Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST37792

Molecular Toxicity Prediction and Machine Learning

Jizhi Zhang

School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, China
10244602476@stu.ecnu.edu.cn

Abstract. With the rapid growth in the number of chemicals, conventional in vivo and in vitro toxicity tests face challenges including high cost, long experimental cycles, and ethical concerns, creating an urgent need for efficient computational toxicology approaches. This review systematically summarizes the application of machine learning to molecular toxicity prediction, with particular emphasis on two core components: algorithm selection and descriptor construction. At the algorithmic level, the principles and applicable scenarios of conventional machine-learning models, including random forest (RF), extreme gradient boosting (XGBoost), and k-nearest neighbors (KNN), as well as deep-learning methods, are introduced in detail, together with the classification framework of supervised, unsupervised, and semi-supervised learning. At the descriptor level, in view of the dependence of conventional computational descriptors on accurate molecular structures, the advantages and recent applications of mass spectrometry (MS) data as a new type of descriptor are highlighted. These applications include molecular-structure recognition (CANOPUS and SIFTER), quantitative compound analysis (an internal-standard replacement strategy based on natural organic matter background peaks and virtual molecular projection CNNs for nanoparticles), and spectrum prediction. Current studies indicate that combining mass-spectrometric descriptors with machine learning can reduce dependence on structural databases and enable rapid classification and quantitative prediction of unknown compounds, showing considerable potential for environmental monitoring and toxicity screening. Nevertheless, current models remain strongly data-dependent and are limited by insufficient interpretability and restricted applicability domains. Future work should therefore focus on expanding datasets, improving model interpretability, and strengthening cross-domain generalization.

Keywords: molecular toxicity prediction, machine learning, mass spectrometry, QSAR

1. Introduction

As of May 8, 2019, more than 150 million chemical substances had been registered in the Chemical Abstracts Service (CAS). Although the widespread use of chemicals has greatly improved human quality of life, it has also introduced potential risks to the environment and human health. Chemical pollution has been shown to threaten human survival on Earth. Many chemicals are persistent, dispersive, and biologically toxic, and through human production and daily activities they have entered the atmosphere, soil, and water. According to a report by The Lancet Commission on

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AST37792

Pollution and Health, an estimated 9 million premature deaths in 2015 were associated with air, water, and soil pollution. Water pollution, in particular, has increasingly become one of the major challenges facing humanity in the 21st century, and most water-pollution incidents are caused by chemical contamination. These data highlight the importance and urgency of rigorous health-risk assessment before compounds are produced and used on a large scale. However, only a small fraction of the tens of thousands of commercial chemicals currently have comprehensive toxicological data. Toxicity assessment still relies primarily on in vivo and in vitro biological experiments, among which aquatic toxicity testing is an important step in evaluating environmental hazards and risks. Conventional biological toxicity tests, however, cannot adequately address the high experimental cost, ethical concerns associated with extensive animal use, and the enormous number of chemicals requiring evaluation. Moreover, the number of newly developed chemicals continues to increase rapidly. Government regulators and industry therefore urgently require new methods capable of rapidly and comprehensively evaluating the toxicity of large numbers of chemicals to support their sale, use, and regulation.

The basic concept of machine learning is to provide a set of input features and allow an algorithm to automatically learn the relationship required to generate predictions. Research related to machine learning can be traced back to the early development of modern computing and entered a period of rapid growth after the 1980s. With continuous advances in computer hardware, particularly graphics processing units (GPUs) capable of performing large numbers of simple computations in parallel, numerous neural-network algorithms have emerged and have been applied to medicine, pharmaceuticals, environmental science, materials science, and other fields because of their impressive predictive performance. For example, AlphaFold2 integrates physical and biological knowledge related to protein structure through multiple-sequence alignment within a deep-learning framework and enables protein-structure prediction at near-atomic accuracy; DoDNet uses a single encoder-decoder network to address partially labeled multi-organ and tumor segmentation in abdominal CT scans. Machine-learning algorithms have thus become an important new approach for solving quantitative and qualitative analytical problems and for uncovering underlying mechanisms in scientific research. Against this background, a variety of computational toxicology methods, such as quantitative structure-activity relationship (QSAR) modeling, have been developed for rapid and efficient toxicity screening. QSAR establishes qualitative or quantitative relationships between chemical structures and molecular properties such as physicochemical characteristics, biological activity, toxicological effects, and environmental behavior using molecular descriptors and modeling algorithms. Machine-learning approaches that associate molecular structures with properties through predictive models can also be regarded as a form of QSAR. QSAR methods have been recognized by industrial and governmental regulatory agencies in many countries and are now widely used in practical applications. Machine-learning-based toxicity prediction and structural-alert approaches have become important components of modern in silico chemical-safety assessment [1].

In 2007, the OECD proposed five principles for QSAR model development: (a) a defined endpoint; (b) an unambiguous algorithm; (c) a defined domain of applicability; (d) appropriate measures of goodness-of-fit, robustness, and predictivity; and (e) a mechanistic interpretation, if possible.

With the development of computer science, the methods used in QSAR to relate chemical structures to biological activities have gradually shifted from multiple linear regression equations to machine-learning algorithms with self-learning capability. These include high-performing conventional machine-learning algorithms, such as RF, SVM, XGBoost, and KNN, which use arrays of feature variables as descriptor inputs in a manner analogous to multiple linear regression, as well

DOI: 10.54254/2755-2721/2026.AST37792

as end-to-end deep-learning algorithms with stronger information-extraction capabilities. Representative examples include convolutional neural networks (CNNs) that identify molecular features from image-like representations, graph neural networks (GNNs) that learn atomic-node information, and recurrent neural networks (RNNs) that learn molecular string representations. Algorithmic innovation has not only improved information extraction and provided new perspectives for mechanistic interpretation, but has also indirectly promoted the diversification of molecular descriptors. Benchmark studies such as MoleculeNet further demonstrate the importance of molecular representations, dataset quality, and appropriate evaluation procedures in molecular machine learning [2].

In addition to algorithm selection, descriptor selection also has a major influence on model performance. Molecular-structure descriptors used in machine-learning studies of molecular properties can generally be divided into experimental descriptors and computational descriptors. Large numbers of experimental descriptors are costly to obtain, whereas computational descriptors require accurate knowledge of molecular structure; therefore, no single type of descriptor can perfectly cover all application scenarios. In recent years, researchers have proposed a variety of innovative descriptors. Such new descriptors can provide alternative perspectives for model interpretation and broaden the range of model applications. Begcr and Wilkes processed spectral data as descriptors and used NMR features to construct QSDAR (quantitative spectroscopic data-activity relationship) models for predicting the binding of organic compounds to proteins. Compared with computational descriptors, mass-spectrometric descriptors do not require prior knowledge of the exact molecular structure and can be obtained using relatively straightforward analytical procedures. Current direct-ionization mass-spectrometry techniques can operate in open environments with little or no sample pretreatment, which is favorable for extending mass-spectrometric descriptors to practical applications.

2. Application of machine learning to the biological toxicity prediction of organic chemicals

2.1. Introduction to common machine-learning algorithms

The early development of QSAR is often associated with Hansch's work in the 1960s. In that framework, differences in biological activity among homologous compounds were related to physicochemical quantities such as hydrophobicity, electronegativity, and steric effects, which were initially treated as independent variables. The first models were fitted by multiple linear regression, and nonlinear terms were later added when simple linear relations were insufficient. A representative form is: $\log(I/C) = a\pi + b\sigma + cEs + k$.

In Equation (1), I/C is the measured activity term. The symbols π, σ, and Es describe hydrophobic, electronic, and steric contributions, respectively; a, b, and c are fitted coefficients, while k is the constant term.

At a more general level, a QSAR model can be written as: $y = f(X)$.

This notation separates the problem into three elements. X contains the information supplied to the model and may range from conventional numerical descriptors to images, molecular graphs, or sequence-like representations; y is the property or endpoint to be estimated; and f is the relationship learned from the data. Modern QSAR work therefore differs from early regression mainly in the flexibility of f and in the form of X. As larger datasets and more expressive algorithms have become available, machine-learning models have increasingly been used to learn this mapping and to extend QSAR beyond simple linear relationships.

DOI: 10.54254/2755-2721/2026.AST37792

2.2. Machine-learning models

2.2.1. Classification of machine-learning models

From the viewpoint of label availability, machine-learning methods used in molecular prediction can be grouped into supervised, unsupervised, and semi-supervised learning. Supervised learning is the most direct setting: each training example is paired with a target value, and model parameters are adjusted so that predicted outputs approach the experimental labels. Linear regression, support vector machines, random forests, and neural networks all fall into this category when trained on labeled data. Unsupervised learning addresses a different question. Rather than fitting a known endpoint, it searches the data for recurring structure, for example by K-means clustering or principal component analysis. This is useful for examining sample similarity, compressing high-dimensional descriptors, and revealing groups or trends that are not specified in advance. Semi-supervised learning combines the two settings by using a relatively small labeled set together with a larger unlabeled set. Its usefulness depends on assumptions about the geometry of the data: nearby samples should usually have similar outputs (smoothness), decision boundaries should preferably avoid dense regions (low density), and samples lying on the same low-dimensional manifold should have compatible labels (manifold consistency). In practice, an initial model is learned from the labeled samples and the unlabeled distribution is then used as additional information, with the aim of improving prediction when experimental labels are scarce.

2.2.2. Decision-tree-based random forest and extreme gradient boosting models

2.2.2.1. Random forest model

Random forest (RF) builds an ensemble of CART decision trees rather than relying on a single tree. Given a training set D = {(x_i, y_i)}_(i=1)^n, each tree is fitted to a bootstrap sample drawn with replacement from D. Because of this resampling step, an individual bootstrap set contains about 63.2% of the unique training observations. Randomness is introduced again at each split: instead of examining all p variables, the algorithm considers only m_try randomly chosen candidates, with common choices including √p and p/3. The split that best improves a criterion such as information gain or the Gini index for classification, or reduces mean squared error for regression, is then selected. After B trees have been grown independently, their predictions are combined by voting in classification or by averaging in regression.

These two sources of randomness - bootstrap sampling and restricted feature selection - make the trees less correlated with one another and reduce the variance that often makes an individual decision tree unstable. The observations omitted from a given bootstrap sample form that tree's out-of-bag (OOB) set, so prediction error can be estimated without reserving a separate validation subset. OOB samples also provide a practical route to feature-importance analysis: when one descriptor is permuted, the resulting increase in OOB error indicates how much the model depended on that descriptor. RF is attractive for heterogeneous molecular data because it can capture nonlinear effects and feature interactions, tolerates outliers reasonably well, and usually does not require prior scaling. Its limitations are equally important: the ensemble is less transparent than a linear model, and a small noisy dataset can still lead to overfitting.

DOI: 10.54254/2755-2721/2026.AST37792

2.2.2.2. Extreme Gradient Boosting model

Extreme Gradient Boosting (XGBoost) follows the gradient boosting decision tree (GBDT) idea but modifies both the optimization and the control of model complexity. Trees are added sequentially, and each new tree is chosen to correct the residual error left by the current ensemble. XGBoost evaluates this update through an objective that includes the prediction loss together with a regularization penalty: $L^{(t)} = \Sigma_i\, l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$. A second-order Taylor expansion of the loss supplies both first- and second-order information. This allows the weight of each leaf and the gain from a proposed split to be evaluated analytically, rather than relying only on the first-order gradient as in a simpler boosting formulation.

The implementation also includes several mechanisms that are useful in practical modeling, including column-wise processing, default directions for missing values, learning-rate shrinkage, and feature subsampling. Together with regularization, these choices make it possible to fit complex nonlinear descriptor-property relationships while limiting unnecessary tree growth. Performance, however, is sensitive to hyperparameter settings. Parameters such as max_depth, eta, subsample, and colsample_bytree are therefore normally selected through cross-validation, often combined with grid search or Bayesian optimization rather than fixed a priori.

2.2.3. Distance-based K-nearest neighbors model

K-nearest neighbors (KNN) takes a much more local view of prediction. It stores the training data and postpones most computation until a new sample is queried, which is why it is usually described as a lazy-learning method. The working assumption is simple: points that occupy nearby positions in feature space are likely to have related target values. For a new sample, distances to the training observations are calculated, the K closest samples are retained, and the output is obtained from their labels - by majority voting for classification and by an arithmetic or distance-weighted mean for regression.

For regression tasks, KNN prediction can be expressed as: $\hat{y}(x) = (1/K)\, \Sigma_{(i \in N_K(x))}\, y_i$

where $N_K(x)$ is the index set of the K nearest neighbors of x. To distinguish the contributions of neighbors at different distances, distance-weighted KNN can be used:$\hat{y}(x) = \Sigma_{(i \in N_K(x))}\, w_i\, y_i / \Sigma_{(i \in N_K(x))}\, w_i$, $w_i = 1/d(x, x_i)^p$

Here, $d(x, x_i)$ denotes the distance from the query x to a neighboring sample x_i, and p controls how strongly distance changes the weight. The distance definition therefore becomes part of the model itself. Euclidean and Manhattan distances are common choices; Chebyshev distance uses the largest coordinate-wise difference, whereas Mahalanobis distance additionally incorporates the covariance among descriptors. The latter can be useful when features are correlated or when the geometry of the descriptor space is not well represented by an ordinary Euclidean metric.

The value of K determines how local the prediction remains. A very small K follows individual observations closely and is consequently sensitive to noise, whereas a large K averages over a wider neighborhood and may smooth away meaningful local structure. Cross-validation is commonly used to find a compromise. Feature scaling is also essential because a descriptor with a large numerical range can dominate a distance calculation; Z-score standardization or Min-Max normalization is therefore generally performed before fitting KNN. The method is straightforward and can represent nonlinear class boundaries without an explicit parametric model, but this simplicity has costs. Prediction requires comparison with the stored training set, giving O(n) query complexity, and the notion of a meaningful 'nearest' sample becomes less reliable as dimensionality increases.

DOI: 10.54254/2755-2721/2026.AST37792

2.3. Descriptor calculation and selection

Spectral analysis provides a natural source of data for machine learning because a measured spectrum contains far more information than is normally used in a single manually derived parameter. Analytical measurements may be chromatographic, optical-spectroscopic, mass-spectrometric, or based on other spectroscopic techniques, and their raw signals rarely correspond directly to the property of interest. Traditional analysis therefore relies on a sequence of processing steps, prior knowledge, and expert interpretation. Machine learning can be inserted into this sequence by treating the processed or even relatively raw signal as a feature representation and learning its association with the target property. This does not remove the need for chemical understanding, but it can reduce some hand-designed processing steps and reveal signal patterns that are difficult to specify in advance. Mass spectrometry (MS) is especially relevant in this context. It is routinely used to characterize unknown molecular samples in organic synthesis, pharmacokinetics, forensic analysis, explosives detection, and remote gas analysis. In electron-ionization MS, molecules are ionized and fragmented, the resulting ions are separated before detection, and the recorded spectrum describes ion abundance as a function of mass-to-charge ratio (m/z). Such a spectrum can therefore be treated not only as an identification pattern but also as a high-dimensional descriptor for subsequent modeling.

3. Application of machine learning to spectral feature extraction

3.1. Molecular structure identification

Dührkop et al. [3] developed CANOPUS for situations in which direct class prediction fails because the test compound belongs to a subclass absent from the spectral training data. Their solution combines high-resolution LC-MS/MS fragmentation spectra with a two-stage computational architecture. An SVM first produces probabilistic molecular fingerprints, which are then passed to a DNN that assigns the compound to one or more of 2,497 classes. The DNN was trained with 4.1 million known structural fingerprints, and probabilistic perturbations were introduced during training to mimic the errors expected in fingerprint prediction. This design matters because the model does not simply memorize a spectral library: if the molecular formula can be inferred from the spectrum, a compound may still be classified even when neither its exact structure nor its subclass appears in the training spectra. Under the subclass-exclusion test reported in the study, CANOPUS reached 97% prediction accuracy. For well-represented classes, the mean MCC was 0.744, compared with 0.692 for CSI:FingerID, 0.482 for MetFrag, and 0.542 for spectral-library KNN; the micro-averaged MCC for sparse classes was 0.603. In practical terms, the method shifts the task from exact database matching toward class-level inference from generalized structural information. A complementary strategy was reported by Picache et al. [4] in SIFTER. Rather than relying on an MS/MS library, SIFTER uses three experimentally measured ion-mobility/mass-spectrometric quantities - collision cross section, mass-to-charge ratio, and Kendrick mass defect - to classify compounds that remain unidentified in non-targeted analysis. Four random-forest classifiers are arranged hierarchically across the ClassyFire levels of kingdom, superclass, class, and subclass; the prediction from one level is passed forward as an input to the next. The output therefore includes both a final assignment and probabilities for alternative assignments, which is useful when the evidence is ambiguous. On an ideal test set, complete four-level classification exceeded 99%. The performance fell to roughly 80% in complex biological samples, however, mainly because the training data did not cover the full chemical space and nearest-neighbor bias

DOI: 10.54254/2755-2721/2026.AST37792

became more important. The comparison highlights a recurring limitation of data-driven identification: database independence at prediction time does not eliminate the need for broad and representative training data.

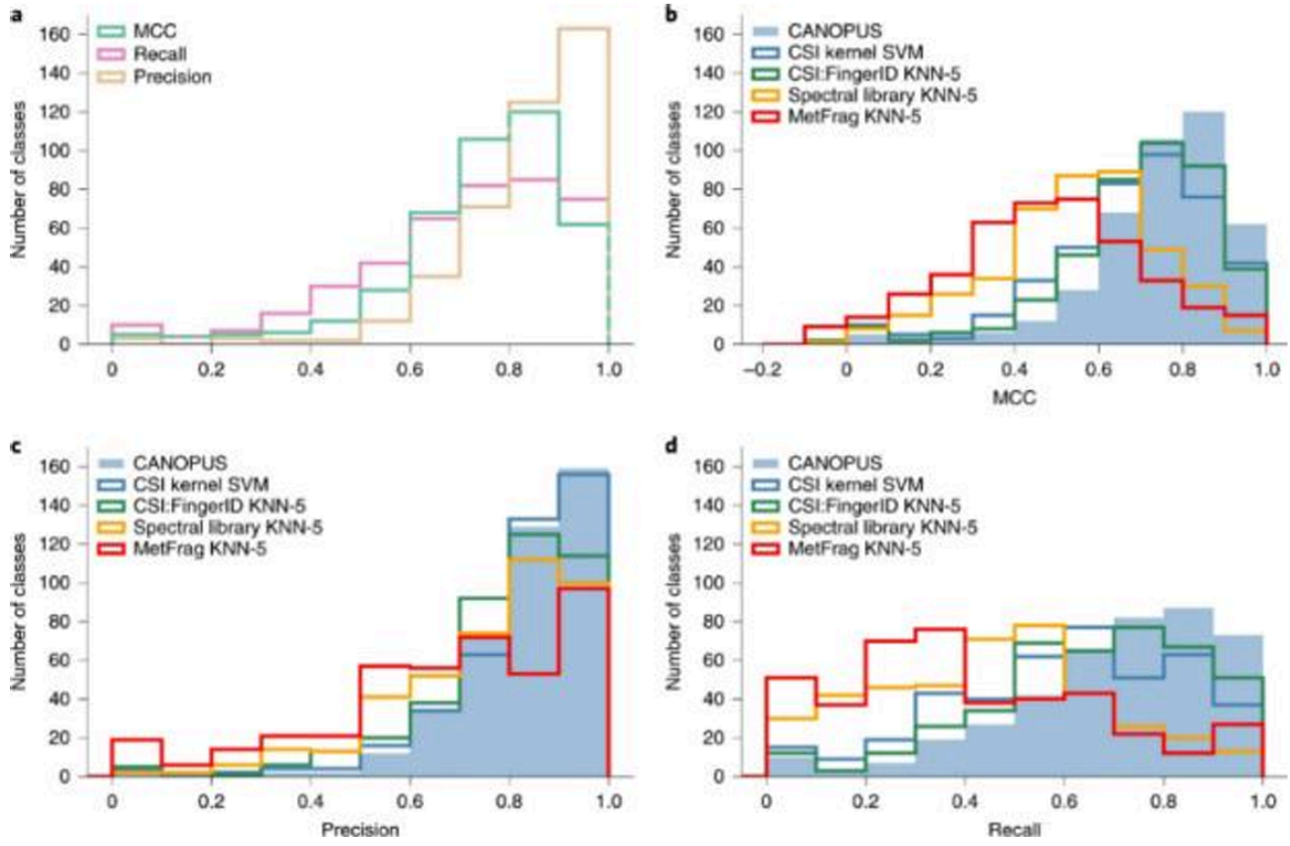


Figure 1. CANOPUS prediction results

3.2. Quantitative analysis of compounds

Baek et al. [5] approached quantitative mass-spectrometric analysis from a different direction. Stable-isotope-labeled (SIL) internal standards give reliable correction of matrix effects but are expensive, so the authors asked whether information already present in the sample matrix could serve a similar role. Full-scan HRMS-Orbitrap data from water samples were used, and fragment peaks associated with natural organic matter (NOM) were treated as candidate surrogate standards. The spectra were divided into more than 230,000 windows; for each window, its integrated signal was compared with the SIL internal-standard peak, and the 35 NOM windows showing the strongest correlations were retained. These inputs were then used to estimate the concentrations of five micropollutants, including sulpiride, metformin, and benzotriazole, with both conventional machine-learning methods and deep CNNs such as ResNet101, as shown in Figure 2. For the CNN representation, retention time and m/z were arranged as a 287 × 216 two-dimensional matrix. This preserves local relationships among neighboring spectral signals that would be less explicit in a one-dimensional feature vector. ResNet101 gave the best results among the four deep models examined, with an average validation R^2 of 0.84 and an average MSE of 0.26 ng/L. The residual architecture is useful here because it permits a deep network to learn weak NOM fingerprint patterns without the same degree of vanishing-gradient difficulty. The practical appeal of the strategy is economic as well

DOI: 10.54254/2755-2721/2026.AST37792

as computational: SIL standards may cost more than USD 1,000 per milligram, whereas the proposed workflow uses naturally occurring NOM peaks already present in the sample. At the same time, the method should not be interpreted as a universal replacement for internal standards. It was trained for only five compounds, its decisions are difficult to interpret chemically, and a large change in NOM composition between regions could alter the learned relationship and reduce quantitative accuracy.

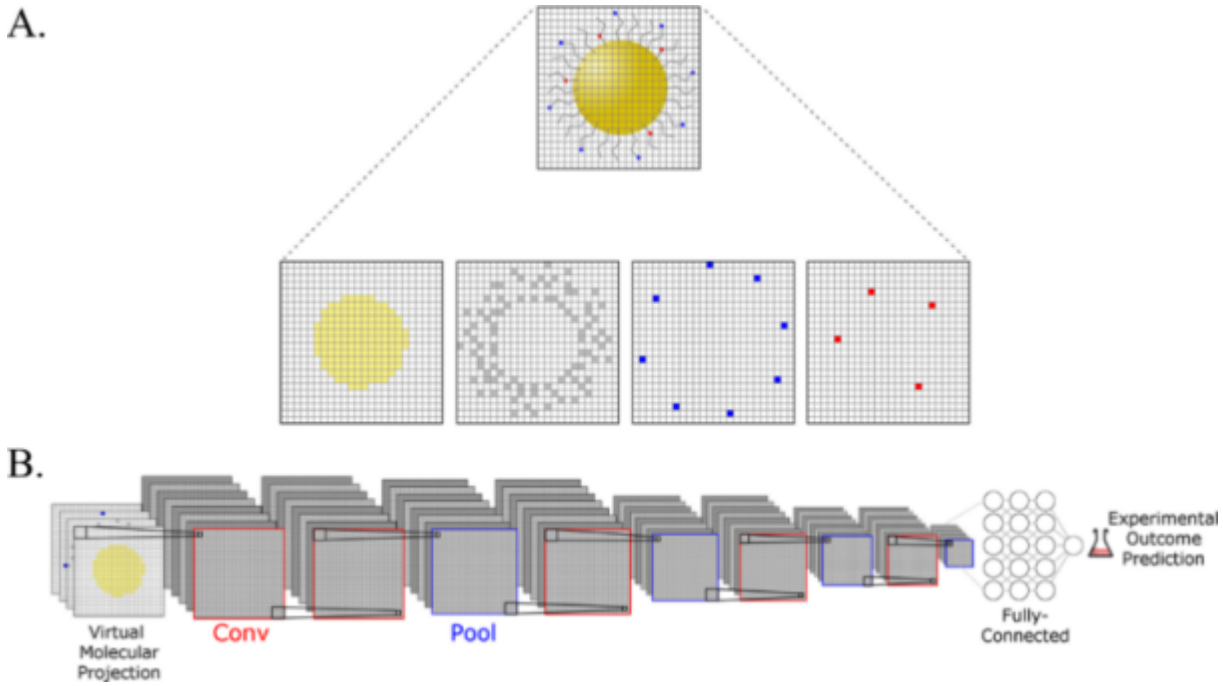


Figure 2. Virtual molecular projection and convolutional neural network (CNN) modeling scheme

Russo et al. [6] used a structural representation designed specifically for nanoparticles, where a ligand-only descriptor can omit information carried by the inorganic core, particle size, or ligand density. Their dataset contained 77 nanoparticles spanning gold, platinum, and palladium cores, diameters from 5.4 to 27.8 nm, and 18 surface ligands. Each particle was converted into a 'virtual molecular projection,' a 100 × 100 × 7 numerical array that retained information from both the core and the surface chemistry. A nine-layer CNN was then trained directly on these arrays to predict cellular uptake, lipophilicity, and zeta potential. The main distinction from a conventional descriptor workflow is that feature definition is largely learned inside the network: successive convolution and pooling operations transform the complete digitalized nanoparticle representation into higher-level structural patterns without requiring a manually selected descriptor set. Neuron-activation maximization was additionally used to examine which structural motifs activated particular neurons, linking responses to features such as core material or surface hydrophobicity and providing a limited window into the model's internal representation. Predictive performance was strongest for lipophilicity ($R^2 = 0.89$) and cellular uptake ($R^2 = 0.845$), while zeta potential was more difficult to model ($R^2 = 0.51$). These results show the value of retaining the whole nanoparticle representation, but the small dataset also defines a narrow applicability domain. Only atom types represented during training are covered, and 77 particles are far fewer than the sample numbers typical of image-recognition tasks; predictions for structurally different nanoparticles should therefore be treated cautiously.

DOI: 10.54254/2755-2721/2026.AST37792

3.3. Spectrum prediction

The CANOPUS workflow also illustrates how spectrum prediction and structural inference can be linked in a single computational chain. Dührkop et al. combined high-resolution LC-MS/MS fragmentation data with the SIRIUS + CSI:FingerID + CANOPUS framework so that an unknown metabolite could be placed into one of 2,497 ClassyFire compound classes without requiring an exact library match. After SIRIUS-related processing, probabilistic fingerprints predicted by an SVM are supplied to a DNN trained on 4.1 million structural fingerprints. Artificial perturbation of these fingerprints during training was used to represent realistic prediction noise. As a consequence, the classifier can still exploit generalized structural signals when the relevant subclass is completely absent from the spectral training set, provided that the molecular formula can be inferred from the measured spectrum. This is the central reason the method is less dependent on existing spectral and structure databases than conventional direct matching. In the stringent subclass-exclusion experiment, direct prediction produced only 8% accuracy, whereas CANOPUS achieved 97%. The mean MCC was 0.744 for well-represented classes and the micro-averaged MCC for sparse classes was 0.603; the corresponding reported values for CSI:FingerID, MetFrag, and spectral-library KNN were 0.692, 0.482, and 0.542. Beyond class assignment, the approach can also reduce annotation-homogenization bias during molecular-network propagation, which broadens the biological information that can be extracted from an otherwise unidentified spectrum.

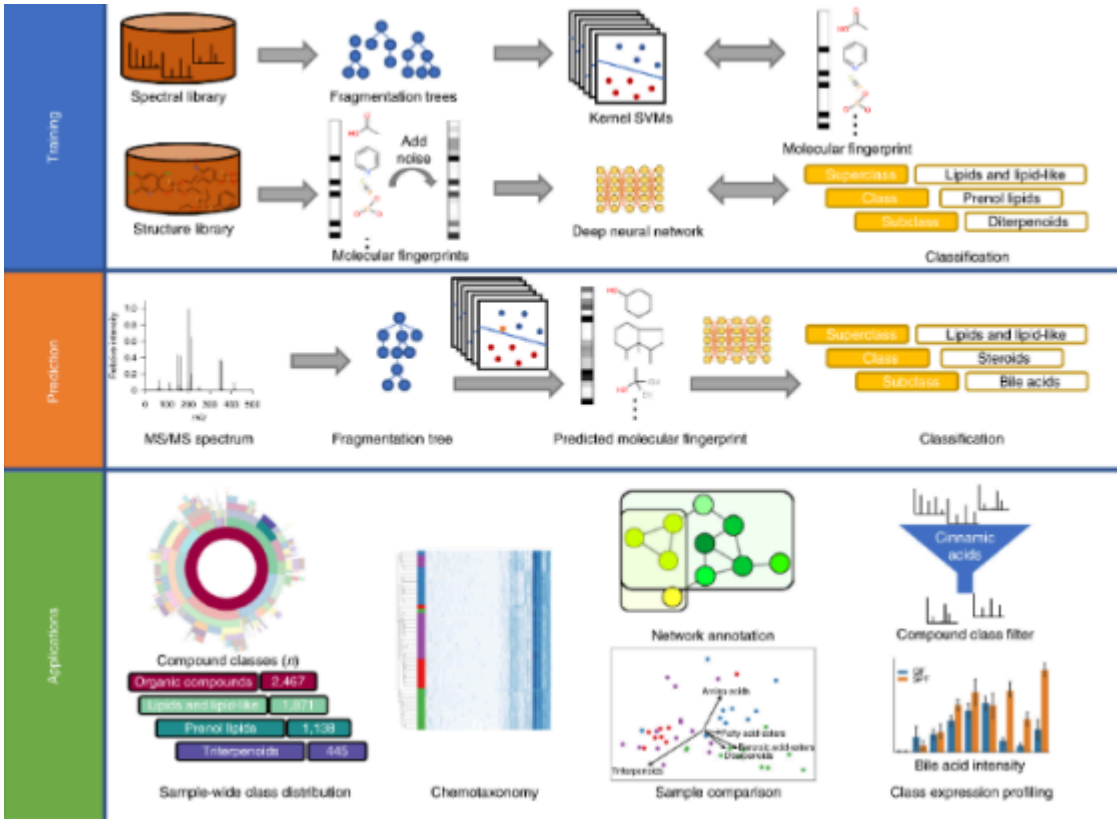


Figure 3. CANOPUS workflow

4. Conclusion

Taken together, the studies reviewed here show that the performance of machine learning in molecular toxicity and spectral analysis is determined by two connected choices: how the model

DOI: 10.54254/2755-2721/2026.AST37792

learns and what information is presented to it. RF, XGBoost, and KNN provide different compromises between robustness, flexibility, computational cost, and interpretability. RF gains stability by averaging many randomized trees and can provide useful feature-importance estimates. XGBoost extracts nonlinear relationships efficiently through sequentially boosted trees while controlling complexity with regularization. KNN requires almost no parametric training and can work well when relevant samples are locally clustered, although its computational cost and sensitivity to high-dimensional distance limit its use on large descriptor spaces. Deep CNN- and GNN-based models extend this idea further by learning representations directly from structured or high-dimensional inputs, which is particularly valuable when manually designed descriptors discard information. The reviewed mass-spectrometric examples make this point concrete. CANOPUS classified unknown metabolites into 2,497 classes and maintained 97% accuracy in the subclass-exclusion experiment; SIFTER showed that multidimensional ion-mobility measurements can support class inference even for compounds without a conventional database match. For quantitative analysis, NOM background peaks were used in place of purchased SIL standards and the best model reached an average validation R^2 of 0.84. The virtual molecular projection approach likewise retained whole-nanoparticle structural information and produced R^2 values of 0.845 for cellular uptake and 0.89 for lipophilicity. These results are promising, but they also identify where current models are weakest. Their accuracy depends heavily on the size and chemical coverage of the training data, predictions become less reliable outside the represented domain, and the internal basis of deep-model decisions is often difficult to connect to a chemical mechanism. In addition, many published demonstrations use restricted compound sets rather than the heterogeneous matrices encountered in routine environmental analysis. Progress will therefore depend less on adding model complexity alone and more on improving the evidence available to the models: broader paired mass-spectrometry/toxicity datasets, clearer applicability-domain assessment, interpretable modeling strategies, and validation on samples that differ from the training distribution. Semi-supervised and self-supervised learning may help use the large amount of unlabeled spectral data, while explainable AI methods can be used to test whether the features learned by a model correspond to meaningful physicochemical relationships. With these constraints addressed, combining machine learning with mass spectrometry could provide a more practical and scalable route for chemical screening and risk assessment.

References

[1] Yang, H., Sun, L., Li, W., Liu, G., & Tang, Y. (2018). In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. *Frontiers in Chemistry, 6*, 30. https://doi.org/10.3389/fchem.2018.00030.

[2] Wu, Z., Ramsundar, B., Feinberg, E. N., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science, 9*(2), 513-530. https://doi.org/10.1039/C7SC02664A.

[3] Dührkop, K., Nothias, L.-F., Fleischauer, M., et al. (2021). Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology, 39*(4), 462-471. https://doi.org/10.1038/s41587-020-0740-8.

[4] Picache, J. A., May, J. C., & McLean, J. A. (2020). Chemical class prediction of unknown biomolecules using ion mobility-mass spectrometry and machine learning: Supervised inference of feature taxonomy from ensemble randomization. *Analytical Chemistry, 92*(15), 10759-10767. https://doi.org/10.1021/acs.analchem.0c02137.

[5] Baek, S.-S., Choi, Y., Jeon, J., et al. (2021). Replacing the internal standard to estimate micropollutants using deep and machine learning. *Water Research, 188*, 116535. https://doi.org/10.1016/j.watres.2020.116535.

[6] Russo, D. P., Yan, X., Shende, S., et al. (2020). Virtual molecular projections and convolutional neural networks for the end-to-end modeling of nanoparticle activities and properties. *Analytical Chemistry, 92*(20), 13971-13979. https://doi.org/10.1021/acs.analchem.0c02878.