Artificial Intelligence in Molecular Physics

[Vol-01, Issue-02, July 2026]
DOI: XX
ISSN: XXXX-XXXX

AI-Assisted Spectroscopy and Molecular Structure Elucidation: Recent Progress, Challenges, and Future Research Directions

Liang Gao

Shanghai Key Laboratory of Advanced Polymeric Materials, Key Laboratory for Ultrafine Materials of Ministry of Education, Frontiers Science Center for Materiobiology and Dynamic Chemistry, School of Materials Science and Engineering, East China University of Science and Technology, Shanghai, 200237 China, Ordid ID: orcid.org/0000-0001-6852-8301

Received: 06/06/2026 **Revised:** 12/07/2026

Accepted: 20/07/2026 **Published:** 27/07/2026

Abstract—*Artificial intelligence is increasingly transforming spectroscopy and molecular structure elucidation by enabling automated spectral processing, pattern recognition, molecular property prediction, candidate ranking, and data-driven interpretation across complex analytical platforms. This review examines recent progress in AI-assisted nuclear magnetic resonance, mass spectrometry, infrared, Raman, ultraviolet-visible, fluorescence, and emerging spectroscopies, together with advances in multimodal data fusion, deep learning, generative modelling, and automated structure generation. Current evidence indicates that machine-learning and deep-learning approaches can improve peak detection, chemical shift prediction, fragmentation analysis, molecular fingerprint classification, spectral simulation, and inverse structure elucidation while reducing interpretive workload. Multimodal models further strengthen structural confidence by integrating complementary spectral and molecular information. Important barriers remain, including small and chemically imbalanced datasets, inconsistent preprocessing, instrument-dependent variability, limited external validation, weak interpretability, uncertain confidence calibration, and the generation of chemically implausible outputs. Open and FAIR spectral repositories, standardized acquisition protocols, chemically informed data augmentation, explainable and uncertainty-aware models, and physics-constrained learning are required to improve reproducibility and generalizability. Prospective validation within real laboratory workflows and closer integration with autonomous experimental systems will be essential. AI should therefore function as an advanced analytical partner that enhances expert reasoning, accelerates molecular characterization, and supports reliable applications in pharmaceutical development, materials science, chemical discovery, clinical diagnostics, and environmental analytical research.*

Keywords—***Artificial intelligence, Molecular structure elucidation, Multimodal spectroscopy, Spectral analysis, Spectroscopic data fusion***

1. INTRODUCTION

The elucidation of the molecular structure is one of the core tasks in analytical chemistry since the molecular properties, such as physicochemical, biological activity, reactivity, and utility, are defined by the identity, connectivity, stereochemistry, and conformational behaviour. Reliable structural characterization is crucial for pharmaceutical development, discovery of natural products, metabolomics, materials research, environmental monitoring, and quality control. The conventional elucidation relies on the interpretation of the complementary data of spectroscopy and spectrometric techniques, such as nuclear magnetic resonance, mass spectrometry, infrared spectroscopy, Raman spectroscopy, ultraviolet-visible spectroscopy, fluorescence analysis, and others. While mass spectrometry is particularly useful to determine molecular mass and elemental composition, substructural features and fragmentation behaviour, the confident identification of small molecules frequently relies on careful analysis of the ionization pathways, adduct formation, isotopic pattern and tandem mass spectra (De Vijlder et al., 2018). The versatility of use and the non-destructive nature of infrared spectroscopy have been proven for structurally different carbon-based materials, offering the ability to quickly and non-destructively characterize bonding environments and functional groups (Ţucureanu et al., 2016).

Even with these significant instrumental developments, the structure elucidation of molecules is not always easy, particularly because of overlapping signals, low-intensity peaks, matrix peaks, conformational effects, stereoisomerism, and lack of reference data. Interpretive uncertainty also increases for biological molecules, which are complex and heterogeneous samples. Cross-linking mass spectrometry can, for instance, provide information on spatial limitations or protein interactions; however, the analysis of data is very complex since the large number of potential cross-links, fragmentation ions, and potential false assignments have to be considered (O'Reilly & Rappsilber, 2018). Similarly, by also combining ion mobility with cryogenic spectroscopy, the resulting multidimensional data sets can be used to gain an even greater level of structural discrimination of biomolecular ions, but will still require further complex processing and comparison strategies (Kamrath & Rizzo, 2018). The importance of computational molecular spectroscopy has, therefore, grown over time to simulate the properties of spectra, correlate the computed results with the experimental ones, and to assign structures to spectra that are not possible to assign from manual inspection (Barone et al., 2021).

AI can provide a strong foundation to tackle these challenges. Machine-learning algorithms are able to recognize patterns in large spectral data sets, automatically recognise peaks, identify compounds, predict chemical shifts and fragmentation, and rank candidate molecular structures. Deep neural networks can be trained to obtain a representation directly from the raw or only partially processed spectrum, thus less relying on hand designed descriptors. The ability of generative models and graph-based architectures can also be used in solving inverse problems, where possible molecular structures are guessed based on observed analytical signals. In the pharmaceutical field, computer-based methods are now being employed to predict the behaviour of spectrograms and chromatograms, which aids in the identification of the analytes, the development of a method, and data interpretation (Malarvannan et al., 2023). The progress of such developments sets AI not only as a classification system, but also as a holistic decision support system for molecular characterization.

With automated laboratories, the use of spectroscopy is further becoming relevant. Large-scale and structured data from automated experimentation can be analysed iteratively to direct next sets of measurements, optimise reaction conditions and enhance the performance of the model (Shi et al., 2021). The electronic systems that are becoming more common in the field of automatic chemical discovery are integrating analytical tools with robotics, optimization algorithms, and machine intelligence to decrease discovery timelines and repetitive manual work (Coley et al., 2020). In materials science, similar strategies show the potential for an automated workflow to speed the synthesis, characterization, and mapping of properties over wide experimental parameter space (Stein & Gregoire, 2019). They rely on standardized data formats, instruments that are interoperable, reliable, and machine-readable metadata, and procedures that are machine-readable. Translation of experimental operations and analytical outputs into reproducible digital workflows has thus emerged as the need of the hour in the field of chemical informatics standardization and chemputation.

Nevertheless, there are ongoing problems with AI-assisted spectroscopy, such as data quality, chemical space, transferability, interpretability, uncertainty quantification, variability, and reproducibility. Data-driven models trained from a limited range of compounds or spectra may not be able to predict the structures of rare compounds, the spectra of mixtures, novel scaffolds, or spectra collected under different experimental conditions. Predictions that are made in a black-box manner can also be challenging to assess if there are chemically implausible candidates that are predicted with a high level of confidence. Human judgement is still needed to validate results and interpret results that are not conclusive or are contradictory. This review revisits the developments in the past few years in the field of spectral analysis and molecular structure elucidation using AI, highlighting the various stages of preprocessing, modality-specific applications, multimodal data fusion, deep learning, generative models, validation, and uncertainty quantification. It also discusses the challenges for

implementation and future research needs for the establishment of transparent, standardized, chemically constrained, and experimentally reliable systems for analytical practice.

Objectives of the review:

1.To summarize recent advances in AI-assisted spectral analysis and molecular structure elucidation.

2.To evaluate current methodological, interpretability, validation, and implementation challenges.

3.To identify future research priorities for reliable and automated spectroscopic workflows.

2. FOUNDATIONS OF AI-ASSISTED SPECTRAL ANALYSIS

Beyond manual feature extraction and rule-based interpretation, AI has emerged as a key element in spectral analysis, complementing traditional chemometric approaches. AI-assisted spectroscopy is a combination of statistical learning, machine learning, deep learning, and pattern recognition techniques to observe relationships between spectral signals and molecular properties. These could be used for classification, regression, clustering, anomaly detection, peak assignment, compound identification, and quantitative prediction in various analytical platforms. They are useful for processing spectra of high dimensionality, identifying subtle patterns in the signal, and minimizing the need for subjective expert analysis, especially when overlapping peaks, background noise, and/or complex multicomponent signatures are present in the data (Xue et al., 2023).

Chemometrics is still key in the methodological backbone of many AI-driven systems, as it affects the performance of these systems through its influence on the preprocessing, dimensionality reduction, calibration, and multivariate modelling. Several recent methods combine traditional methods like principal component analysis and partial least-squares regression with support vector machines as well as ensemble algorithms and neural networks. The integration of these techniques has enhanced substance classification and compositional analysis in Raman spectroscopy by leveraging discriminative information from the correlated spectral features, as shown by Pan et al. (2022).

Deep-learning models can also be used for end-to-end learning from raw or only slightly processed spectra, but are often not as interpretable as other models. Explainable AI is thus becoming more relevant to the definition of influential wavelengths, peaks, and spectral regions, and for guaranteeing that chemical plausibility and transparency are maintained during the decision-making process (Workman & Mark, 2025). In biomedical applications, especially cancer diagnostics, AI-assisted vibrational spectroscopy holds promise for quick and label-free characterization of tissues and biofluids, indicating its wider clinical relevance and automated spectral interpretation (Rehman et al., 2020). The flow of chemometric processing to AI-based uses of spectra is illustrated in Figure 1.

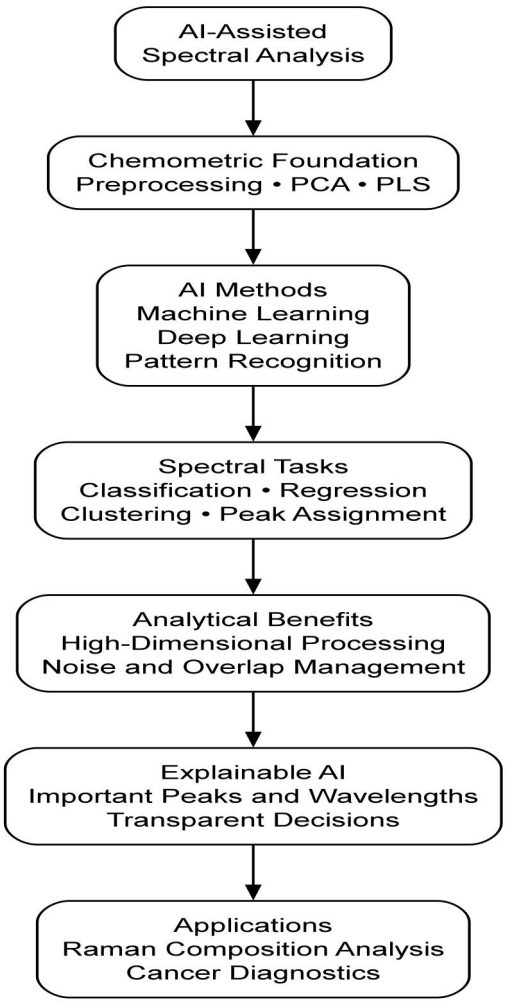

Figure 1. Workflow of AI-Assisted Spectral Analysis From Chemometric Preprocessing to Analytical Applications

3. SPECTRAL DATA ACQUISITION, PREPROCESSING, AND STANDARDIZATION

The accuracy of AI spectroscopy relies significantly on the quality, uniformity, and representativeness of the data used. Raw spectra often exhibit baseline drift, instrumental noise, scattering effects, fluorescence interferences, peak shifts, and intensity variations due to variations in sample preparation, measurement conditions, etc. To enhance the quality of the signal and extract chemically relevant information, pre-processing techniques like smoothing, normalization, baseline correction, derivative transformation, peak alignment, dimensionality reduction, etc., are needed. The choice of a preprocessing pipeline needs to be consistent with the analytical goal, as too much processing can result in the loss of weak but significant spectral signatures or create artificial patterns that can affect model predictions (Yan, 2025).

Multidimensional Raman and IR data have to be further processed to handle the complexity of spatial, temporal, and spectral features. There are methods such as principal component analysis, multivariate curve resolution, cluster analysis, and hyperspectral decomposition that can help to separate overlapping signals and uncover relationships that cannot be easily found by examining individual spectra (univariate) (Gautam et al., 2015). While Deep-learning models can learn representations from spectra without relying on manually engineered features, they can suffer from data imbalance and acquisition variability and are not yet well validated for external use (Cai et al., 2025).

In situations where the amount of data is limited, or the chemical composition of the data is unevenly distributed, data augmentation has come to the rescue as a practical solution to enrich training data diversity. While generated data can be used to create synthetic spectra, simulate realistic perturbations, and increase underrepresented classes, it is important to ensure that generated data has physical and chemical plausibility to avoid model bias (Flanagan et al., 2025). Reproducible modelling and transfer of data between different instruments require standardized acquisition protocols, documentation of metadata, file format, calibration procedures, and quality-control criteria. Harmonization of preprocessing and reporting frameworks can improve the comparability, transparency, and clinical or

industrial deployment. The major steps in spectral preprocessing, augmentation, and standardization are summarized in Table 1.

Table 1. Spectral Data Processing and AI Model Development Workflow

Workflow stage	Common methods	Role in AI-assisted analysis	Potential risks	Recommended practice	Supporting citation
Data acquisition	Instrument calibration, replicate measurement, controlled sample preparation, and standardized acquisition parameters	Generates consistent spectra suitable for model training and comparison	Instrument drift, batch effects, inconsistent sampling, and laboratory-specific bias	Use harmonized protocols, calibration standards, quality-control samples, and complete metadata	(Yan, 2025)
Noise reduction	Smoothing, filtering, wavelet transformation, and deep-learning denoising	Improves signal-to-noise ratio and facilitates weak-peak detection	Excessive smoothing may remove chemically meaningful features	Evaluate denoising parameters against raw spectra and preserve weak but reproducible signals	(Yan, 2025; Cai et al., 2025)
Baseline correction	Polynomial fitting, asymmetric least squares, derivative methods, and automated correction	Removes fluorescence, scattering, and background variation	Incorrect correction may distort peak intensities and spectral shape	Select modality-specific correction methods and report all processing parameters	(Yan, 2025)
Normalization and scaling	Vector normalization, standard normal variate, min–max scaling, and autoscaling	Reduces non-chemical intensity variation and enables cross-sample comparison	May suppress genuine concentration-dependent information	Match normalization strategy to classification or quantitative objectives	(Yan, 2025)
Peak detection and alignment	Local maxima detection, curve fitting, correlation alignment, and warping methods	Identifies relevant signals and corrects spectral shifts	False peaks, missed signals, and alignment-induced artefacts	Validate detected peaks against known standards or expert-reviewed spectra	(Gautam et al., 2015)
Dimensionality reduction	Principal component analysis, multivariate curve resolution, autoencoders, and feature selection	Reduces computational complexity and isolates informative spectral variation	Important low-variance features may be discarded	Combine statistical criteria with chemical interpretation	(Gautam et al., 2015)
Data augmentation	Noise injection, spectral shifting, intensity scaling, simulation, and generative models	Expands limited datasets and improves model robustness	Synthetic spectra may be chemically unrealistic or reinforce existing bias	Validate augmented data for physical plausibility and distributional similarity	(Flanagan et al., 2025)
Representation learning	Convolutional neural networks, recurrent networks, transformers, and latent embeddings	Learns task-specific features directly from raw or minimally processed spectra	Requires large datasets and may reduce interpretability	Use external validation, regularization, and explainability methods	(Cai et al., 2025)
Standardization and reporting	Common file formats, ontologies, preprocessing logs, and FAIR metadata	Supports reproducibility, interoperability, and model transfer	Inconsistent terminology and incomplete experimental documentation	Adopt shared data standards and publish processing code and metadata	(Hammer et al., 2021; Coca-Lopez et al., 2025)

4. AI APPLICATIONS IN NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

AI is becoming a more and more important part of NMR, with the most recent developments focused on the enhancement of spectral processing, signal assignment, chemical shift prediction, and evaluation of candidate structures. NMR is one of the most useful techniques for the elucidation of molecular structures, as it gives detailed information about the atomic environment, connectivity, stereochemistry, and molecular dynamics. In some cases,

spectra may have overlapping resonances, weak signals, complicated coupling patterns, or conformational effects, which can make manual interpretation difficult. Consequently, machine-learning models can be used to identify peaks, classify spectral features and predict chemical shifts from molecular descriptors or graph-based representations, which can help the comparison of experimental and calculated data (Jonas et al., 2022).

The capacities have grown with deep-learning architectures that enable automated denoising, reconstruction of the spectra, peak picking, and multidimensional NMR analysis. Convolutional and recurrent neural networks can be trained directly from large data sets to learn the patterns in the signals and can be more sensitive in the case of incomplete spectra or noisy spectra. However, their performance varies based on the quality of their training data sets, the consistency of their data sets, and the extent to which diverse classes of molecules are represented in their training sets (Chen et al., 2020).

Computational NMR combines quantum-chemical calculations, statistical, and AI-based methods to provide predictions of chemical shifts, coupling constants, and conformational profiles. These methods have been designed to enhance structural discrimination, but they still have limitations such as computational cost (particularly for MD simulations), the effects of the solvent, conformational sampling, and stereochemical ambiguity that merit attention (Das & Merz, 2025). The use of machine learning is also extending its application into computer-aided structure elucidation pipelines, to prioritize candidates, check for inconsistencies, and differentiate between similar natural products. These systems have the potential to decrease the interpretative workload, but they must have clear measures of confidence in the results and be verified by experts to avoid chemically implausible conclusions (Cortés et al., 2023). AI-assisted NMR is thus best suited to be used as a complement to and support for, rather than a substitute for, the process of spectroscopic reasoning. In Figure 2, the NMR spectrum interpretation and structure elucidation workflow is shown, which can be accomplished using artificial intelligence.

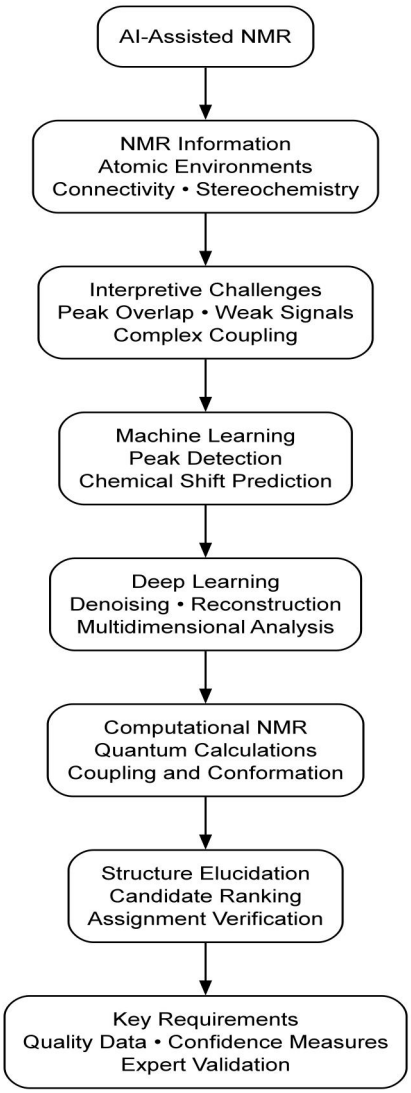

Figure 2. AI-Assisted NMR Workflow for Spectral Interpretation and Molecular Structure Elucidation

5. AI APPLICATIONS IN MASS SPECTROMETRY

Artificial intelligence has increased the analytical capabilities of MS, such as automated peak interpretation, prediction of fragments, assignment of molecular formula, compound annotation, and candidate ranking. Small-molecule mass spectrometry is very complex because the ionization, adduct formation, isotopic patterns, collision energy, and instrument type can all affect the observed spectra. Machine-learning models can be used to establish structure–spectral association, which enhances the classification, retention behavior prediction, and annotation of unknown compounds from large data sets (Hong et al., 2025).

Fragmentation is a key component in AI-assisted compound identification that employs in silico approaches. These methods attempt to simulate or learn likely pathways of bond cleavage, and then match the predicted fragments to the experimentally obtained tandem mass spectra. These can limit the number of potential candidates, but require diverse training sets as well as representation of uncommon chemical scaffolds for the rule-based or combinatorial fragmentation systems, or the deep-learning architectures (Krettler & Thallinger, 2021).

Liquid chromatography – high-resolution mass spectrometry is another area where machine learning is aiding, as it can be used in non-targeted screening situations where thousands of features might need to be prioritized and assigned structure. While exact mass, isotope distribution, retention time, collision cross section, and fragmentation evidence can be incorporated into algorithms, the current state of incomplete spectral libraries and inconsistent confidence reporting is a significant challenge (Hupatz et al., 2025).

Another difficulty with natural product annotation is the fact that structurally similar analogues can yield similar mass spectra. When used in combination with MS data, coupling MS with NMR data enables machine-learning systems to leverage complementary data sources for substructure identification and candidate discrimination to boost confidence in dereplication and novel compound characterization. (Hu & Qiu, 2023). AI mass spectrometry can thus help to speed up molecular identification, but only if the results are carefully validated, the score is explained, and the structures are confirmed by experts. AI in mass spectrometry (MS) has a pathway for structural annotation and molecular identification, as shown in Figure 3.

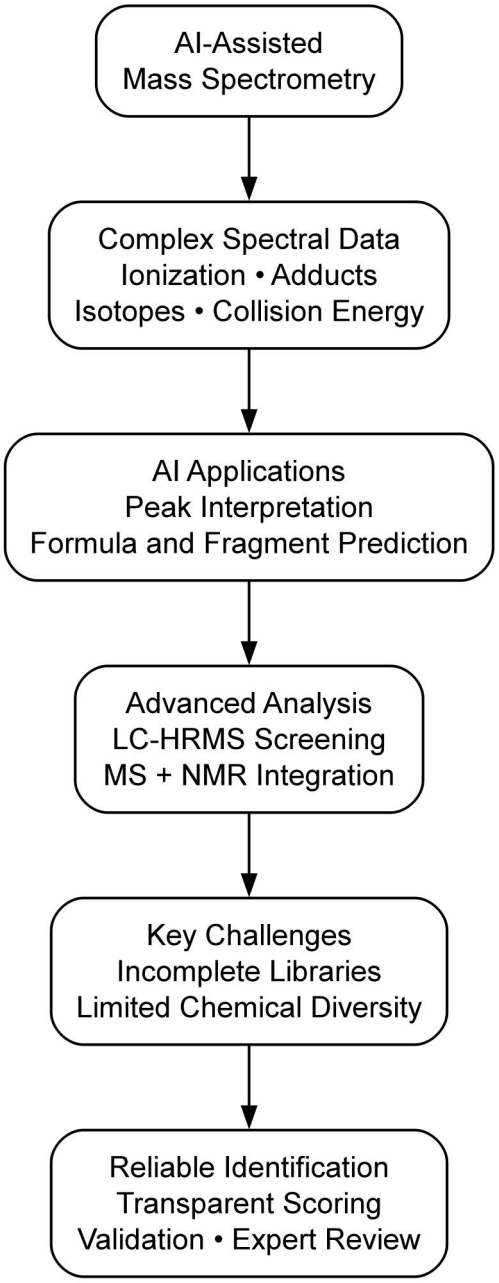

Figure 3. AI-Assisted Mass Spectrometry Workflow for Structural Annotation and Reliable Molecular Identification

6. AI APPLICATIONS IN INFRARED, RAMAN, AND VIBRATIONAL SPECTROSCOPY

AI has improved vibrational spectroscopic techniques, including IR and Raman, by creating automated spectral classification, quantitative prediction, molecular fingerprint recognition, and functional-group identification. The techniques produce chemically rich spectral signatures derived from molecular vibrations, but the information they provide can be difficult to interpret because of overlapping bands, fluorescent spectra, baseline fluctuations, and slight variations between structurally similar samples. Machine-learning algorithms can learn from data to recognize those areas of the spectrum that are discriminatory, and to identify the nonlinear relationship between spectral features and sample properties, which enhances the analytical sensitivity and decreases the need to rely on subjective visual inspection (Meza Ramirez et al., 2021).

Raman and infrared spectroscopy have specific applications in biomedical analyses, where they give rapid, label-free molecular fingerprints of cells, tissues and biofluids. AI models are able to identify the differences in proteins, lipids, nucleic acids and carbohydrates between malignant and

non-malignant samples that are not obvious in the traditional interpretation of the data. Their use in cancer detection illustrates the potential of the vibrational spectroscopy technique for minimally invasive diagnosis and real-time tissue characterisation (Zhang et al., 2023).

Machine learning can also be used in theoretical vibrational spectroscopy, where the infrared and Raman spectra can be predicted from the structure of the molecule. These models can be used to approximate more complex quantum-chemical calculations, allowing for easier simulation and comparison of the spectra with experimental data. The reliability of the model is still, however, reliant on molecular diversity, conformational sampling, and accurate reference calculations (Han et al., 2022).

Vibrational analysis extended by combining spatial and spectral information: hyperspectral imaging. The use of AI for the analysis of cereal quality has been limited to the identification of compositional variation, contamination, adulteration, and structural defects using infrared images (An et al., 2023). The applications highlight that AI-powered vibrational spectroscopy has the potential to aid in the rapid and non-destructive analysis in various clinical, agricultural, pharmaceutical, and materials scenarios.

7. AI APPLICATIONS IN ULTRAVIOLET–VISIBLE, FLUORESCENCE, AND EMERGING SPECTROSCOPIES

The use of AI is growing for assisting the advancement of ultraviolet–visible and fluorescence spectroscopy, as well as advanced optical spectroscopy, to enhance the prediction of spectra, classification of chromophores, estimation of photophysical properties, and molecular design. These techniques are useful for studies of electronic transitions, excited-state behaviour, fluorescence emission and charge transfer processes: they have potential applications in photochemistry, sensing, bioimaging, optoelectronics and materials development. The relationship between a molecular structure and its maximum absorption and emission wavelengths, quantum yields, and other optical properties can be established using quantitative structure–property relationship models and machine-learning algorithms, which will allow screening of the candidate photonic compounds much faster (Buglak et al., 2025).

As the performance of organic fluorescent materials is governed by complicated molecular structure, aggregation, solvent environment, excited-state dynamics and other factors, this is an ideal field for AI. Machine-learning models can predict the fluorescence colour, brightness, stability and aggregation-induced emission behaviour, which helps to decrease the number of compounds that must be synthesized and tested in experiments (Zhong et al., 2025).

Explainable deep-learning methods have also been extended to the traditional empirical rules applied to ultraviolet–visible spectra. These models can be used to derive interpretable design principles for property-oriented chromophores, and enhance rational molecular optimization (Joung et al., 2022).

New spectroscopic techniques are also being developed and coupled with computational modelling. The molecular simulations together with ultrafast spectroscopy can help to interpret transient electronic states, intramolecular charge transfer and excited state relaxation processes, which are not easily accessible based on steady state measurements (Patrizi et al., 2020). AI can complement these workflows by speeding up spectral deconvolution and the identification of kinetic components, and by connecting temporal signals to molecular mechanisms. These demonstrate the ability of AI-powered optical spectroscopy to enable molecular characterisation and targeted design, but validation and predictions from experiments remain crucial to ensure the reliability of the model. It can be inferred from Table 2 that there are some differences with respect to the modality used in AI applications, outputs, advantages, and limitations.

Table 2. AI Applications Across Major Spectroscopic and Spectrometric Modalities

Analytical modality	Primary AI applications	Typical analytical outputs	Principal advantages	Key limitations	Supporting citation
Nuclear magnetic resonance spectroscopy	Peak picking, spectral reconstruction, chemical-shift prediction, signal assignment, and candidate ranking	Atomic environments, molecular connectivity, coupling constants, stereochemistry, and conformational profiles	Reduces manual interpretation and improves comparison between experimental and calculated spectra	Overlapping resonances, conformational effects, solvent dependence, and limited molecular diversity in training data	(Jonas et al., 2022; Chen et al., 2020; Das & Merz, 2025; Cortés et al., 2023)
Mass spectrometry	Molecular formula assignment, fragmentation prediction, compound annotation, retention prediction, and	Exact mass, isotope patterns, fragment ions, substructures, and probable molecular identities	Supports rapid screening of large datasets and improves annotation of unknown	Incomplete spectral libraries, instrument dependence, uncommon scaffolds, and	(Hong et al., 2025; Krettler & Thallinger, 2021; Hupatz et

	candidate prioritization		compounds	inconsistent confidence reporting	al., 2025)
Infrared spectroscopy	Functional-group identification, spectral classification, quantitative prediction, and hyperspectral analysis	Bonding environments, compositional profiles, functional groups, and spatial chemical distributions	Enables rapid, non-destructive, and label-free analysis	Baseline drift, water interference, overlapping bands, and preprocessing sensitivity	(Meza Ramirez et al., 2021; An et al., 2023)
Raman spectroscopy	Molecular fingerprint classification, disease detection, substance identification, and compositional analysis	Vibrational fingerprints, molecular composition, tissue characteristics, and material identity	Requires minimal sample preparation and supports real-time analysis	Fluorescence interference, weak signals, instrument variability, and limited standardized datasets	(Pan et al., 2022; Zhang et al., 2023)
Ultraviolet-visible spectroscopy	Absorption prediction, chromophore classification, and structure–property modelling	Absorption maxima, electronic transitions, and chromophore characteristics	Accelerates screening and rational design of optically active compounds	Limited structural specificity and strong dependence on solvent and molecular environment	(Buglak et al., 2025; Joung et al., 2022)
Fluorescence spectroscopy	Emission-property prediction, fluorophore classification, sensing, and materials optimization	Emission wavelength, quantum yield, fluorescence intensity, and aggregation behaviour	Provides high sensitivity and supports bioimaging and sensing applications	Environmental sensitivity, photobleaching, aggregation effects, and limited generalizability	(Zhong et al., 2025)
Ultrafast and emerging optical spectroscopy	Spectral deconvolution, transient-state recognition, and excited-state kinetic analysis	Charge-transfer mechanisms, transient species, and relaxation pathways	Resolves rapid molecular processes inaccessible to steady-state methods	Complex instrumentation, high-dimensional data, and limited annotated training datasets	(Patrizi et al., 2020)

8. MULTIMODAL SPECTRAL FUSION FOR MOLECULAR STRUCTURE ELUCIDATION

Multimodal spectral fusion aims to integrate complementary data from nuclear magnetic resonance, mass spectrometry, infrared, Raman, ultraviolet–visible and fluorescence spectroscopy, and related methods, to enhance the molecular identification and minimize ambiguity in molecular structure. Molecular mass, fragmentation behaviour, functional groups, electronic transitions, and local atomic environments can be correlated and interpreted in an integrated fashion by individual modalities, but not by all modalities. This integration can be achieved through machine-learning models, which can learn relationships across different datasets, and identify cross-modal patterns that are not captured by analysing the data sets individually (Chen et al., 2023).

Data fusion can be done at early, intermediate or late stage. Early fusion is a fusion of preprocessed features, intermediate is a fusion of learned representations and late is a fusion of prediction from different models. Depending on data dimensionality, the size of the samples and the measurement compatibility, as well as the amount of missing information, different strategies are optimal. In traditional Chinese medicine, it is shown how the combination of chromatographic, spectroscopic and physicochemical data can enhance the quality evaluation and compound discrimination of chemically complex samples (Huang et al., 2023).

The use of deep-learning models for establishing spectrum–structure correlations from a variety of analytical platforms is becoming more common. Shared latent representations can be used to correlate the spectral signals with corresponding molecular graphs, substructures, and candidate identities, which can be beneficial for both forward prediction and inverse structure elucidation (Lu et al., 2024). The abilities are especially applicable to inverse chemical issues, such as the determination of probable molecular structures or development of design candidates

sought to have particular properties, given observed signals (Sridharan et al., 2022).

Despite this, a careful calibration of multimodal systems, ensuring standardisation of metadata, balancing of the datasets, and transparency of the weighting of each modality is required. Strong validation is needed to make sure that fused models enhance the confidence in the structure, not just the correlated errors, or any modality-specific bias.

9. DEEP LEARNING, GENERATIVE MODELS, AND AUTOMATED STRUCTURE GENERATION

Deep learning and generative modelling are not only driving the capacity to recognize patterns in molecular structures but also to generate candidates and even reverse-design molecules. Neural architectures may be used to learn complex relationships between molecular representations and physicochemical properties and spectral outputs, and thus propose structures that are consistent with a set of analytical constraints, as defined a priori. The typical autonomous molecular design systems often employ generative algorithms, property predictors, optimization schemes, and computational validation to traverse large chemical spaces more efficiently, as compared to the conventional trial-and-error methods (Joshi & Kumar, 2021).

Examples of generative models are variational autoencoders, generative adversarial networks, recurrent neural networks, transformers, graph-based networks and diffusion models. These systems can produce molecular graphs, simplified molecular-input line-entry system strings, or three dimensional conformations with or without optimizing the structural novelty and structural validity, the synthetic feasibility and the target properties. They can be integrated with spectroscopic data and allow to perform spectrum-to-structure workflows where the candidate molecules are generated and prioritized based on their model prediction of agreement with the experimental data. Although significant strides are being made, generative AI can generate chemically unstable, synthetically inaccessible or even too limited of a range of structures, necessitating chemical constraints and expert verification (Sun et al., 2025).

Interpretability and uncertainty awareness are important aspects for trustworthy deployment. Confidence needs to be indicated in the structure proposed and models should show which spectral features, fragments or learned representations have been used. For example, by assisting an analyst in differentiating between high confidence predictions and those that need further experimental verification, an interpretable and uncertainty-aware system can help to improve the calibration of trust in the system by the user (Tomsett et al., 2020).

Uncertainty communication continues to be difficult as the level of confidence that can be ascribed to the model may not be the same as that felt by the people. The representation of uncertainty in ExAIFrameworks should thus be in a form that enables the evaluation of the predictions in an informed manner, but not at the expense of suggesting more than the predictions are certain (Chiaburu et al., 2024). Generative systems are most useful in iterative processes that include human validation of the spectra, computational filtering, and generative systems. Table 3 shows the comparison of the advanced AI methods involved in spectral interpretation and structure generation.

Table 3. Advanced AI Strategies for Molecular Structure Elucidation

AI strategy	Core function	Spectral or molecular inputs	Contribution to structure elucidation	Major concern	Supporting citation
Classical chemometrics	Multivariate classification, regression, clustering, and calibration	Preprocessed spectral intensities and manually selected variables	Establishes interpretable relationships between spectra and molecular properties	Limited capacity to model highly nonlinear spectral relationships	(Xue et al., 2023; Workman & Mark, 2025)
Convolutional neural networks	Automated local-feature extraction and pattern recognition	One-dimensional spectra, spectral images, and hyperspectral datasets	Detects peaks, bands, and spatially localized molecular signatures	Sensitivity to acquisition differences and limited interpretability	(Chen et al., 2020; Cai et al., 2025)
Recurrent neural networks and transformers	Sequential modelling and long-range dependency learning	Spectral sequences, molecular strings, and peak lists	Links distributed spectral features with molecular representations	High computational requirements and dependence on large training datasets	(Lu et al., 2024; Sun et al., 2025)
Graph neural networks	Learning from molecular connectivity and atomic environments	Molecular graphs, atom features, bonds, and experimental spectra	Predicts chemical shifts, fragmentation patterns, and structure–spectrum relationships	Difficulty representing stereochemistry, conformational flexibility, and	(Cortés et al., 2023; Lu et al., 2024)

				unusual bonding	
Generative models	Creation of candidate molecular structures under analytical constraints	Spectra, molecular graphs, SMILES strings, descriptors, and target properties	Generates and ranks structures compatible with observed spectral evidence	Chemically invalid, unstable, or synthetically inaccessible candidates	(Joshi & Kumar, 2021; Sun et al., 2025)
In silico fragmentation models	Prediction of bond cleavage and fragment-ion formation	Molecular structures, collision conditions, and tandem mass spectra	Narrows candidate lists and supports mass-spectral annotation	Limited accuracy for uncommon chemical scaffolds and rearrangement pathways	(Krettler & Thallinger, 2021; Hupatz et al., 2025)
Multimodal data fusion	Integration of complementary analytical evidence	NMR, MS, infrared, Raman, UV-visible, fluorescence, chromatography, and metadata	Reduces structural ambiguity by combining mass, functional-group, connectivity, and electronic information	Missing modalities, incompatible scales, and amplification of correlated errors	(Chen et al., 2023; Huang et al., 2023)
Spectrum-to-structure correlation	Mapping experimental signals to molecular candidates	Spectral embeddings, molecular graphs, and structural databases	Supports inverse structure elucidation and automated candidate prioritization	Non-unique mappings and insufficient uncertainty estimation	(Lu et al., 2024; Sridharan et al., 2022)
Autonomous analytical systems	Closed-loop experimentation, analysis, and decision-making	Robotic platforms, analytical instruments, optimization models, and spectral feedback	Iteratively tests structural hypotheses and selects subsequent experiments	Error propagation, integration failures, and unclear accountability	(Coley et al., 2020; Tom et al., 2024)

10. MODEL VALIDATION, INTERPRETABILITY, AND UNCERTAINTY QUANTIFICATION

It is essential to carefully validate models, interpret results in a transparent manner, and explicitly use statistics to estimate the uncertainty of predictions when using AI-assisted spectroscopy. High internal accuracy is not enough to guarantee analytical reliability as the performance may be affected when spectra of the model are presented to spectra from other instruments, other laboratories, other sample populations or other chemical classes. Independent external validation should thus be based on external data sets, chemically appropriate partitioning of train–test datasets, repeated cross-validation and validation against realistic baseline methods. If the molecules are structurally similar or if the same spectrum is performed twice, for example, the random splitting of data may lead to artificial over-optimism when measuring performance.

Explainable AI can help to give insight into how the parts of the spectrum, molecular descriptors, peaks or latent features contribute to the prediction. Common techniques are feature importance, saliency maps, local explanation methods, attention visualisation and surrogate modelling. The necessity to differentiate between the global behaviour of a complex classification and prediction system and the explanations that can be given to individual predictions, as shown in explainability frameworks developed for these types of systems (Nascita et al., 2024). In the field of molecular research, explanation methods must be chemically meaningful and they should provide insights into whether the model is based on a valid structure–property relationship or on experimental artefacts, or on dataset bias (Ponzoni et al., 2023).

Uncertainty quantification is also crucial to determine predictions that lie outside the domain of applicability of the model. Calibration curves, (pseudo) confidence intervals, ensemble variation, Bayesian approaches and conformal prediction can suggest when further measurements or expert evaluation are needed. Further, open datasets, standardized benchmarks, interoperable formats, and FAIR principles can enhance the reproducibility and allow for independent comparison of AI-enabled Raman models in research settings (Coca-Lopez et al., 2025). Therefore, it is crucial to have robust validation and uncertainty reporting for translating predictive performance into reliable analytical evidence. Overall, the validation, explainability, and uncertainty is summarized in figure 4 in the context of reliable AI-assisted spectroscopy.

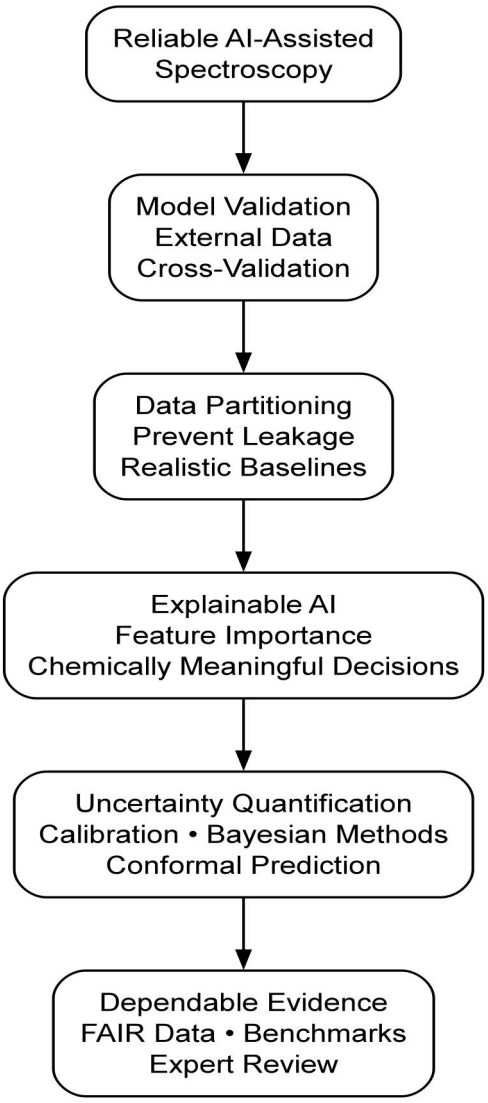

Figure 4. Framework for Validation, Explainability, and Uncertainty in Reliable AI-Assisted Spectroscopy

11. CURRENT CHALLENGES IN AI-ASSISTED STRUCTURE ELUCIDATION

There are several challenges to the implementation of AI for molecular structure elucidation, including limited availability of spectra, lack of spectral standardization, limited generalizability of the models, lack of model interpretation, and the need for integration of AI tools within the lab. There are many small, chemically imbalanced, instrument-specific and partially annotated datasets that preclude the possibility of models learning rare compounds and new molecular scaffolds. Additionally, the inability to reproduce results across labs is further reduced because of differences in acquisition settings, sample preparation and pre-processing procedures as well as metadata reporting. The application of open repositories and the principles of FAIR data can enhance the findability, accessibility, interoperability and reusability of spectral data, but requires a community and the development of shared spectral formats, validated annotations, standardised ontologies and community participation (Coca-Lopez et al., 2025).

The translation to independent analytical process has some technical difficulties, as well. Self-driving laboratories combine robotics experimentation, analytical instrumentation, machine learning and adaptive decision making, but to operate reliably, instrument communication must be reliable, quality control must be automated, failure detection must be reliable and the decision pathways must be traceable. When experiments are not properly verified, errors can be cascaded in closed-loop systems if incorrect structural assignments are made which lead to further experiments. Human intervention is thus still necessary for some spectra that are ambiguous, contain complex mixtures, are stereochemical assignments, and are beyond the applicability domains (Tom et al., 2024).

Large language models can be used as spectrographic assistants to summarize results, create analysis scripts, connect observation with the knowledge of chemistry, and aid in the documentation of the laboratories. They may be able to give answers that are constructed, wrong answers or overly confident answers, particularly in the absence of all the spectral evidence. To use effectively, the information needs to be retrieved from credible databases, identified as the source of the information, be domain specific and have been reviewed by an expert (Fu et al., 2025). Other obstacles include the cost of computation, datasets which are available only to the proprietary user, cybersecurity issues, intellectual property questions, and questions of accountability. These limitations can be overcome by having benchmark datasets, external validation, uncertainty-aware predictions, interoperable infrastructures, as well as governance standards for research, clinical and industrial settings.

12. LIMITATIONS AND FUTURE DIRECTIONS

The current AI-enabled spectroscopy is restricted by small, imbalanced and instrument specific datasets, inconsistent pre-processing, incomplete metadata and lack of external validation. Models can be good at predicting spectra of test sets of compounds but have poor performance on unseen compounds, mixed samples, rare scaffolds, and spectra collected in different laboratories. The limited interpretability, uncertainty of confidence calibration, proprietary databases and the potential for chemically implausible predictions further limit the routine use. Also many studies don't have a common framework for comparison, reproducible code, or a prospective analysis of the performance in a real analytical workflow.

Going forward, modern spectral repositories that are open and FAIR, acquisition standards that are harmonized, chemically informed data augmentation, and benchmark sets with wider molecules should be targeted. The NMR, mass spectrometry, Raman and infrared data, along with molecular graph and experimental metadata should be included in the multimodal foundation models. Improved reliability with Physics-informed learning, Conformal prediction, Explainable AI and uncertainty-aware candidate ranking. For safe clinical, pharmaceutical and industrial applications, prospective validation, human–AI co-working, interoperable laboratory infrastructure, and regulatory guidance will be crucial.

13. CONCLUSION

AI is transforming the challenges of spectroscopy and structure elucidation of molecules through the ability to quickly process and visualize spectra, extract features automatically, predict chemical shifts and fragmentations, fuse multimodal data, generate candidate structures, and provide uncertainty-aware decision support. From NMR to mass spectrometry, from infrared to Raman, from ultraviolet-visible to fluorescence and beyond, the promise of machine learning and deep learning to boost the analytical efficiency, sensitivity, and consistency of scientific analysis is evident in applications across a wide range of spectroscopic platforms. These are further enhanced by generative models and autonomous laboratory systems, which can "connect spectral evidence to molecular design, and iterative experimentation. Some challenges that have not yet been overcome are the lack of diverse sets, differences in data processing, instrument-related variability, limited interpretability, poor external validation and the possibility of producing chemically unrealistic results. These will be achieved with the help of open and FAIR spectral repositories, standardized acquisition and reporting protocols, multimodal foundation models, physics-informed learning, transparent uncertainty quantification and chemically meaningful explainability. Wherever possible, human expertise should be an integral part of the assignment of structures that are ambiguous or high stakes. AI can thus be seen as an advanced assistant to the experts, rather than a replacement. By carefully benchmarking, integrating different disciplines, and taking a responsible approach, AI-driven spectroscopy can be used as a solid basis to characterize molecules, discover new compounds, develop drugs, research materials, and diagnose diseases.

REFERENCES

1.An, D., Zhang, L., Liu, Z., Liu, J., & Wei, Y. (2023). Advances in infrared spectroscopy and hyperspectral imaging combined with artificial intelligence for the detection of cereals quality. Critical Reviews in Food Science and Nutrition, 63(29), 9766-9796.

2.Barone, V., Alessandrini, S., Biczysko, M., Cheeseman, J. R., Clary, D. C., McCoy, A. B., ... & Puzzarini, C. (2021). Computational molecular spectroscopy. Nature Reviews Methods Primers, 1(1), 38.

3.Buglak, A. A., Chebotaev, P. P., & Filatov, M. A. (2025). Applications of QSPR and Machine Learning in Molecular Photonics. Advanced Optical Materials, 13(33), e01713.

4.Cai, Y., Lin, Y., Cai, H., & Ni, H. (2025). Deep learning in vibrational spectroscopy: Benefits, limitations, and recent progress. Journal of the Chinese Chemical Society, 72(6), 611-626.

5.Chen, D., Wang, Z., Guo, D., Orekhov, V., & Qu, X. (2020). Review and prospect: deep learning in nuclear magnetic resonance spectroscopy. Chemistry–A European Journal, 26(46), 10391-10401.

6.Chen, X., Shu, W., Zhao, L., & Wan, J. (2023). Advanced mass spectrometric and spectroscopic methods coupled with machine learning for in vitro diagnosis. View, 4(1), 20220038.

7.Chiaburu, T., Haußer, F., & Bießmann, F. (2024).

Uncertainty in xai: Human perception and modeling approaches. Machine Learning and Knowledge Extraction, 6(2), 1170-1192.

8.Coca-Lopez, N., Alcolea-Rodriguez, V., Bañares, M. A., Brockhauser, S., Gorenflot, J., Henderson, A., ... & Portela, R. (2025). Artificial intelligence-powered Raman spectroscopy through open science and FAIR principles. ACS nano, 19(44), 38189-38218.

9.Coca-Lopez, N., Alcolea-Rodriguez, V., Bañares, M. A., Brockhauser, S., Gorenflot, J., Henderson, A., ... & Portela, R. (2025). Open and FAIR Raman spectroscopy. Paving the way for artificial intelligence.

10.Coley, C. W., Eyke, N. S., & Jensen, K. F. (2020). Autonomous discovery in the chemical sciences part I: Progress. Angewandte Chemie International Edition, 59(51), 22858-22893.

11.Cortés, I., Cuadrado, C., Hernández Daranas, A., & Sarotti, A. M. (2023). Machine learning in computational NMR-aided structural elucidation. Frontiers in Natural Products, 2, 1122426.

12.Das, S., & Merz Jr, K. M. (2025). Exploring the frontiers of computational nmr: Methods, applications, and challenges. Chemical reviews, 125(19), 9256-9295.

13.De Vijlder, T., Valkenborg, D., Lemière, F., Romijn, E. P., Laukens, K., & Cuyckens, F. (2018). A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. Mass spectrometry reviews, 37(5), 607-629.

14.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.

15.Fu, L., Zhou, Q., Jin, M., & Wu, W. (2025). Large language models as spectrographic assistants: Opportunities and challenges in laboratory data analysis. Environmental Chemistry and Safety, 1(1), 9600002.

16.Gautam, R., Vanga, S., Ariese, F., & Umapathy, S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. EPJ Techniques and Instrumentation, 2(1), 8.

17.Hammer, A. J., Leonov, A. I., Bell, N. L., & Cronin, L. (2021). Chemputation and the standardization of chemical informatics. JACS Au, 1(10), 1572-1587.

18.Han, R., Ketkaew, R., & Luber, S. (2022). A concise review on recent developments of machine learning for the prediction of vibrational spectra. The Journal of Physical Chemistry A, 126(6), 801-812.

19.Hong, Y., Ye, Y., & Tang, H. (2025). Machine learning in small-molecule mass spectrometry. Annual Review of Analytical Chemistry, 18.

20.Hu, G., & Qiu, M. (2023). Machine learning-assisted structure annotation of natural products based on MS and NMR data. Natural Product Reports, 40(11), 1735-1753.

21.Huang, R., Ma, S., Dai, S., & Zheng, J. (2023). Application of data fusion in traditional Chinese medicine: a review. Sensors, 24(1), 106.

22.Hupatz, H., Rahu, I., Wang, W. C., Peets, P., Palm, E. H., & Kruve, A. (2025). Critical review on in silico methods for structural annotation of chemicals detected with LC/HRMS non-targeted screening. Analytical and Bioanalytical Chemistry, 417(3), 473-493.

23.Jonas, E., Kuhn, S., & Schlörer, N. (2022). Prediction of chemical shift in NMR: A review. Magnetic Resonance in Chemistry, 60(11), 1021-1031.

24.Joshi, R. P., & Kumar, N. (2021). Artificial intelligence for autonomous molecular design: A perspective. Molecules, 26(22), 6761.

25.Joung, J. F., Han, M., Jeong, M., & Park, S. (2022). Beyond Woodward–Fieser rules: design principles of property-oriented chromophores based on explainable deep learning optical spectroscopy. Journal of Chemical Information and Modeling, 62(12), 2933-2942.

26.Kamrath, M. Z., & Rizzo, T. R. (2018). Combining ion mobility and cryogenic spectroscopy for structural and analytical studies of biomolecular ions. Accounts of chemical research, 51(6), 1487-1495.

27.Krettler, C. A., & Thallinger, G. G. (2021). A map of mass spectrometry-based in silico fragmentation prediction and compound identification in metabolomics. Briefings in Bioinformatics, 22(6), bbab073.

28.Lu, X. Y., Wu, H. P., Ma, H., Li, H., Li, J., Liu, Y. T., ... & Liu, G. K. (2024). Deep learning-assisted spectrum–structure correlation: state-of-the-art and perspectives. Analytical Chemistry, 96(20), 7959-7975.

29.Malarvannan, M., Kumar, K. V., Reddy, Y. P., Nikhil, P., Aishwarya, D., Ravichandiran, V., & Ramalingam, P. (2023). Assessment of computational approaches in the prediction of spectrogram and chromatogram behaviours of analytes in pharmaceutical analysis: assessment review. Future Journal of Pharmaceutical Sciences, 9(1), 86.

30.Meza Ramirez, C. A., Greenop, M., Ashton, L., & Rehman, I. U. (2021). Applications of machine learning in spectroscopy. Applied Spectroscopy Reviews, 56(8-10), 733-763.

31.Nascita, A., Aceto, G., Ciuonzo, D., Montieri, A., Persico, V., & Pescapé, A. (2024). A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection. IEEE Communications Surveys & Tutorials, 27(5), 3165-3198.

32.O'reilly, F. J., & Rappsilber, J. (2018). Cross-linking

mass spectrometry: methods and applications in structural, molecular and systems biology. Nature structural & molecular biology, 25(11), 1000-1008.

33.Pan, L., Zhang, P., Daengngam, C., Peng, S., & Chongcheawchamnan, M. (2022). A review of artificial intelligence methods combined with Raman spectroscopy to identify the composition of substances. Journal of Raman Spectroscopy, 53(1), 6-19.

34.Patrizi, B., Cozza, C., Pietropaolo, A., Foggi, P., & Siciliani de Cumis, M. (2020). Synergistic approach of ultrafast spectroscopy and molecular simulations in the characterization of intramolecular charge transfer in push-pull molecules. Molecules, 25(2), 430.

35.Ponzoni, I., Páez Prosper, J. A., & Campillo, N. E. (2023). Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. Wiley Interdisciplinary Reviews: Computational Molecular Science, 13(6), e1681.

36.Rehman, I. U., Khan, R. S., & Rehman, S. (2020). Role of artificial intelligence and vibrational spectroscopy in cancer diagnostics. Expert Review of Molecular Diagnostics, 20(8), 749-755.

37.Shi, Y., Prieto, P. L., Zepel, T., Grunert, S., & Hein, J. E. (2021). Automated experimentation powers data science in chemistry. Accounts of Chemical Research, 54(3), 546-555.

38.Sridharan, B., Goel, M., & Priyakumar, U. D. (2022). Modern machine learning for tackling inverse problems in chemistry: molecular design to realization. Chemical communications, 58(35), 5316-5331.

39.Stein, H. S., & Gregoire, J. M. (2019). Progress and prospects for accelerating materials science with automated and autonomous workflows. Chemical science, 10(42), 9640-9649.

40.Sun, Y., Chen, L., Jing, Z., Li, Y. Y., Kim, D., Gao, J. Y., ... & Hu, P. (2025). Generative AI for the design of molecules: advances and challenges. Journal of Chemical Information and Modeling, 65(23), 12668-12690.

41.Tom, G., Schmid, S. P., Baird, S. G., Cao, Y., Darvish, K., Hao, H., ... & Aspuru-Guzik, A. (2024). Self-driving laboratories for chemistry and materials science. Chemical Reviews, 124(16), 9633-9732.

42.Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., ... & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns, 1(4).

43.Ţucureanu, V., Matei, A., & Avram, A. M. (2016). FTIR spectroscopy for carbon family study. Critical reviews in analytical chemistry, 46(6), 502-520.

44.Workman Jr, J., & Mark, H. (2025). Recent Research in Chemometrics and AI for Spectroscopy, Part II: Emerging Applications, Explainable AI, and Future Trends. Spectroscopy, 40(7), 6-10.

45.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.

46.Yan, C. (2025). A review on spectral data preprocessing techniques for machine learning and quantitative analysis. IScience, 28(7).

47.Zhang, S., Qi, Y., Tan, S. P. H., Bi, R., & Olivo, M. (2023). Molecular fingerprint detection using Raman and infrared spectroscopy technologies for cancer detection: a progress review. Biosensors, 13(5), 557.

48.Zhong, J., Zhu, W., Shen, S., Zhou, N., Xi, M., Du, K., ... & Tang, B. Z. (2025). Machine learning for organic fluorescent materials. Aggregate, 6(9), e70089.