optimization, AI models simultaneously evaluate potency, selectivity, solubility, permeability, metabolic stability, synthetic feasibility, and structural novelty. Active learning can select the most informative compounds for synthesis and testing, while Bayesian optimization can balance exploration of new chemical space with exploitation of known structure–activity relationships. Drug repurposing applies related methods to approved or investigational compounds, integrating drug targets, disease genes, pathways, phenotypes, contraindications, and clinical evidence. TxGNN used a medical knowledge graph and graph foundation model to rank potential indications and contraindications, including for diseases with limited established treatments (Huang *et al.,*2024). ADMET prediction is critical because promising potency does not ensure clinically useful exposure or safety. Machine-learning models estimate intestinal absorption, permeability, plasma-protein binding, tissue distribution, blood–brain barrier penetration, cytochrome P450 metabolism, transporter interactions, clearance, half-life, and routes of excretion. They also predict solubility, lipophilicity, chemical stability, and medicinal-chemistry liabilities. ADMETlab 3.0 integrates multitask graph-based modelling with molecular descriptors to predict broad physicochemical, pharmacokinetic, and toxicity endpoints while providing uncertainty estimates for decision support (Fu *et al.,*2024). Predictions should guide compound prioritization rather than replace validated in vitro, in vivo, or clinical pharmacokinetic studies.

Safety modelling includes hepatotoxicity, cardiotoxicity, mutagenicity, carcinogenicity, reproductive toxicity, skin sensitization, and organ-specific adverse effects. Off-target prediction identifies unintended protein interactions that may cause toxicity or reveal additional therapeutic mechanisms. Drug–drug interaction models further evaluate whether co-administered agents alter metabolism, transport, exposure, efficacy, or adverse-event risk. Zitnik et al. (2018) introduced Decagon, a graph-convolutional framework integrating protein–protein, drug–protein, and drug–drug relationships to predict specific polypharmacy side effects. Such predictions can prioritize high-risk combinations, but they may reflect prescribing patterns, reporting bias, missing negative evidence, or confounding clinical variables (Zitnik *et al.,*2018). Ultimately, AI should function as an experimentally connected decision-support system rather than an autonomous substitute for medicinal chemistry and pharmacology. Robust implementation requires chemically realistic data splitting, prospective testing, applicability-domain assessment, uncertainty estimation, interpretability, and validation against newly synthesized compounds. Models should also be examined for activity cliffs, imbalanced labels, assay heterogeneity, data leakage, and overrepresentation of well-studied targets and chemical scaffolds. By combining molecular graphs, protein sequences, three-dimensional structures, biological networks, and ADMET evidence, AI can reduce unproductive experiments and improve candidate prioritization, but reliable translation depends on iterative cycles of prediction, synthesis, biological testing, and model refinement (Sadybekov & Katritch, 2023; Vamathevan *et al.,*2019).

Critical Challenges, Research Gaps, and Future Directions

Data Limitations, Reproducibility, and Generalizability

Despite rapid advances in AI-driven analytical and molecular modelling, high reported accuracy does not necessarily indicate scientific reliability or real-world applicability. Model performance is fundamentally constrained by the quality, diversity, provenance, and representativeness of the underlying data. Environmental and pharmaceutical datasets are frequently assembled from individual laboratories, published studies, public repositories, or proprietary databases that differ in experimental design and analytical quality. Consequently, models may learn laboratory-specific patterns, database conventions, or sampling biases rather than transferable relationships between chemical structure, analytical signals, biological activity, and environmental behaviour (Zhu *et al.,*2023).

Small datasets remain a major limitation because obtaining experimentally validated chemical, toxicological, pharmacokinetic, or spectroscopic measurements is expensive and time-consuming. In high-dimensional datasets, the number of molecular descriptors, spectral variables, mass-spectrometric features, or image pixels may greatly exceed the number of independent samples. Complex neural networks can therefore fit noise or chance correlations while appearing accurate during internal validation. Small test sets also produce unstable performance estimates, making differences between competing models difficult to interpret without confidence intervals or repeated evaluation (Jiang *et al.,*2025; Zhu *et al.,*2023). Class imbalance creates an additional problem. Active compounds, toxic contaminants, degradation products, counterfeit samples, or adverse biological responses may represent only a small fraction of the available observations. A classifier can consequently achieve high overall accuracy by predicting the majority class while failing to identify the scientifically important minority class. Accuracy should therefore be supplemented with sensitivity, specificity, balanced accuracy, precision, recall, F1 score, precision–recall curves, and class-specific calibration. Resampling and synthetic augmentation may improve numerical balance, but they must be performed only within the training data to avoid leakage and cannot replace genuine experimental diversity (Jiang *et al.,*2025).

Limited chemical diversity further restricts generalizability. Frequently used datasets may be dominated by drug-like molecules, commercially available compounds, well-studied pollutants, or common pharmaceutical scaffolds. Rare elements, highly ionizable chemicals, complex natural products, polymers, transformation products, and structurally unusual contaminants are often underrepresented. Kretschmer et al. (2025) demonstrated that commonly used small-molecule datasets do not uniformly cover known biomolecular structure space. Models trained on such collections may perform well for compounds similar to the training set but become unreliable at the boundaries of chemical space, where novel discovery is most valuable. Reference labels are also less certain than they commonly appear. Biological activity can vary with assay format, cell line, species, exposure period, concentration range, endpoint definition, and statistical threshold. Solubility, toxicity, degradation, adsorption, and bioavailability values may be measured under different pH, temperature, ionic-strength, matrix, or redox conditions. Values reported as inactive may instead represent untested concentration ranges, measurements below detection limits, or failed experiments. Combining these records without harmonization introduces label noise and may teach models relationships that reflect methodological differences rather than intrinsic chemical properties.

Inconsistent sample preparation presents a parallel challenge for spectroscopy, chromatography, mass spectrometry, and sensor-based modelling. Extraction solvent, filtration, dilution, storage, particle size, moisture, derivatization, chromatographic

conditions, ionization mode, collision energy, and spectral preprocessing can all alter the recorded signal. If preparation methods are correlated with sample classes, a model may classify procedural artefacts rather than the intended chemical attribute. Raw data should therefore be accompanied by complete metadata describing sampling, preparation, instrument settings, quality-control procedures, preprocessing operations, and reference methods (Wilkinson *et al.,*2016; Zhu *et al.,*2023). Instrument and laboratory variability can cause substantial distribution shifts. Spectrometers from different vendors may differ in wavelength alignment, resolution, detector response, illumination geometry, and signal-to-noise ratio. Chromatographic and mass-spectrometric data vary with column condition, source contamination, calibration, gradient profile, instrument tuning, and software version. Even instruments of the same model may not produce identical measurements. Workman (2018) emphasized that successful calibration transfer depends on instrument reproducibility, reference-value accuracy, mathematical standardization, and repeatable sample presentation. A model validated on one device or laboratory should not therefore be assumed to transfer to another without calibration-transfer assessment or independent testing.

Missing metadata makes these shifts difficult to identify or correct. Public datasets may contain a chemical identifier and endpoint while omitting batch, sampling location, analytical platform, operator, date, replicate structure, uncertainty, or detection limit. These variables are essential for determining whether observations are independent and whether model errors arise from chemistry, instrumentation, geography, or time. Adoption of FAIR principles—making data findable, accessible, interoperable, and reusable—would improve dataset integration and machine actionability. FAIR implementation also requires persistent identifiers, structured metadata, provenance, versioning, licensing, and documentation, not merely depositing an unexplained spreadsheet online (Wilkinson *et al.,*2016). Proprietary databases create another research gap. Industrial datasets may contain valuable negative experiments, failed formulations, inactive compounds, long-term stability results, and uncommon chemical structures that are absent from publications. However, licensing restrictions and commercial confidentiality prevent independent researchers from reconstructing datasets or testing reported models. Public literature is also affected by publication bias toward successful experiments and positive outcomes. Privacy-preserving collaboration, federated learning, standardized data-sharing agreements, and release of carefully anonymized benchmark subsets may permit broader model development without disclosing commercially sensitive information.

Data leakage is among the most serious causes of inflated performance. Leakage occurs when related compounds, replicate spectra, measurements from the same sample, analytical batch, patient, sampling site, or experimental series appear in both training and test sets. It can also occur when normalization, feature selection, augmentation, or missing-value imputation is performed before data splitting. In molecular modelling, random splits may place closely related analogues or identical scaffolds on both sides of the split, allowing structural memorization. Wallach and Heifets (2018) showed that several ligand-classification benchmarks rewarded memorization rather than genuine generalization. More recent leakage-reduced splitting approaches similarly demonstrate the importance of grouping related observations before model evaluation (Joeres *et al.,*2025).

Evaluation strategies must reflect the intended application. Chemical models may require scaffold, cluster, or temporal splits; analytical models may require batch-, instrument-, or laboratory-level separation; and environmental models may require site- or region-level holdouts. Hyperparameter optimization should be confined to training and validation data through nested cross-validation, leaving the final test set untouched. Performance should be compared with simple baselines and reported across repeated splits rather than presenting only the best-performing run. Applicability domains and uncertainty estimates are particularly important when predictions concern chemicals unlike those represented during training (Joeres *et al.,*2025; Zhu *et al.,*2023). Reproducibility is further weakened by inadequate methodological reporting and limited code availability. Studies may omit preprocessing parameters, descriptor definitions, software versions, random seeds, hyperparameter-search spaces, class-balancing procedures, dataset exclusions, or exact train–test assignments. Without these details, apparently identical implementations can generate different results. Open-source workflows such as QSPRpred illustrate how serializable datasets, preserved preprocessing steps, documented model configurations, and deployable prediction pipelines can improve reproducibility and practical transferability (van den Maagdenberg *et al.,*2024).

Future studies should prioritize prospective and independent validation rather than repeatedly benchmarking new algorithms on familiar datasets. Models should be tested on data produced after model development and, where possible, by independent laboratories, instruments, geographic regions, or research teams. Multicentre benchmark datasets should include negative results, rare chemical classes, realistic matrices, technical replicates, uncertainty estimates, and standardized metadata. Progress will depend less on increasingly complex architectures than on better experimental design, transparent reporting, open and versioned resources, leakage-resistant evaluation, and continuous monitoring after deployment. Only models that retain accuracy under realistic chemical,

instrumental, spatial, and temporal shifts can be considered genuinely generalizable.

Interpretability, Uncertainty, and Regulatory Readiness

Complex neural networks and ensemble models can achieve strong predictive performance while providing limited insight into how particular chemical structures, spectral regions, process variables, or biological features influence an output. Explainable-AI methods such as SHAP values, feature attribution, counterfactual explanations, and attention maps can reveal influential inputs, but these post hoc explanations do not automatically establish chemical or biological mechanisms. Explanations should therefore be evaluated against established toxicological pathways, physicochemical principles, analytical evidence, and expert knowledge rather than accepted solely because they are visually convincing (Jiménez-Luna *et al.,*2020).

Uncertainty is frequently reported inadequately or reduced to an uncalibrated probability score. A model may assign high confidence to an incorrect prediction, particularly when the input differs from its training distribution. Reliable systems should distinguish uncertainty arising from experimental noise from uncertainty caused by insufficient model knowledge and should report calibrated probabilities, prediction intervals, or confidence categories for individual predictions. Because uncertainty methods can behave differently across chemical datasets and applications, their calibration must be evaluated rather than assumed (Rasmussen *et al.,*2023). Applicability domains must also be defined explicitly. Predictions for new chemical classes, environmental matrices, instruments, manufacturing conditions, or biological populations may be unreliable when these inputs fall outside the model's structural or experimental domain. Out-of-distribution detection, similarity assessment, leverage analysis, ensemble disagreement, and distance-based methods can identify such cases and trigger expert review or additional testing. OECD guidance requires regulatory QSAR models to have a defined endpoint, unambiguous algorithm, applicability domain, demonstrated robustness and predictivity, and mechanistic interpretation where possible (OECD, 2024).

Regulatory readiness further requires complete traceability and auditability. Model documentation should preserve data provenance, inclusion and exclusion decisions, preprocessing steps, software versions, model architecture, hyperparameters, validation results, user access, prediction history, and reasons for model updates. Audit trails must allow regulators to reconstruct how an environmental-risk classification, pharmaceutical-quality decision, or safety prediction was produced. The intended context of use should be specified because the evidence required for an exploratory screening model differs from that required for batch release, clinical development, or regulatory hazard assessment. Cybersecurity and data integrity are equally important because corrupted sensor data, unauthorized database changes, compromised software libraries, or manipulated model files can alter regulated decisions. Risk controls should include access management, secure data storage, version control, backup and recovery procedures, verification of third-party software, anomaly detection, and protection of connected laboratory and manufacturing systems. The NIST AI Risk Management Framework emphasizes governance, measurement, monitoring, documentation, and continuous management of AI-related risks throughout deployment (Tabassi, 2023).

Models must therefore be controlled throughout their lifecycle rather than validated only once. Changes in instruments, formulations, chemical populations, sampling locations, analytical methods, or manufacturing conditions can cause model drift. Predetermined acceptance criteria, periodic performance reviews, change-control procedures, recalibration rules, independent revalidation, human oversight, and formal decommissioning plans are necessary. EMA recommends risk-based management from early development through deployment and decommissioning, including monitoring for performance degradation, while the FDA's 2025 draft guidance proposes assessing model credibility in relation to a clearly defined context of use (European Medicines Agency, 2024; U.S. Food and Drug Administration, 2025). Ultimately, regulatory acceptance will depend less on algorithmic complexity than on evidence that a model is scientifically valid, reproducible, interpretable, secure, and fit for its intended purpose. AI outputs are most likely to be accepted when they form part of a transparent weight-of-evidence framework supported by independent validation, uncertainty estimates, applicability-domain checks, confirmatory experiments, and accountable human decision-making.

Multimodal Foundation Models and Autonomous Scientific Systems

Multimodal foundation models represent a major future direction for AI-driven environmental and pharmaceutical research. Rather than learning from a single input type, these systems can align molecular structures, spectra, chromatograms, biological assays, protein sequences, images, process measurements, and scientific text within a shared representation. Large-scale self-supervised pretraining may reduce dependence on manually labelled datasets and enable transfer to chemical identification, toxicity prediction, formulation design, and molecular-property estimation. For example, DreaMS learned transferable molecular representations from millions of unannotated tandem mass spectra, demonstrating the potential of domain-specific foundation models for analytical chemistry (Bushuiev *et al.,*2026). Bidirectional spectrum–structure modelling could further connect instrumental measurements with molecular discovery. Spectrum-to-structure systems

infer molecular formulas, substructures, fingerprints, or complete candidate structures from MS, NMR, IR, or Raman data. Spec2Mol demonstrated end-to-end generation of candidate molecular structures directly from tandem mass spectra, including compounds absent from reference databases (Litsa *et al.,*2023). Conversely, structure-to-spectrum models such as MassFormer predict fragmentation spectra from molecular graphs, potentially expanding spectral-library coverage and enabling consistency checks between proposed structures and experimental signals (Young *et al.,*2024).

Knowledge graphs can organize relationships among chemicals, targets, pathways, analytical evidence, environmental sources, toxicity outcomes, formulations, and manufacturing processes. When combined with retrieval-augmented generation, they can provide language models with traceable external evidence instead of relying exclusively on memorized training information. Chemistry-focused question-answering systems have already demonstrated retrieval across embedded knowledge graphs, while tool-augmented agents such as ChemCrow connect language models with chemical databases, structure converters, reaction predictors, and computational tools (Zhou *et al.,*2023; Bran *et al.,*2024). These approaches may improve chemical reasoning, although retrieved evidence, database versions, and reasoning steps must remain visible and auditable. Autonomous laboratories extend AI from prediction to closed-loop experimentation. Robotic systems can prepare samples, conduct reactions, operate instruments, analyse results, and select subsequent experiments. Active learning identifies experiments expected to provide the greatest information, whereas Bayesian optimization balances exploration of uncertain regions with exploitation of promising conditions. A mobile robotic chemist completed 688 experiments over eight days using Bayesian search, illustrating how automation can navigate multidimensional experimental spaces (Burger *et al.,*2020). Coscientist further showed that a tool-enabled language model could plan, optimize, and execute chemical experiments through automated laboratory interfaces (Boiko *et al.,*2023).

Digital twins can provide continuously updated virtual representations of analytical instruments, formulations, treatment systems, or manufacturing lines. By combining mechanistic equations, sensor data, and machine learning, they can simulate interventions, forecast failures, and optimize operating conditions before changes are applied physically. An AI-supported pharmaceutical digital twin has already been demonstrated for modelling and optimizing milling and spray-drying processes used to manufacture solidified nanosuspensions (Davidopoulou & Ouranidis, 2022). Federated learning could complement these systems by allowing laboratories, companies, or monitoring networks to train shared models without centrally exchanging confidential raw data, although heterogeneous local datasets and privacy risks remain important challenges (Huang *et al.,*2023). Human-in-the-loop governance must remain central to autonomous science. Experts should approve high-risk experiments, inspect unexpected predictions, evaluate mechanistic plausibility, and intervene when models encounter out-of-distribution samples. Edge AI may support portable environmental sensors, point-of-care devices, and manufacturing instruments by processing data locally, reducing latency, connectivity dependence, and unnecessary data transfer. However, compact models must still meet requirements for calibration, cybersecurity, traceability, and lifecycle monitoring.

Finally, scientific autonomy should not be achieved at the expense of environmental sustainability. Training and repeatedly tuning large foundation models can consume substantial computational energy. Future systems should therefore favour efficient architectures, transfer learning, model compression, carbon-aware computing, reusable pretrained models, and reporting of energy and hardware requirements. Autonomous laboratories should also optimize solvent use, experimental scale, waste generation, and energy consumption alongside scientific performance. Sustainable, human-supervised systems that unite multimodal evidence, grounded reasoning, robotic experimentation, privacy-preserving collaboration, and efficient computing are more likely to produce reliable and socially responsible scientific advances (Strubell *et al.,*2019).

CONCLUSION

Artificial intelligence is transforming analytical and molecular sciences by creating a direct computational connection between experimental measurements, molecular representations, biological responses, environmental behaviour, pharmaceutical performance, and decision-making. This review demonstrates that the field has progressed from conventional chemometric approaches, including PCA and PLSR, toward nonlinear machine learning, deep neural networks, graph-based models, transformers, self-supervised learning, multimodal architectures, foundation models, and generative AI. These developments have expanded the capacity to analyse high-dimensional spectra, chromatograms, mass-spectrometric features, sensor responses, molecular graphs, protein information, toxicological endpoints, formulation variables, and manufacturing data. Nevertheless, classical chemometrics remains highly relevant, particularly when datasets are small, structured, or require transparent interpretation. The most scientifically defensible strategy is therefore not the unrestricted replacement of established methods, but their integration with appropriately selected AI models.

In environmental applications, AI enables more rapid and comprehensive detection, classification, quantification, and prioritization of contaminants.

Machine-learning models can resolve overlapping analytical signals, assist suspect and non-target screening, rank candidate structures, identify transformation products, attribute pollution sources, and support semi-quantification when authentic standards are unavailable. When analytical information is integrated with molecular descriptors, environmental conditions, bioassays, and monitoring data, AI can further predict degradation, persistence, mobility, adsorption, bioaccumulation, wastewater-treatment removal, ecotoxicity, mixture effects, and ecological risk. These capabilities can move environmental analysis beyond determining which chemicals are present toward assessing where they originate, how they behave, and which substances require urgent monitoring or intervention. Within pharmaceutical science, AI supports an equally broad range of activities across the product and molecular lifecycles. Spectroscopy, chromatography, mass spectrometry, dissolution testing, and process-monitoring data can be used for raw-material authentication, API quantification, counterfeit detection, polymorph classification, content-uniformity testing, impurity annotation, degradation-product identification, and stability assessment. Molecular and formulation data can guide excipient selection, solubility enhancement, particle engineering, drug loading, encapsulation efficiency, release kinetics, permeability, bioavailability, and formulation stability. Process analytical technology, digital twins, continuous manufacturing, and real-time release testing additionally provide opportunities for adaptive process control and more consistent pharmaceutical quality. At the drug-discovery level, graph neural networks, protein-language models, generative systems, and knowledge graphs can assist target identification, virtual screening, binding-affinity prediction, drug repurposing, lead optimization, ADMET estimation, off-target analysis, and drug–drug interaction assessment. The convergence of environmental and pharmaceutical applications is a central contribution of this review. Both fields depend on the interpretation of complex chemical mixtures, high-dimensional instrumental measurements, molecular structures, biological effects, and uncertain outcomes. Techniques developed for pharmaceutical impurity profiling and degradation-product identification can strengthen environmental unknown screening, while environmental approaches to mixture toxicity, source attribution, and risk prioritization can inform pharmaceutical safety evaluation. Multimodal models that integrate spectra, structures, biological assays, metadata, and scientific text may consequently provide more complete and transferable predictions than isolated single-modality systems.

However, technological sophistication does not ensure scientific validity. Small and imbalanced datasets, narrow chemical-space coverage, uncertain labels, inconsistent sample preparation, missing metadata, instrument variability, proprietary databases, data leakage, inadequate reporting, and limited external validation remain major barriers. Black-box predictions, poorly calibrated probabilities, undefined applicability domains, and weak uncertainty reporting further limit trust and regulatory acceptance. Models intended for environmental or pharmaceutical decision-making must therefore be evaluated using chemically and operationally meaningful data splits, independent laboratories, new instruments, unseen molecular scaffolds, realistic matrices, prospective experiments, and clearly defined contexts of use. Future progress should prioritize FAIR and carefully curated multimodal datasets, shared benchmarks, open and version-controlled workflows, explainable modelling, uncertainty calibration, out-of-distribution detection, lifecycle monitoring, cybersecurity, and accountable human oversight. Foundation models, retrieval-augmented chemical reasoning, federated learning, edge AI, active learning, Bayesian optimization, digital twins, and autonomous laboratories offer substantial opportunities, but their implementation must remain traceable, experimentally grounded, and environmentally sustainable. Closed-loop systems in which AI proposes experiments, robotic platforms generate evidence, and validated observations update the model may ultimately accelerate both environmental assessment and pharmaceutical development. Overall, AI should be viewed as a scientifically supervised decision-support framework rather than a substitute for analytical quality control, mechanistic understanding, authentic standards, confirmatory experiments, or expert judgement. Its long-term value will depend on whether computational advances produce models that are not only accurate, but also reproducible, interpretable, uncertainty-aware, transferable, secure, and fit for their intended purpose. By integrating analytical measurements with molecular and biological knowledge, AI can support a more connected, efficient, and evidence-driven future for environmental protection, pharmaceutical innovation, and responsible chemical decision-making.

REFERENCES

- Alberts, M., Laino, T., & Vaucher, A. C. (2024). Leveraging infrared spectroscopy for automated structure elucidation. Communications Chemistry, 7, Article 268. doi:10.1038/s42004-024-01341-w
- Anisimov, D. S., Chekusova, V. P., Trul, A. A., Abramov, A. A., Borshchev, O. V., Agina, E. V., & Ponomarenko, S. A. (2021). Fully integrated ultra-sensitive electronic nose based on organic field-effect transistors. Scientific Reports, 11, Article 10683. doi:10.1038/s41598-021-88569-x
- Arturi, K., & Hollender, J. (2023). Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data. *Environmental Science & Technology, 57*(46), 18067–18079. doi: 10.1021/acs.est.3c00304

- Awotunde, O., Roseboom, N., Cai, J., Hayes, K., Rajane, R., Chen, R., Yusuf, A., & Lieberman, M. (2022). Discrimination of substandard and falsified formulations from genuine pharmaceuticals using NIR spectra and machine learning. Analytical Chemistry, 94(37), 12586–12594. doi: 10.1021/acs.analchem.2c00998
- Bae, S.-Y., Lee, J., Jeong, J., Lim, C., & Choi, J. (2021). Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints. *Computational Toxicology, 20*, Article 100178. doi: 10.1016/j.comtox.2021.100178
- Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., & Allen, C. (2021). Machine learning directed drug formulation development. Advanced Drug Delivery Reviews, 175, 113806. doi: 10.1016/j.addr.2021.05.016
- Bannigan, P., Bao, Z., Hickman, R. J., Aldeghi, M., Häse, F., Aspuru-Guzik, A., & Allen, C. (2023). Machine learning models to accelerate the design of polymeric long-acting injectables. *Nature Communications, 14*, Article 35. doi:10.1038/s41467-022-35343-w
- Bao, Z., Bufton, J., Hickman, R. J., Aspuru-Guzik, A., Bannigan, P., & Allen, C. (2023). Revolutionizing drug formulation development: The increasing impact of machine learning. Advanced Drug Delivery Reviews, 202, 115108. doi: 10.1016/j.addr.2023.115108
- Beck, A. G., Muhoberac, M., Randolph, C. E., Beveridge, C. H., Wijewardhane, P. R., Kenttämaa, H. I., & Chopra, G. (2024). Recent developments in machine learning for mass spectrometry. ACS Measurement Science Au, 4(3), 233–246. doi:10.1021/acsmeasuresciau.3c00060
- Beck, A. G., Muhoberac, M., Randolph, C. E., Beveridge, C. H., Wijewardhane, P. R., Kenttämaa, H. I., & Chopra, G. (2024). Recent developments in machine learning for mass spectrometry. *ACS Measurement Science Au, 4*(3), 233–246. doi:10.1021/acsmeasuresciau.3c00060
- Bhargava, A., Sachdeva, A., Sharma, K., Alsharif, M. H., Uthansakul, P., & Uthansakul, M. (2024). Hyperspectral imaging and its applications: A review. *Heliyon, 10*(12), Article e33208. doi: 10.1016/j.heliyon. 2024.e33208
- Biancolillo, A., & Marini, F. (2018). Chemometric methods for spectroscopy-based pharmaceutical analysis. Frontiers in Chemistry, 6, Article 576. doi:10.3389/fchem.2018.00576
- Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624, 570–578. doi:10.1038/s41586-023-06792-0
- Bonciarelli, S., Desantis, J., Goracci, L., Siragusa, L., Zamora, I., & Ortega-Carrasco, E. (2021). Automatic identification of lansoprazole degradants under stress conditions by LC-HRMS with MassChemSite and WebChembase. Journal of Chemical Information and Modeling, 61(6), 2706–2719. doi: 10.1021/acs.jcim.1c00226
- Bramer, W. M., de Jonge, G. B., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). A systematic approach to searching: An efficient and complete method to develop literature searches. *Journal of the Medical Library Association, 106*(4), 531–541. doi:10.5195/jmla.2018.283
- Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. Nature Machine Intelligence, 6, 525–535. doi:10.1038/s42256-024-00832-8
- Brandt, J., Mattsson, K., & Hassellöv, M. (2021). Deep learning for reconstructing low-quality FTIR and Raman spectra—A case study in microplastic analyses. *Analytical Chemistry, 93*(49), 16360–16368. doi: 10.1021/acs.analchem.1c02618
- Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., Roger, J.-M., Walczak, B., & Tauler, R. (2018). Chemometrics in analytical chemistry—Part II: Modeling, validation, and applications. Analytical and Bioanalytical Chemistry, 410(26), 6691–6704. doi:10.1007/s00216-018-1283-4
- Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S., & Cooper, A. I. (2020). A mobile robotic chemist. Nature, 583, 237–241. doi:10.1038/s41586-020-2442-2
- Bushuiev, R., Bushuiev, A., Samusevich, R., Brungs, C., Sivic, J., & Pluskal, T. (2026). Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Nature Biotechnology, 44, 630–640. doi:10.1038/s41587-025-02663-3
- Butler, H. J., Smith, B. R., Fritzsch, R., Radhakrishnan, P., Palmer, D. S., & Baker, M. J. (2018). Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. *Analyst, 143*(24), 6121–6134. doi:10.1039/C8AN01384E
- Cavasotto, C. N., & Scardino, V. (2022). Machine learning toxicity prediction: Latest advances by toxicity end point. *ACS Omega, 7*(51), 47536–47546. doi:10.1021/acsomega.2c05693
- Celma, A., Bade, R., Sancho, J. V., Hernández, F., Humphries, M., & Bijlsma, L. (2022). Prediction of retention time and collision cross section (CCS H+, CCS H−, and CCS Na+) of emerging contaminants using multiple adaptive regression splines. Journal of Chemical Information and Modeling, 62(22), 5425–5434. doi: 10.1021/acs.jcim.2c00847
- Chen, J., Wang, B., Huang, J., Deng, S., Wang, Y., Blaney, L., Brennan, G. L., Cagnetta, G., Jia, Q., & Yu, G. (2022). A machine-learning approach clarifies interactions between contaminants of

emerging concern. One Earth, 5(11), 1239–1249. doi: 10.1016/j.oneear.2022.10.006

- Chen, S., Xue, D., Chuai, G., Yang, Q., & Liu, Q. (2020). FL-QSAR: A federated learning-based QSAR prototype for collaborative drug discovery. Bioinformatics, 36(22–23), 5492–5498. doi:10.1093/bioinformatics/btaa1006
- Choi, J. M., Manthapuri, V., Keenum, I., Brown, C. L., Xia, K., Chen, C., Vikesland, P. J., Blair, M. F., Bott, C., Pruden, A., & Zhang, L. (2025). A machine learning framework to predict PPCP removal through various wastewater and water reuse treatment trains. Environmental Science: Water Research & Technology, 11, 481–493. doi:10.1039/D4EW00892H
- Cui, H.-L., Gao, S.-H., Wang, H.-C., Zhang, L.-Y., Luo, Y., Ying, G.-G., Sun, W.-L., Yu, Y.-J., Liang, B., & Wang, A.-J. (2026). Big data integration for environmental risk assessment of emerging contaminants. *Nature Sustainability, 9*, 196–206. doi:10.1038/s41893-025-01718
- David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: A review and practical guide. *Journal of Cheminformatics, 12*, Article 56. doi:10.1186/s13321-020-00460-5
- Davidopoulou, C., & Ouranidis, A. (2022). Pharma 4.0—Artificially intelligent digital twins for solidified nanosuspensions. Pharmaceutics, 14(10), 2113. doi:10.3390/pharmaceutics14102113
- Dávila-Santiago, E., Shi, C., Mahadwar, G., Medeghini, B., Insinga, L., Hutchinson, R., Good, S., & Jones, G. D. (2022). Machine learning applications for chemical fingerprinting and environmental source tracking using non-target chemical data. *Environmental Science & Technology, 56*(7), 4080–4090. doi: 10.1021/acs.est.1c06655
- de Juan, A., & Tauler, R. (2021). Multivariate curve resolution: 50 years addressing the mixture analysis problem—A review. Analytica Chimica Acta, 1145, 59–78. doi: 10.1016/j.aca.2020.10.051
- Dekermanjian, J. P., Shaddox, E., Nandy, D., Ghosh, D., & Kechris, K. (2022). Mechanism-aware imputation: A two-step approach in handling missing values in metabolomics. *BMC Bioinformatics, 23*, Article 179. doi:10.1186/s12859-022-04659-1
- Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., & Wang, F. (2023). A systematic study of key elements underlying molecular property prediction. Nature Communications, 14, Article 6395. doi:10.1038/s41467-023-41948-6
- Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., Dorrestein, P. C., Rousu, J., & Böcker, S. (2019). SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. Nature Methods, 16(4), 299–302. doi:10.1038/s41592-019-0344-8
- Eugster, R., Orsi, M., Buttitta, G., Serafini, N., Tiboni, M., Casettari, L., Reymond, J.-L., Aleandri, S., & Luciani, P. (2024). Leveraging machine learning to streamline the development of liposomal drug delivery systems. Journal of Controlled Release, 376, 1025–1038. doi: 10.1016/j.jconrel.2024.10.065
- European Medicines Agency. (2024). Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023.
- Fan, D., Yang, H., Li, F., Sun, L., Di, P., Li, W., Tang, Y., & Liu, G. (2018). In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. Toxicology Research, 7(2), 211–220. doi:10.1039/C7TX00259A
- Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., & Wang, H. (2022). Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence, 4*, 127–134. doi:10.1038/s42256-021-00438-4
- Feshuk, M., Kolaczkowski, L., Dunham, K., Davidson-Fritz, S. E., Carstens, K. E., Brown, J., Judson, R. S., & Paul Friedman, K. (2023). The ToxCast pipeline: Updates to curve-fitting approaches and database structure. *Frontiers in Toxicology, 5*, Article 1275980. doi:10.3389/ftox.2023.1275980
- Fisher, C. M., Peter, K. T., Newton, S. R., Schaub, A. J., & Sobus, J. R. (2022). Approaches for assessing performance of high-resolution mass spectrometry–based non-targeted analysis methods. Analytical and Bioanalytical Chemistry, 414, 6455–6471. doi:10.1007/s00216-022-04203-3
- Flanagan, A. R., Dalal, D., & Glavin, F. G. (2025). Exploring generative artificial intelligence and data augmentation techniques for spectroscopy analysis. Chemical Reviews, 125(13), 6130–6155. doi: 10.1021/acs.chemrev.4c00815
- Fu, L., Shi, S., Yi, J., Wang, N., He, Y., Wu, Z., Peng, J., Deng, Y., Wang, W., Wu, C., Lyu, A., Zeng, X., Zhao, W., Hou, T., & Cao, D. (2024). ADMETlab 3.0: An updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Research, 52(W1), W422–W431. doi:10.1093/nar/gkae236
- Galata, D. L., Könyves, Z., Nagy, B., Novák, M., Mészáros, L. A., Szabó, E., Farkas, A., Marosi, G., & Nagy, Z. K. (2021). Real-time release testing of dissolution based on surrogate models developed by machine learning algorithms using NIR spectra, compression force and particle size distribution as input data. International Journal of Pharmaceutics, 597, 120338. doi: 10.1016/j.ijpharm.2021.120338
- Gloaguen, Y., Kirwan, J. A., & Beule, D. (2022). Deep learning-assisted peak curation for large-scale

LC–MS metabolomics. *Analytical Chemistry, 94*(12), 4930–4937. doi: 10.1021/acs.analchem.1c02220

- González-Domínguez, Á., Estanyol-Torres, N., Brunius, C., Landberg, R., & González-Domínguez, R. (2024). QComics: Recommendations and guidelines for robust, easily implementable and reportable quality control of metabolomics data. *Analytical Chemistry, 96*(3), 1064–1072. doi: 10.1021/acs.analchem.3c03660
- Graff, D. E., Shakhnovich, E. I., & Coley, C. W. (2021). Accelerating high-throughput virtual screening through molecular pool-based active learning. Chemical Science, 12(22), 7866–7881. doi:10.1039/D0SC06805E
- Grisoni, F., Huisman, B. J. H., Button, A. L., Moret, M., Atz, K., Merk, D., & Schneider, G. (2021). Combining generative artificial intelligence and on-chip synthesis for de novo drug design. Science Advances, 7(24), eabg3338. doi:10.1126/sciadv. abg3338
- Guo, J., & Huan, T. (2023). Mechanistic understanding of the discrepancies between common peak-picking algorithms in liquid chromatography–mass spectrometry-based metabolomics. *Analytical Chemistry, 95*(14), 5894–5902. doi: 10.1021/acs.analchem.2c04887
- Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217. doi:10.1002/jrsm.1378
- Haddaway, N. R., Gray, C. T., & Grainger, M. (2021). Novel tools and methods for designing and wrangling multifunctional, machine-readable evidence synthesis databases. *Environmental Evidence, 10*, Article 5. doi:10.1186/s13750-021-00219-x
- Han, R., Xiong, H., Ye, Z., Yang, Y., Huang, T., Jing, Q., Lu, J., Pan, H., Ren, F., & Ouyang, D. (2019). Predicting physical stability of solid dispersions by machine learning techniques. Journal of Controlled Release, 311–312, 16–25. doi: 10.1016/j.jconrel.2019.08.030
- Harms, Z. D., Shi, Z., Kulkarni, R. A., & Myers, D. P. (2019). Characterization of near-infrared and Raman spectroscopy for in-line monitoring of a low-drug-load formulation in a continuous manufacturing process. Analytical Chemistry, 91(13), 8045–8053. doi: 10.1021/acs.analchem.8b05002
- He, H., Yan, S., Lyu, D., Xu, M., Ye, R., Zheng, P., Lu, X., Wang, L., & Ren, B. (2021). Deep learning for biospectroscopy and biospectral imaging: State-of-the-art and perspectives. Analytical Chemistry, 93(8), 3653–3665. doi: 10.1021/acs.analchem.0c04671
- Heid, E., McGill, C. J., Vermeire, F. H., & Green, W. H. (2023). Characterizing uncertainty in machine learning for chemistry. Journal of Chemical Information and Modeling, 63(13), 4012–4029. doi: 10.1021/acs.jcim.3c00373
- Helmus, R., ter Laak, T. L., van Wezel, A. P., de Voogt, P., & Schymanski, E. L. (2021). patRoon: Open-source software platform for environmental mass spectrometry based non-target screening. Journal of Cheminformatics, 13, Article 1. doi:10.1186/s13321-020-00477-w
- Hemingway, R., Baertschi, S. W., Benstead, D., Campbell, J. M., Coombs, M., Huang, Z., Khalaf, R., Ott, M. A., Ponting, D. J., Reid, D. L., Stevenson, N. G., Webb, S. J., White, A., & Zelensky, T. (2024). In silico prediction of pharmaceutical degradation pathways: A benchmarking study using the software program Zeneth. Organic Process Research & Development, 28(3), 674–692. doi: 10.1021/acs.oprd.3c00344
- Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., & Coley, C. W. (2020). Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling, 60*(8), 3770–3780. doi: 10.1021/acs.jcim.0c00502
- Hohrenk, L. L., Itzel, F., Baetz, N., Tuerk, J., Vosough, M., & Schmidt, T. C. (2020). Comparison of software tools for liquid chromatography–high-resolution mass spectrometry data processing in nontarget screening of environmental samples. Analytical Chemistry, 92(2), 1898–1907. doi: 10.1021/acs.analchem.9b04095
- Hollender, J., Schymanski, E. L., Singer, H. P., & Ferguson, P. L. (2017). Nontarget screening with high resolution mass spectrometry in the environment: Ready to go? Environmental Science & Technology, 51(20), 11505–11512. doi: 10.1021/acs.est.7b02184
- Houhou, R., & Bocklitz, T. (2021). Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. *Analytical Science Advances, 2*(3–4), 128–141. doi:10.1002/ansa.202000162.
- Huanbutta, K., Burapapadh, K., Kraisit, P., Sriamornsak, P., Ganokratanaa, T., Suwanpitak, K., & Sangnim, T. (2024). Artificial intelligence-driven pharmaceutical industry: A paradigm shift in drug discovery, formulation development, manufacturing, quality control, and post-market surveillance. *European Journal of Pharmaceutical Sciences, 203*, Article 106938. doi: 10.1016/j.ejps.2024.106938
- Huang, D., Ye, X., Zhang, Y., & Sakurai, T. (2023). Collaborative analysis for drug discovery by federated learning on non-IID data. Methods, 219, 1–7. doi: 10.1016/j.ymeth.2023.09.001
- Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., Nadkarni, G. N., Glicksberg,

B. S., Gehlenborg, N., & Zitnik, M. (2024). A foundation model for clinician-centered drug repurposing. Nature Medicine, 30, 3601–3613. doi:10.1038/s41591-024-03233-x

- Hufnagl, B., Stibi, M., Martirosyan, H., Wilczek, U., Möller, J. N., Löder, M. G. J., Laforsch, C., & Lohninger, H. (2022). Computer-assisted analysis of microplastics in environmental samples based on µFTIR imaging in combination with machine learning. Environmental Science & Technology Letters, 9(1), 90–95. doi: 10.1021/acs.estlett.1c00851
- Inoue, M., Hisada, H., Koide, T., Fukami, T., Roy, A., Carriere, J., & Heyler, R. (2019). Transmission low-frequency Raman spectroscopy for quantification of crystalline polymorphs in pharmaceutical tablets. Analytical Chemistry, 91(3), 1997–2003. doi: 10.1021/acs.analchem.8b04365
- Jiang, J., Zhang, C., Ke, L., Hayes, N., Zhu, Y., Qiu, H., Zhang, B., Zhou, T., & Wei, G. (2025). A review of machine learning methods for imbalanced data challenges in chemistry. Chemical Science, 16, 7637–7658. doi:10.1039/D5SC00270B
- Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence, 2*, 573–584. doi:10.1038/s42256-020-00236-4
- Joeres, R., Blumenthal, D. B., & Kalinina, O. V. (2025). Data splitting to avoid information leakage with DataSAIL. Nature Communications, 16, Article 3337. doi:10.1038/s41467-025-58606-8
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3, 422–440. doi:10.1038/s42254-021-00314-5
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2023). PubChem 2023 update. *Nucleic Acids Research, 51*(D1), D1373–D1380. doi:10.1093/nar/gkac956
- Kiseleva, O., Kurbatov, I., Ilgisonis, E., & Poverennaya, E. (2022). Defining blood plasma and serum metabolome by GC–MS. *Metabolites, 12*(1), Article 15. doi:10.3390/metabo12010015
- Krenn, M., Häse, F., Nigam, A. K., Friederich, P., & Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology, 1*(4), Article 045024. doi:10.1088/2632-2153/aba947
- Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W., & Böcker, S. (2025). Coverage bias in small molecule machine learning. Nature Communications, 16, Article 554. doi:10.1038/s41467-024-55462-w
- Leach, D. T., Stratton, K. G., Irvahn, J., Richardson, R., Webb-Robertson, B.-J. M., & Bramer, L. M. (2023). malbacR: A package for standardized implementation of batch correction methods for omics data. *Analytical Chemistry, 95*(33), 12195–12199. doi: 10.1021/acs.analchem.3c01289
- Li, D., Chen, Q., Ouyang, Q., & Liu, Z. (2025). Advances of Vis/NIRS and imaging techniques assisted by AI for tea processing. Critical Reviews in Food Science and Nutrition, 65(31), 7575–7593. doi:10.1080/10408398.2025.2474183
- Li, T., & Su, C. (2018). Authenticity identification and classification of Rhodiola species in traditional Tibetan medicine based on Fourier transform near-infrared spectroscopy and chemometrics analysis. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 204, 131–140. doi: 10.1016/j.saa.2018.06.004
- Liigand, J., Wang, T., Kellogg, J., Smedsgaard, J., Cech, N., & Kruve, A. (2020). Quantification for non-targeted LC/MS screening without standard substances. Scientific Reports, 10, Article 5808. doi:10.1038/s41598-020-62573-z
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123–1130. doi:10.1126/science.ade2574
- Litsa, E. E., Chenthamarakshan, V., Das, P., & Kavraki, L. E. (2023). An end-to-end deep learning framework for translating mass spectra to de novo molecules. Communications Chemistry, 6, Article 132. doi:10.1038/s42004-023-00932-3
- Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., & Anandkumar, A. (2023). Multi-modal molecule structure–text model for text-based retrieval and editing. Nature Machine Intelligence, 5(12), 1447–1457. doi:10.1038/s42256-023-00759-6
- Liu, Y., Yao, W., Qin, F., Zhou, L., & Zheng, Y. (2023). Spectral classification of large-scale blended (micro)plastics using FT-IR raw spectra and image-based machine learning. Environmental Science & Technology, 57(16), 6656–6663. doi: 10.1021/acs.est.2c08952
- Mandal, S., Paul, D., Saha, S., & Das, P. (2022). Deep learning assisted detection of toxic heavy metal ions based on visual fluorescence responses from a carbon nanoparticle array. Environmental Science: Nano, 9, 2596–2606. doi:10.1039/D2EN00077F
- Markl, D., Warman, M., Dumarey, M., Bergman, E.-L., Folestad, S., Shi, Z., Manley, L. F., Goodwin, D. J., & Zeitler, J. A. (2020). Review of real-time release testing of pharmaceutical tablets: State-of-the art, challenges and future perspective. International Journal of Pharmaceutics, 582, 119353. doi: 10.1016/j.ijpharm.2020.119353
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera,

J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., . . . Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research, 47*(D1), D930–D940. doi:10.1093/nar/gky1075

- Méndez-Lucio, O., Nicolaou, C. A., & Earnshaw, B. (2024). MolE: A foundation model for molecular graphs using disentangled attention. Nature Communications, 15, Article 9431. doi:10.1038/s41467-024-53751-y
- Mirza, A., Patiny, L., & Jablonka, K. M. (2026). End-to-end multimodal structure elucidation from raw spectra combining contrastive learning and evolutionary algorithms. *Nature Communications, 17*, Article 5013. doi:10.1038/s41467-026-73846-y
- Moreno-Benito, M., Lee, K. T., Kaydanov, D., Verrier, H. M., Blackwood, D. O., & Doshi, P. (2022). Digital twin of a continuous direct compression line for drug product and process design using a hybrid flowsheet modelling approach. International Journal of Pharmaceutics, 628, 122336. doi: 10.1016/j.ijpharm.2022.122336
- Neguț, I., & Bita, B. (2023). Exploring the potential of artificial intelligence for hydrogel development—A short review. Gels, 9(11), 845. doi:10.3390/gels9110845
- Nene, L., Flepisi, B. T., Brand, S. J., Basson, C., & Balmith, M. (2024). Evolution of drug development and regulatory affairs: The demonstrated power of artificial intelligence. *Clinical Therapeutics, 46*(8), e6–e14. doi: 10.1016/j.clinthera.2024.05.012
- Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2021). GraphDTA: Predicting drug–target binding affinity with graph neural networks. Bioinformatics, 37(8), 1140–1147. doi:10.1093/bioinformatics/btaa921
- Organisation for Economic Co-operation and Development. (2024). (Q)SAR assessment framework: Guidance for the regulatory assessment of (quantitative) structure–activity relationship models and predictions (2nd ed.). OECD Publishing. doi:10.1787/bbdac345-en
- Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: Deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829. doi:10.1093/bioinformatics/bty593
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. doi:10.1136/bmj. n71
- Pesciullesi, G., Schwaller, P., Laino, T., & Reymond, J.-L. (2020). Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. Nature Communications, 11, Article 4874. doi:10.1038/s41467-020-18671-7
- Puthongkham, P., Wirojsaengthong, S., & Suea-Ngam, A. (2021). Machine learning and chemometrics for electrochemical sensors: Moving forward to the future of analytical chemistry. *Analyst, 146*(21), 6351–6364. doi:10.1039/D1AN01148K
- Qiu, Z., Zhuang, Z., Jia, X., Li, H., Chen, C., Shen, J., & Yu, B. (2025). Machine learning-driven advances in water treatment. Journal of Environmental Chemical Engineering, 13(6), Article 120441. doi: 10.1016/j.jece.2025.120441
- Rad, M., Abtahi, A., Berndtsson, R., McKnight, U. S., & Aminifar, A. (2024). Interpretable machine learning for predicting the fate and transport of pentachlorophenol in groundwater. Environmental Pollution, 345, Article 123449. doi: 10.1016/j.envpol.2024.123449
- Rasmussen, M. H., Duan, C., Kulik, H. J., & Jensen, J. H. (2023). Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets. Journal of Cheminformatics, 15, Article 121. doi:10.1186/s13321-023-00790-0
- Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews, 10*, Article 39. doi:10.1186/s13643-020-01542-z
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences, 118*(15), Article e2016239118. doi:10.1073/pnas.2016239118
- Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence, 4*, 1256–1264. doi:10.1038/s42256-022-00580-7
- Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So you think you can PLS-DA? BMC Bioinformatics, 21(Suppl. 1), Article 2. doi:10.1186/s12859-019-3310-7
- Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. Journal of Cheminformatics, 8, Article 3. doi:10.1186/s13321-016-0115-9
- Sadybekov, A. V., & Katritch, V. (2023). Computational approaches streamlining drug discovery. Nature, 616(7958), 673–685. doi:10.1038/s41586-023-05905-z

- Sahin, F., Celik, N., Camdal, A., Sakir, M., Ceylan, A., Ruzi, M., & Onses, M. S. (2022). Machine learning-assisted pesticide detection on a flexible surface-enhanced Raman scattering substrate prepared by silver nanoparticles. ACS Applied Nano Materials, 5(9), 13112–13122. doi:10.1021/acsanm.2c02897
- Schneuing, A., Harris, C., Du, Y., Didi, K., Jamasb, A., Igashov, I., Du, W., Gomes, C., Blundell, T. L., Lio, P., Welling, M., Bronstein, M., & Correia, B. (2024). Structure-based drug design with equivariant diffusion models. Nature Computational Science, 4(12), 899–909. doi:10.1038/s43588-024-00737-x
- Schulze, B., Heffernan, A. L., Samanipour, S., Gómez Ramos, M. J., Veal, C., Thomas, K. V., & Kaserzon, S. L. (2023). Is nontarget analysis ready for regulatory application? Influence of peak-picking algorithms on data analysis. *Analytical Chemistry, 95*(50), 18361–18369. doi: 10.1021/acs.analchem.3c03003
- Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Central Science, 5(9), 1572–1583. doi:10.1021/acscentsci.9b00576
- Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environmental Science & Technology, 48(4), 2097–2098. doi:10.1021/es5002105
- Sepman, H., Malm, L., Peets, P., MacLeod, M., Martin, J., Breitholtz, M., & Kruve, A. (2023). Bypassing the identification: MS2Quant for concentration estimations of chemicals detected with nontarget LC-HRMS from MS^2 data. Analytical Chemistry, 95(33), 12329–12338. doi: 10.1021/acs.analchem.3c01744.
- Shukla, V. K., Heller, G. T., & Hansen, D. F. (2023). Biomolecular NMR spectroscopy in the era of artificial intelligence. *Structure, 31*(11), 1360–1374. doi: 10.1016/j.str.2023.09.011
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. doi: 10.1016/j.jbusres.2019.07.039
- Soares, T. A., Nunes-Alves, A., Mazzolari, A., Ruggiu, F., Wei, G.-W., & Merz, K. M., Jr. (2022). The (re)-evolution of quantitative structure–activity relationship studies propelled by the surge of machine learning methods. Journal of Chemical Information and Modeling, 62(22), 5317–5320. doi: 10.1021/acs.jcim.2c01422
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics. doi:10.18653/v1/P19-1355
- Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology. doi:10.6028/NIST.AI.100-1
- Takata, M., Lin, B.-L., Xue, M., Zushi, Y., Terada, A., & Hosomi, M. (2020). Predicting the acute ecotoxicity of chemical substances by machine learning using graph theory. Chemosphere, 238, Article 124604. doi: 10.1016/j.chemosphere.2019.124604
- Tian, T., Li, S., Fang, M., Zhao, D., & Zeng, J. (2024). MolSHAP: Interpreting quantitative structure–activity relationships using Shapley values of R-groups. Journal of Chemical Information and Modeling, 64(7), 2236–2249. doi: 10.1021/acs.jcim.3c00465
- U.S. Food and Drug Administration. (2025). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: Draft guidance for industry and other interested parties.
- Uluseker, C., Lofts, S., & Harrison, S. (2025). Advances and challenges in modelling the environmental fate and exposure of pharmaceuticals: A comprehensive review. Environmental Science: Processes & Impacts, 27, 3700–3724. doi:10.1039/D5EM00449G
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery, 18*, 463–477. doi:10.1038/s41573-019-0024-5
- van den Maagdenberg, H. W., Šícho, M., Araripe, D. A., Luukkonen, S., Schoenmaker, L., Jespers, M., Béquignon, O. J. M., González, M. G., van den Broek, R. L., Bernatavicius, A., van Hasselt, J. G. C., van der Graaf, P. H., & van Westen, G. J. P. (2024). QSPRpred: A flexible open-source quantitative structure–property relationship modelling tool. Journal of Cheminformatics, 16, Article 128. doi:10.1186/s13321-024-00908-y
- Wallach, I., & Heifets, A. (2018). Most ligand-based classification benchmarks reward memorization rather than generalization. Journal of Chemical Information and Modeling, 58(5), 916–932. doi: 10.1021/acs.jcim.7b00403
- Wang, S., Zhang, R., Li, X., Cai, F., Ma, X., Tang, Y., Xu, C., Wang, L., Ren, P., Liu, L., Wu, S., Qian, Q., & Bai, F. (2025). Recent advances in molecular representation methods and their applications in scaffold hopping. npj Drug Discovery, 2, Article 14. doi:10.1038/s44386-025-00017-2
- Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature

Machine Intelligence, 4, 279–287. doi:10.1038/s42256-022-00447-x

- Wang, Z., Chen, J., & Hong, H. (2021). Developing QSAR models with defined applicability domains on PPARγ binding affinity using large data sets and machine learning algorithms. *Environmental Science & Technology, 55*(10), 6857–6866. doi: 10.1021/acs.est.0c07040
- Wang, Z., Chen, J., & Hong, H. (2021). Developing QSAR models with defined applicability domains on PPARγ binding affinity using large data sets and machine learning algorithms. Environmental Science & Technology, 55(10), 6857–6866. doi: 10.1021/acs.est.0c07040
- Wellawatte, G. P., Gandhi, H. A., Seshadri, A., & White, A. D. (2023). A perspective on explanations of molecular prediction models. Journal of Chemical Theory and Computation, 19(8), 2149–2160. doi: 10.1021/acs.jctc.2c01235
- Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. Drug Discovery Today: Technologies, 37, 1–12. doi: 10.1016/j.ddtec.2020.11.009
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, Article 160018. doi:10.1038/sdata.2016.18
- Workman, J. J., Jr. (2018). A review of calibration transfer practices and instrument differences in spectroscopy. Applied Spectroscopy, 72(3), 340–365. doi:10.1177/0003702817736064
- Wu, X., Zhou, Q., Mu, L., & Hu, X. (2022). Machine learning in the identification, prediction and exploration of environmental toxicology: Challenges and perspectives. Journal of Hazardous Materials, 438, Article 129487. doi: 10.1016/j.jhazmat.2022.129487
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science, 9*(2), 513–530. doi:10.1039/C7SC02664A
- Xue, X., Sun, H., Yang, M., Liu, X., Hu, H.-Y., Deng, Y., & Wang, X. (2023). Advances in the application of artificial intelligence-based spectral data interpretation: A perspective. Analytical Chemistry, 95(37), 13733–13745. doi: 10.1021/acs.analchem.3c02540
- Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370–3388. doi: 10.1021/acs.jcim.9b00237
- Young, A., Wang, B., & Röst, H. (2024). Tandem mass spectrum prediction for small molecules using graph transformers. Nature Machine Intelligence, 6, 404–416. doi:10.1038/s42256-024-00816-8
- Zhang, J., Li, H., Zhang, Y., Huang, J., Ren, L., Zhang, C., Zou, Q., & Zhang, Y. (2025). Computational toxicology in drug discovery: Applications of artificial intelligence in ADMET and toxicity prediction. Briefings in Bioinformatics, 26(5), Article bbaf533. doi:10.1093/bib/bbaf533
- Zhou, X., Zhang, S., Agarwal, M., Akroyd, J., Mosbach, S., & Kraft, M. (2023). Marie and BERT—A knowledge graph embedding-based question-answering system for chemistry. ACS Omega, 8(36), 33039–33057. doi:10.1021/acsomega.3c05114
- Zhou, Z., Li, Y., Hong, P., & Xu, H. (2025). Multimodal fusion with relational learning for molecular property prediction. *Communications Chemistry, 8*, Article 200. doi:10.1038/s42004-025-01586-z
- Zhu, J.-J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: Common pitfalls and best practices. Environmental Science & Technology, 57(46), 17671–17689. doi: 10.1021/acs.est.3c00026
- Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457–i466. doi:10.1093/bioinformatics/bty294
- Zuo, J., Li, H.-S., Tang, W., Zhao, X., Cheng, M.-Y., Yang, Z., Tian, S., Li, P., Xie, X., Luo, D., & Qiu, K. (2026). Machine learning assisted single-molecule sensing towards standard-free quantification of per- and polyfluoroalkyl carboxylic acids. Nature Communications, 17, Article 3923. doi:10.1038/s41467-026-70718-3