Chem Soc Rev

REVIEW ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d4cs01293c

Intelligent understanding of spectra: from structural elucidation to property design

Shuo Feng, † Meng Huang, † Yanbo Li,† Aoran Cai, Xiaoyu Yue, Song Wang, Linjiang Chen, Jun Jiang * and Yi Luo *

Spectroscopy serves as a bridge between experimental observations and quantum mechanical principles, linking molecular microstructure to macroscopic material properties. Despite its central importance, establishing quantitative structure–property relationships from spectral data remains challenging, typically requiring expensive quantum chemistry calculations and specialized expertise. The integration of artificial intelligence (AI) with spectroscopy presents a transformative opportunity to overcome these limitations. AI models can leverage spectral data as molecular descriptors to construct predictive relationships—both spectrum-to-structure and spectrum-to-property mappings. This review presents representative advances at the AI–spectroscopy intersection, highlighting how these approaches address challenges in spectroscopic analysis: automated spectral interpretation, efficient spectral prediction, and accurate property determination from spectroscopic fingerprints. Beyond individual applications, we demonstrate how AI enables the development of unified spectrum–structure–property frameworks capable of predicting functional properties directly from spectral data. This integrated approach opens pathways for spectrum-guided, AI-driven inverse design of functional matters. In addition, we emphasize the importance of model interpretability, which can illuminate the fundamental physics underlying spectrum–structure–property relationships. Looking forward, we propose that integrating large-scale AI architectures with spectroscopic descriptors could establish universal spectrum–structure–property relationships, potentially revolutionizing chemical theory.

Received 15th June 2025

DOI: 10.1039/d4cs01293c

rsc.li/chem-soc-rev

1. Introduction

Spectroscopy is the study of how matter interacts with electromagnetic radiation. By analysing how atoms and molecules absorb, emit, or scatter light, spectroscopy provides detailed information about their energy levels, structures, and dynamic behaviours, thereby advancing our understanding of the

State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China. E-mail: jiangj1@ustc.edu.cn, yiluo@ustc.edu.cn

† S. Feng, M. Huang and Y. Li contributed equally to this work.

Shuo Feng

Shuo Feng serves as a Senior Research Associate at the State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China. He earned his PhD in Solid Mechanics from the University of Science and Technology of China in 2018. His research is focused on material design and performance regulation, integrating spectroscopy and artificial intelligence technology.

Meng Huang

Meng Huang is a Senior Research Associate in Professor Jun Jiang's research group at the University of Science and Technology of China (USTC). He earned his PhD in Chemical Physics from The Ohio State University in 2018 and subsequently completed postdoctoral fellowships at the University of New Mexico and Emory University. His research is focused on electronic structure theory and intelligent spectroscopy.

microscopic world. A notable example of this is the analysis of the spectral lines of the hydrogen atom, on the basis of which Niels Bohr proposed the groundbreaking quantum mechanical model,[1,2] while Paul Dirac, after studying the finer structure of these spectral lines, proposed a relativistic quantum framework incorporating electron spin.[3,4] Modern spectroscopic methods bridge the gap between quantum mechanics theory and experimental observations, serving as a digital fingerprint of the quantum physical world.

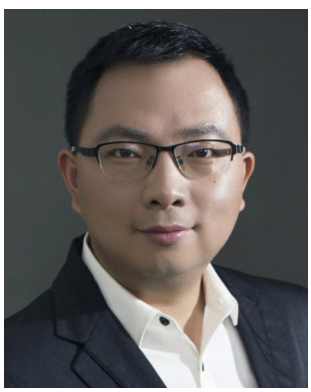
Leveraging its dual role as both a diagnostic tool and a bridge between microscopic and macroscopic scales, spectroscopy offers unique advantages that artificial intelligence (AI) can harness to deepen our understanding of matter (Fig. 1). Spectra possess key attributes: they are closely related to the microstructure of matter and can be obtained through theoretical calculations; they are experimentally measurable quantities that directly reflect macroscopic properties, and they digitally portray the physical world, serving as effective descriptors. These characteristics facilitate the association of spectroscopy with the structure and properties of substances. The introduction of AI technology enables the full exploitation of spectroscopic descriptors, facilitating a comprehensive understanding of the spectrum–structure–property relationship.

By leveraging the inherent characteristics of spectra that digitally represent matter, AI can effectively extract structural information of substances from spectral data, thereby facilitating efficient spectral interpretation. Furthermore, by capitalising on the unique attributes of spectra that bridge the gap between theory and experiments, AI can precisely predict spectra in the experimental environment from the theoretical structure. Additionally, with the aid of spectra that correlate with macroscopic properties, AI can also achieve accurate prediction from spectra to the functional attributes of substances. Due to its capacity to connect micro and macro characteristics, AI can invert the microstructure from macro effectiveness with the assistance of spectra. Given that different kinds of spectra portray the same material entity in different dimensions, AI is also capable of understanding and constructing the correlation between such dimensions and realising the complementation of information in different regions of the same spectrum, as well as between different spectra. By modelling the relationship among spectra, structures and properties, AI technology can facilitate the prediction of spectra from structures, the prediction of properties from spectra, the generation of spectra from properties, and the generation of structures from spectra (Fig. 1). Furthermore, it can provide insights into the complementarity of information between spectra. Ultimately, this enables a more profound comprehension of the spectrum–structure–property relationship.

The advent of AI in spectroscopic research has given rise to a wealth of reviews examining the intersection of spectroscopy,[5–20] AI,[21–23] and the deployment of AI in conjunction with specific spectroscopic techniques.[24–27] Thus, this

Yanbo Li

Yanbo Li is a postdoctoral researcher in Prof. Jun Jiang's group at the University of Science and Technology of China (USTC). She received her PhD degree in Nuclear Science and Technology from USTC in 2021, with research focused on combustion reaction kinetics. Her current research interests are focused on integrating theoretical simulations with machine learning techniques to investigate complex mechanisms in gas-phase and on-surface chemical reactions.



Jun Jiang

Jun Jiang holds a Chair Professorship at the University of Science and Technology of China (USTC), and is the Vice Director of the State Key Laboratory of Precision and Intelligent Chemistry. He has long been engaged in theoretical chemistry research combined with artificial intelligence and robotic chemistry. He is a recipient of the National Science Fund for Distinguished Young Scholars, the Young Theoretical Chemistry Investigator Award of the Chinese Chemistry Society, and the Distinguished Lectureship Award of the Chemical Society of Japan. He is also the founding editor-in-chief of the journal Artificial Intelligence Chemistry.

Yi Luo

Yi Luo currently holds a Distinguished Chair Professorship at the University of Science and Technology of China (USTC), and is the Director of the Hefei National Research Center for Physical Sciences at the Microscale. His research is focused on theoretical chemistry and intelligent spectroscopy. He was honoured with the Group Award for Outstanding Scientific and Technological Achievement Prize by the Chinese Academy of Sciences (CAS), and led a team of robotic AI-Chemists that was recognized as the Team of the Year by the CAS in 2022.

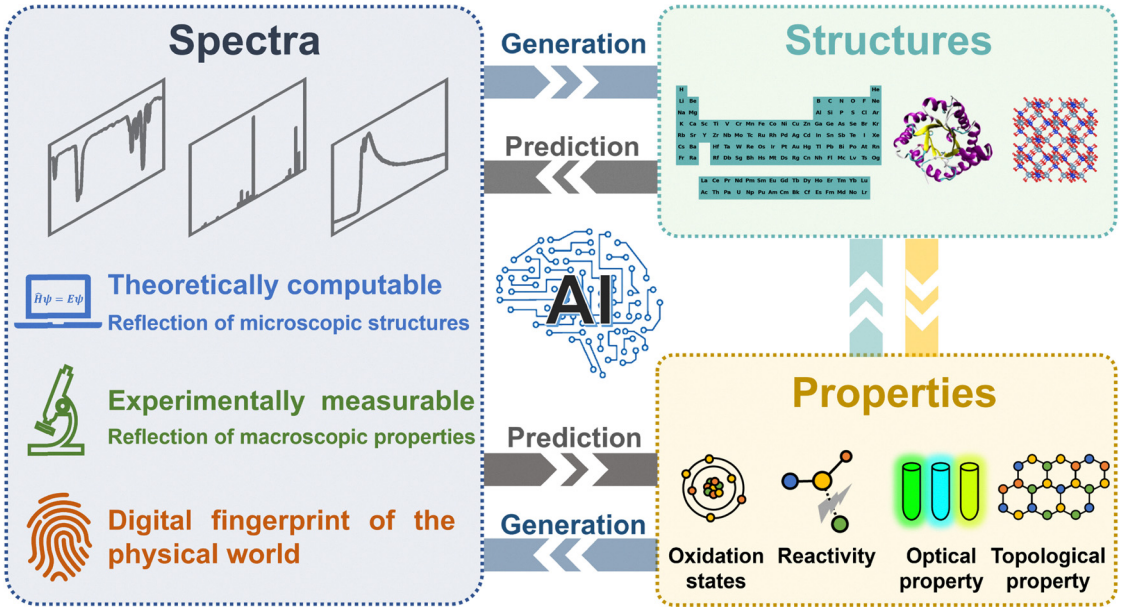


Fig. 1 The characteristics of spectra and the facility of AI in establishing the spectrum–structure–property relationship. As a theoretically computable and experimentally measurable descriptor, the spectrum can link the microscopic structure with macroscopic properties, and can be viewed as the digital fingerprint of the physical world. These characteristics enable spectra to act as a conduit between structures and properties, facilitating the establishment of the spectrum–structure–property relationship by AI. Using this relationship, machine learning models can make predictions from structures to spectra and from spectra to properties, as well as generate spectra from properties and structures from spectra.

review will focus on how AI has transformed the relationship among spectra, structures, and properties. It begins with a brief introduction to the major spectroscopy and AI techniques mentioned in this review. Then, we will review how AI has been employed to tackle conventional challenges in spectroscopy, such as efficiently calculating spectra from structures, resolving structures from spectra, and predicting properties based on spectra. Following this, we will explore the use of AI in addressing scientific tasks that are difficult to accomplish using traditional methods. These include the design of molecules and materials with specified properties based on spectra, as well as generating comprehensive spectral data from limited spectral data. Finally, we will examine the prospective advancements in intelligent spectroscopy with the aid of large models.

2. Background: spectroscopy and AI techniques

2.1 Spectroscopic techniques

The spectroscopic techniques used for matter research have expanded to include the entire electromagnetic spectrum, from infrared light to X-rays. These techniques are typically based on the interaction between matter and light. According to quantum mechanics theory, matter possesses discrete energy levels, and it can absorb or emit a photon whose energy exactly matches the gap between two energy levels. This absorption results in characteristic differences in the absorption or emission ratios of light at various frequencies, forming an absorption spectrum. These absorption and emission phenomena together provide valuable information about the molecular and electronic structure of matter. In various spectroscopic techniques, the observed transitions are generally caused by different particle motions, such as the vibrations and rotations of nuclei or electronic transitions between different orbitals.

In vibrational spectroscopy, the absorption of a photon excites a molecule between two different vibrational states, such as the $\nu = 0 \rightarrow 1$ transition in the O–H stretching vibration mode shown in Fig. 2. The corresponding spectra are typically observed in the infrared (IR) region of the electromagnetic spectrum. Both IR and Raman spectroscopy provide valuable information about molecular vibrations, but they differ in their underlying principles. IR spectroscopy measures the absorption of infrared light by molecules, where the vibration of a molecule must induce a dipole moment to interact with the infrared light. In contrast, Raman spectroscopy relies on the inelastic scattering of light, where the scattering intensity is related to the change in the polarizability of the molecule during vibration.

Ultraviolet-visible (UV-vis) spectroscopy measures the absorption of ultraviolet and visible light by molecules. The photon energies in this spectral region, typically on the order of electron volts (eV), are sufficient to promote valence electrons from occupied to unoccupied molecular orbitals (Fig. 2). This electronic excitation process makes UV-vis spectroscopy a valuable tool for probing the electronic structure of molecules, particularly that of valence electrons. The technique is especially sensitive to conjugated systems—molecules containing alternating single and multiple bonds or aromatic rings---which exhibit characteristic absorption patterns due to their delocalized π-electron systems.

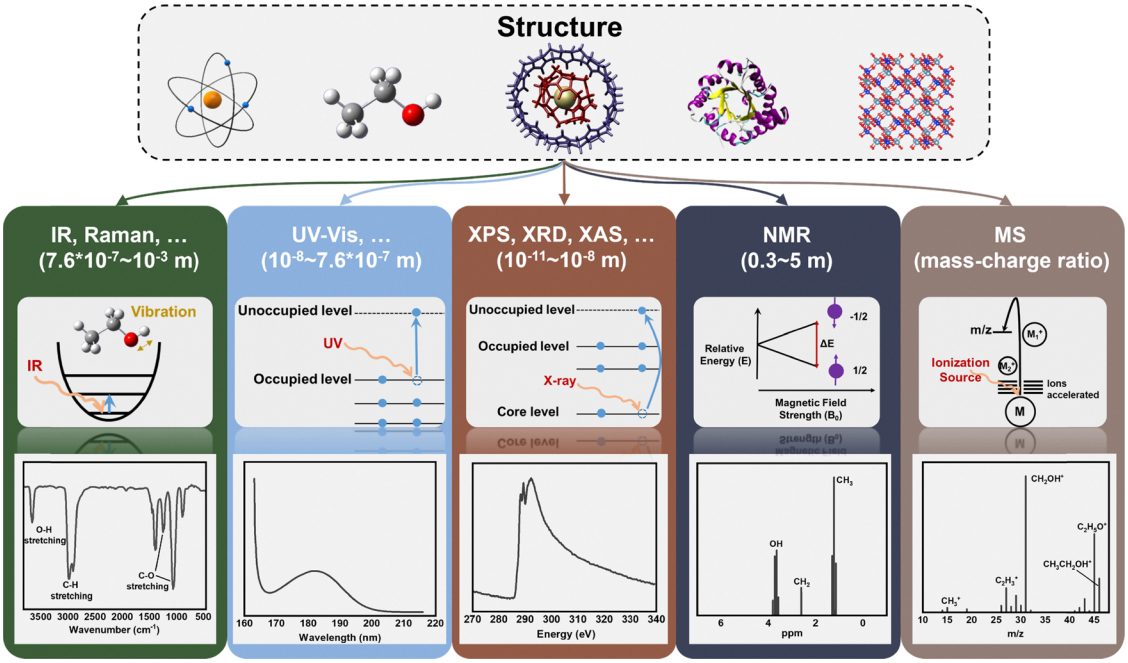


Fig. 2 Overview of the main spectroscopic techniques covered in this review. The IR,[28] UV,[29] XAS,[30] NMR[31] and MS[31] spectral data for ethanol are reproduced from refs.

Circular dichroism (CD) spectroscopy measures the differential absorption of left- and right-circularly polarized light by chiral molecules, providing unique information about molecular chirality that cannot be obtained from conventional absorption spectroscopy. CD can be divided into electronic circular dichroism (ECD) and vibrational circular dichroism (VCD). In ECD, the absorption of circularly polarised light by chiral compounds arises from transitions between electronic states. In contrast, in VCD, the absorption of circularly polarised light by chiral compounds corresponds to transitions between vibrational energy levels within the same electronic state. For chiral substances, the absorption of left-handed and right-handed circular polarised light differs. By measuring this difference, CD can detect the chirality of a substance without interference from non-chiral compounds.

Unlike the absorption spectroscopy methods discussed previously, fluorescence spectroscopy is an emission technique. When a sample absorbs a photon at a certain wavelength (usually in the ultraviolet or visible spectrum), the electrons inside the sample transition to an excited state. Subsequently, the molecule undergoes internal vibrational relaxation, emitting a photon of lower energy as it returns to the ground state—a process known as fluorescence emission. Fluorescence spectroscopy reveals information about not only the molecule's excited state properties, but also its molecular conformation, local environment, and intermolecular interactions. This technique is particularly useful for detecting compounds with fluorescence activity, such as molecules containing conjugated systems, aromatic ring structures, or specific fluorophores. Therefore, it has a wide range of applications in fields such as environmental analysis, chemical sensing, and biomarker analysis.

X-ray spectroscopy encompasses several techniques that utilise X-ray radiation to study the electronic, structural, and chemical properties of matter, which include X-ray absorption spectroscopy (XAS), X-ray photoelectron spectroscopy (XPS), and X-ray diffraction (XRD). The mechanism of XAS is similar to that of UV-vis spectroscopy, but the excited electron is from the inner shell of matter (Fig. 2). In XPS measurement, inner shell electrons in matter are ionised, and the kinetic energy distribution of these photoelectrons is measured. Both XAS and XPS have the advantage of element-specific sensitivity and the ability to detect local structural information, providing insights into the atomic-level composition and arrangement of molecules and materials. XRD, in contrast, exploits the wave nature of X-rays through diffraction by crystalline materials. When X-rays interact with the periodic atomic arrangement in crystals, they produce characteristic diffraction patterns. Analysis of these patterns enables the determination of crystal structure, unit cell parameters, and crystallite size and morphology.

Nuclear magnetic resonance (NMR) spectroscopy exploits the magnetic properties of atomic nuclei. In a strong external magnetic field, nuclear spin states split into discrete energy levels. When radiofrequency electromagnetic radiation matching the energy gap between these levels is applied, nuclei absorb this energy and transition between spin states (Fig. 2). Neighbouring atoms and electron density distributions create subtle magnetic field variations that shift resonance frequencies—a phenomenon known as chemical shift. This sensitivity to the local environment makes NMR a powerful tool for structural elucidation. Consequently, NMR has become an indispensable tool for identifying and characterizing organic compounds, determining protein structures, and studying molecular dynamics.

Mass spectrometry (MS) is an analytical method for measuring the mass-to-charge ratio of ions. It entails the ionisation of the components of a sample, generating charged ions with varying charge-to-mass ratios. These ions are then subjected to accelerating electric and magnetic fields, resulting in the formation of a spectrum (Fig. 2). Mass spectrometry's exceptional sensitivity and specificity make it invaluable for identifying organic molecules. Its analytical power is often enhanced through hyphenation with separation techniques: gas chromatography–mass spectrometry (GC-MS) for volatile compounds, liquid chromatography–mass spectrometry (LC-MS) for non-volatile and thermally labile molecules, and tandem mass spectrometry (MS/MS) for structural elucidation through controlled fragmentation. These combined approaches enable comprehensive molecular characterization across a wide range of applications, from pharmaceutical analysis to proteomics.

Numerous reviews have covered the fundamental principles, recent advancements, and applications of IR,[5,6] Raman,[7–10] UV-vis,[11] CD,[12] fluorescence spectroscopy,[13] NMR,[14] MS,[15–17] and X-ray spectroscopy.[18–20] Additionally, several reviews have examined the application of AI with specific spectroscopic methods, including aiding signal denoising and compound identification of IR[25] and Raman spectral data,[26] predicting IR and Raman spectra,[27] accelerated simulations of UV-vis spectra,[24] analysis of fluorescence spectra,[32–36] and the prediction and elucidation of NMR signals.[37,38]

2.2 Machine learning methods in spectroscopy

In recent years, a variety of AI techniques have been employed in the field of spectroscopy. These techniques encompass unsupervised learning algorithms, supervised learning algorithms, and neural network models, each of which has made a distinctive contribution to the field. In the field of AI, many excellent reviews have been published, offering valuable insights into various aspects of machine learning (ML). For instance, Kotsiantis *et al.*[21] provided a thorough review of supervised classification algorithms, covering logic-based methods, perceptron models, statistical techniques, and support vector machines (SVMs). LeCun *et al.*[22] authored a seminal work on deep learning detailing advanced neural network architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs). Jordan *et al.*[23] offer a comprehensive overview of ML applications, discussing current trends and future prospects. Collectively, these reviews showcase the transformative impact of advanced learning methods across various domains. This section will briefly outline several commonly used AI techniques for spectral research.

When dealing with spectral-related problems, the selection of machine learning methods is often not dependent on the type of spectra, but on the type of task to be solved. For instance, unsupervised learning is suitable for cluster analysis when the intrinsic correlation of unlabelled data needs to be mined. Conversely, when the correlation between variables needs to be constructed for target value prediction, supervised learning methods are often used to train the model based on labelled data. Traditional machine learning methods are suitable for small sample sizes and simple tasks. However, deep learning methods are often more appropriate for handling large amounts of data and modelling complex relationships between inputs and outputs.

Unsupervised learning techniques like Principal Component Analysis (PCA)[39,40] and hierarchical clustering[41,42] are fundamental for discovering patterns in spectral data without predefined labels. PCA is a widely used technique for reducing the dimensionality of complex datasets while preserving their essential patterns (Fig. 3A). In spectroscopy, PCA transforms the spectral data into principal components and identifies significant variations and trends, enhancing data interpretability. Hierarchical clustering, on the other hand, groups similar data points into clusters based on their features (Fig. 3B). This method can help identify relationships within spectral data, offering insights into underlying patterns. PCA and hierarchical clustering are unsupervised learning techniques, but PCA focuses on dimensionality reduction, while hierarchical clustering emphasises grouping and pattern discovery. Although both methods often fail to provide definitive conclusions directly, they can be used for initial data processing to establish a basis for subsequent careful analysis.

Supervised learning algorithms, including SVM,[43] Gaussian process regression (GPR),[44,45] decision trees,[46,47] and ensemble methods, enhance the precision of spectral data analysis by learning from labelled data. SVMs are effective classifiers that utilise the optimal hyperplane to differentiate data into categories, rendering them valuable for identifying molecular signatures and conducting quantitative analysis (Fig. 3C). GPR is a probabilistic model that provides predictions along with uncertainty estimates, rendering it an optimal choice for spectral data regression tasks where the confidence of predictions is of paramount importance (Fig. 3D). Decision trees are intuitive models that predict outcomes based on input features by learning simple decision rules inferred from the data (Fig. 3E). In spectroscopy, a decision tree divides the spectral data into branches based on specific features (*e.g.*, intensity at certain wavelengths), resulting in a decision at each branch until a final prediction is made at the leaf nodes. Each path from the root node to a leaf node represents a classification rule or a regression output. Random forests (RFs),[48] which combine multiple decision trees, enhance prediction accuracy and handle large spectral datasets by reducing overfitting (Fig. 3G). Gradient boosting[49,50] is a stagewise construction method for a model, whereby a sequence of decision trees is added sequentially (Fig. 3F). The construction of each new tree is informed by the errors made by the previous trees to minimise the loss function and thereby improve accuracy. XGBoost (eXtreme gradient boosting)[51] represents an advanced implementation of gradient boosting that incorporates several enhancements. These include regularisation, which helps prevent overfitting, and optimised computational speed through parallel processing. Similar to XGBoost, several proposals of gradient boosting algorithms have been developed to improve both speed and accuracy, such as LightGBM[52] and CatBoost.[53]

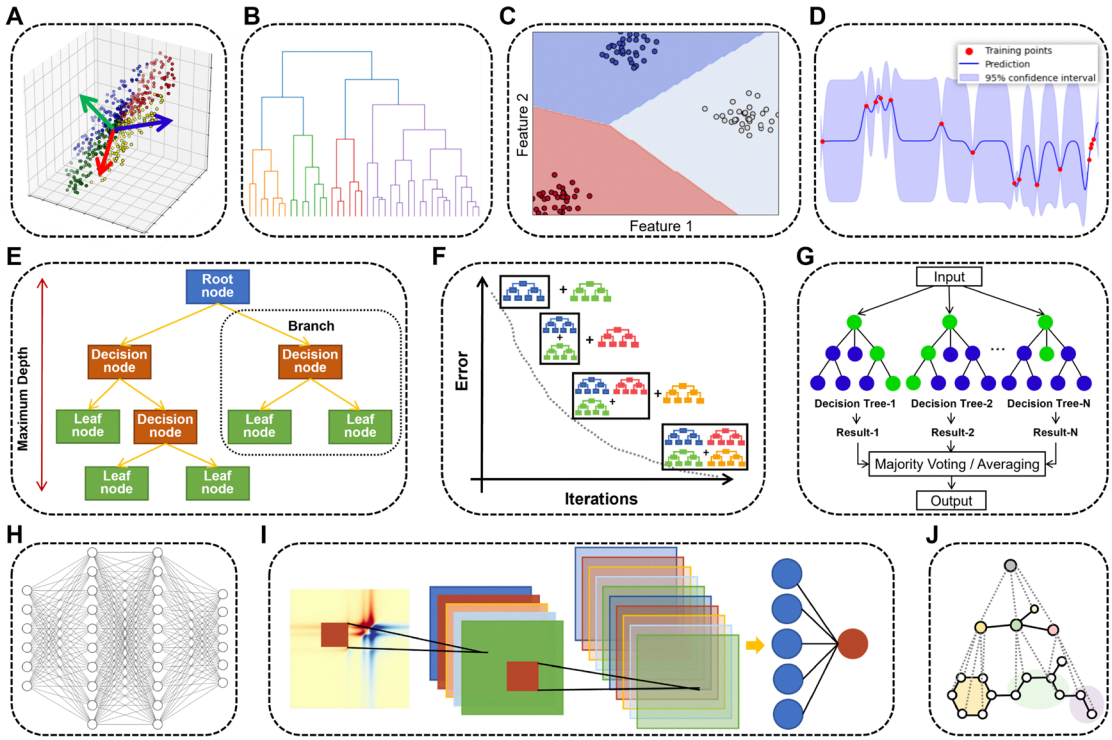


Fig. 3 Overview of AI techniques commonly used in spectroscopy. (A)–(J) Representative ML algorithms including PCA, hierarchical clustering, SVM, GPR, decision tree, gradient boosting, RF, MLP, CNN, and GNN, respectively.

These methods are highly efficient and often perform better on large and complex datasets than traditional gradient boosting methods.[54] These supervised learning models are suitable for situations where the output length is relatively small, *i.e.* where there are not many types of variables to be predicted. They are therefore more often used in studies that use spectra to discriminate between the types of substances.

Neural network models, including Multi-Layer Perceptron (MLP),[55,56] CNN,[57,58] Graph Neural Networks (GNNs),[59,60] and autoencoders,[61] demonstrate exceptional capabilities in capturing intricate patterns within data sets. A feedforward neural network, MLP is a type of artificial neural network that learns complex relationships within data through multiple layers of interconnected neurons (Fig. 3H). The structure of MLP is such that it is also referred to as an artificial neural network (ANN), a fully connected neural network (FCNN), or a deep neural network (DNN). CNN models specialise in processing data with grid-like topology (Fig. 3I), making them particularly effective for analysing both one-dimensional (1D) and two-dimensional (2D) spectral data, such as IR, Raman, and 2D NMR spectra. GNN models are particularly adept at modelling connected data and are, therefore, commonly employed to represent molecular structures (Fig. 3J). These models can capture the crucial atomic connectivity in predicting spectra. Autoencoders are neural networks designed for unsupervised learning tasks such as data denoising and dimensionality reduction. They improve spectral interpretation by learning compact representations of input data, leading to more accurate measurements. While each of these models is adept at learning complex data representations, selecting the appropriate model depends on the specific characteristics of the problem being addressed. For example, 1D CNN models excel at capturing intrinsic correlations between curves, making them well-suited to almost any spectral-based prediction problem, including predicting structural information[62] and interaction properties[63] from spectra, and structure-based spectral prediction.[64] In contrast, the structures of molecules and materials are more easily represented as graphs than spectra and other properties, making GNNs almost exclusively used in structure-based spectral prediction problems.

In addition to the models above, generative models are increasingly being used in spectroscopy research. The most common architectures are GANs and transformers. A GAN consists of a generator and a discriminator. The generator takes Gaussian-sampled random noise as input to generate samples that closely resemble real data. The discriminator then determines whether the generator's output meets the required standard and feeds the loss back to the generator. After the iterative training of the generator and discriminator, the GAN network can generate the required data. This architecture has been used to generate high signal-to-noise ratio spectra from low-quality spectra. Transformer, on the other hand, is based on the encoder–decoder architecture. The encoder embeds the sequence into a high-dimensional vector in the latent space and the decoder performs sequence generation based on this vector. The model is characterised by the introduction of the attention mechanism, which enables the model to establish complex associations between sequences. Since spectra can be

viewed as sequence information, and structures can also be represented as sequences using SMILES encoding, the transformer model has been widely used for both generating spectra from structures and the generation of structures from spectra.

Decisions regarding whether to use supervised or unsupervised learning, whether to handle a classification or regression task, and how to balance performance and interpretability when choosing a particular model are all tightly centred around specific task goals and are not dependent on the types of spectra themselves. For this reason, it is essential to retain the flexibility to apply suitable ML tools to different spectral analysis scenarios in order to solve problems in the most appropriate way.

2.3 Spectroscopy data

The success of AI models in spectroscopic analysis depends on the availability of high-quality, representative datasets. However, the spectroscopic community faces significant data challenges that vary considerably across different techniques. These challenges stem from the fundamental trade-offs between experimental authenticity and theoretical accessibility, along with the need for data augmentation strategies to bridge these gaps.

Experimental spectroscopic data remain essential as they capture real systems under actual measurement conditions. Beyond training AI models on privately collected data, various spectroscopy databases can be accessed for different types of spectra. Several multi-technique databases provide comprehensive coverage across multiple spectroscopic techniques, offering significant advantages for researchers requiring integrated analytical resources. The NIST Chemistry WebBook[28] stands as perhaps the most valuable free resource, encompassing IR, UV-vis, NMR, and MS data with thermochemical properties, making it an essential starting point for many analytical investigations. SDBS (spectral database for organic compounds)[31] provides integrated IR, NMR, and MS data for organic compounds with consistently high-quality spectra and reliable assignments. Wiley SpectraBase[65] also offers extensive multi-technique coverage, including IR, Raman, UV-vis, NMR, and MS data with powerful search capabilities and structural identification features. Beyond these comprehensive resources, different types of specialized databases have been developed for each spectroscopy technique. For example, the RRUFF database[66] serves as an exceptional free resource for Raman spectroscopy of minerals and crystalline materials, particularly valuable for geological applications. For CD spectra, the PCDDB (Protein Circular Dichroism Data Bank)[67] stands as the primary resource for protein secondary structure analysis, containing thousands of protein CD spectra with associated structural and experimental metadata. Nuclear magnetic resonance database platforms include specialized resources like NMRShiftDB,[68] an open-source platform combining chemical structures with experimental and predicted NMR data. Biological applications benefit from specialized databases, including Human Metabolome Database (HMDB),[69] for metabolite identification with comprehensive coverage. For mass spectra, the NIST Mass Spectral Library[28] remains the gold standard for electron ionization spectra, containing over 350 000 compounds with reliable fragmentation patterns and search algorithms. Open-access databases like MassBank,[70] CASMI[71] and GNPS[72] offer high-resolution MS/MS data across various ionization modes. Metabolomics applications benefit from specialized databases including METLIN[73] with over a million metabolites, HMDB for human metabolites, and LipidMaps[74] specifically for lipid analysis. X-ray diffraction databases centre around the International Centre for Diffraction Data[75] (ICDD) and their comprehensive powder diffraction file collection, containing over 900 000 entries covering inorganic, organic, and mineral phases with continuously updated releases. Free alternatives include the Crystallography Open Database (COD)[76] providing open-source crystal structures that can be used to generate calculated diffraction patterns, and the American Mineralogist Crystal Structure Database[77] specifically for geological applications. For X-ray spectroscopy, the NIST XPS Database[78] provides the most comprehensive free resource for X-ray photoelectron spectroscopy, offering binding energies, chemical shifts, and spectral features for elements across the periodic table in various chemical environments. XANES/NEXAFS databases are often maintained by individual synchrotron facilities, with resources available from major facilities. Other than these online spectroscopy databases, many instrumental manufacturers also provide commercial spectroscopic databases bundled with their equipment, which have been increasingly used for machine learning studies.

Experimental datasets face several inherent limitations, including scarcity due to the substantial time and cost involved in data collection and variability arising from differences in experimental conditions, instrumentation, and sample preparation protocols. The quality and quantity of experimental data can be particularly challenging to maintain consistently across different platforms and laboratories. Theoretical simulations have emerged as robust solutions to address these experimental data limitations by providing large quantities of reproducible spectroscopic data. Moreover, they can systematically generate extensive datasets, precisely control specific molecular or structural features, and provide insights into spectroscopic behaviour under ideal conditions. DFT calculations, which provide quantum mechanical solutions to the electronic structure, have been widely used to generate spectroscopic data, including vibrational, UV-vis, CD, NMR, and X-ray spectra. High-level wavefunction methods provide accurate reference data for Δ-machine learning studies to balance accuracy and computational efficiency. In contrast, semi-empirical or empirical approaches enable high-throughput calculations for larger systems. In addition to standard DFT calculations, complementary methods have been developed for specific spectroscopic applications. Classical molecular dynamics simulations offer complementary advantages for vibrational spectroscopy by accounting for anharmonic effects absent in the normal mode analysis. For NMR spectra, modern prediction tools have become increasingly sophisticated, with platforms such as ChemDraw,[79] MestReNova (Mnova),[80] and ACD/NMR

Predictor[81] offering structure-based spectral simulation capabilities that effectively complement experimental databases. In X-ray spectroscopy, real-space Green's function methods such as FEFF[82] have been successfully employed for high-throughput data generation. Since X-ray diffraction (XRD) spectra calculation requires minimal effort from accurate crystal structures, XRD data are generally evaluated using databases such as the Cambridge Structural Database (CSD)[83] for small molecule organic and organometallic structures, and the Inorganic Crystal Structure Database (ICSD)[84] from FIZ Karlsruhe for inorganic materials.

Data augmentation strategies have become particularly important for bridging the gap between theoretical predictions and experimental reality, especially for theoretical spectra where noise must be added to mimic experimental conditions. Different spectroscopic techniques employ technique-specific augmentation approaches tailored to their unique characteristics. In IR spectroscopy, three primary augmentation methods are commonly used to increase the performance of ML models: horizontal shifting, vertical noise addition, and linear combination techniques.[62] Data augmentation is particularly crucial in Raman analysis due to insufficient data per class, often requiring interpolation to ensure all spectra are sampled at common wavelengths for consistent model input.[85] Additionally, random uniform noise proportional to the square root of the signal is added to mimic dispersive detector noise, increasing sample variance and producing more robust networks.[86] For X-ray diffraction, physics-informed data augmentation techniques incorporate experimental artifacts such as strain, texture, and domain size to perturb diffraction peaks,[87–91] with noise typically introduced using random signals drawn from uniform distributions.[92,93] X-ray absorption spectroscopy employs noise-based data augmentation techniques to simulate realistic experimental conditions. In mass spectrometry, training datasets can be generated based on natural abundance ratios without requiring human labelling, with workflows validated on experimental datasets from atom probe tomography (APT) and secondary ion mass spectrometry (SIMS).[94] These augmentation strategies are essential for expanding limited datasets and making theoretical spectra more representative of experimental measurements, ultimately improving the robustness and generalizability of AI models across different spectroscopic techniques.

3. From spectra to structures

Efforts have been dedicated to efficiently extracting structural information from matter based on its spectra using AI models. Depending on the specific tasks, various AI models have been employed. For distinguishing molecular types and identifying absolute configurations from vibrational or CD spectra, simpler machine learning models like SVM and RF can be effective. More complex models, such as FCNN and CNN, enable more detailed spectral analyses. They can determine information such as the presence of functional groups and the composition of nanoclusters. For intricate systems, including mixtures and unknown substances, AI models can even identify phase and phase-fraction information, thereby extending the scope and depth of spectroscopic studies (Fig. 4). The following section provides a comprehensive overview of the recent advancements in utilizing ML to extract structural information from different types of spectra.

3.1 Vibrational spectra

IR and Raman spectra provide valuable information regarding the interactions between molecular vibrations and light. The vibrational features of a molecule are closely connected to its functional groups, such as C═O and O–H, which have characteristic absorption peaks in the infrared range. Many intelligence models[62,98,99] have centred on identifying specific functional groups within a molecule using its IR spectra. This capability allows ML-assisted vibrational spectroscopy to surpass the abilities of traditional methods in determining the presence of substances. Additionally, it can detect and identify unknown substances, determine the functional group structure and surface microstructure of molecules, and even extract compositional information about substances from the spectra.

Building on the valuable insights into vibrations provided by IR and Raman spectroscopy, emerging ML strategies have focused on identifying functional groups. An early approach by Ren *et al.*[98] revealed that the FCNN model based on the structural descriptors can predict the IR/Raman frequencies and intensities of O–H and C═O stretches. They further built a long short-term memory (LSTM) network to identify hydroxyl or carbonyl groups in the molecule based on the IR or Raman spectra, with recognition accuracies of 99.4% and 98.5%, respectively. More importantly, recognition accuracy can be significantly improved by correlating the information extracted from IR and Raman spectra. Also, their model can emulate human-expert-based structure recognition by interrelating chemical information extracted from IR and Raman spectra. LSTM model has the capability to integrate supplementary spectral or non-spectral data that encompass additional structure–property relationships, thereby boosting its performance. The two methods, which have been designed for the purpose of spectrum prediction and structure recognition, demonstrated robust transferability, thus indicating significant promise for their application in a range of spectroscopic analyses and chemical identification tasks. CNN models were also reported to effectively classify the presence or absence of functional groups from the IR spectra.[62,99] CNNs represent a specialized architecture, designed to process data with a grid-like topology, such as spectral data. CNNs employ convolutional layers that apply localized filters across the input data, enabling them to detect local spectral features regardless of their position in the spectrum. In this model, Rieger *et al.* converted human-understandable features into the final output classification, allowing the model to learn the important features automatically.[99] In addition to excellent predictive performance, this approach offers the model robustness and interpretability. This interpretability is further supported by the

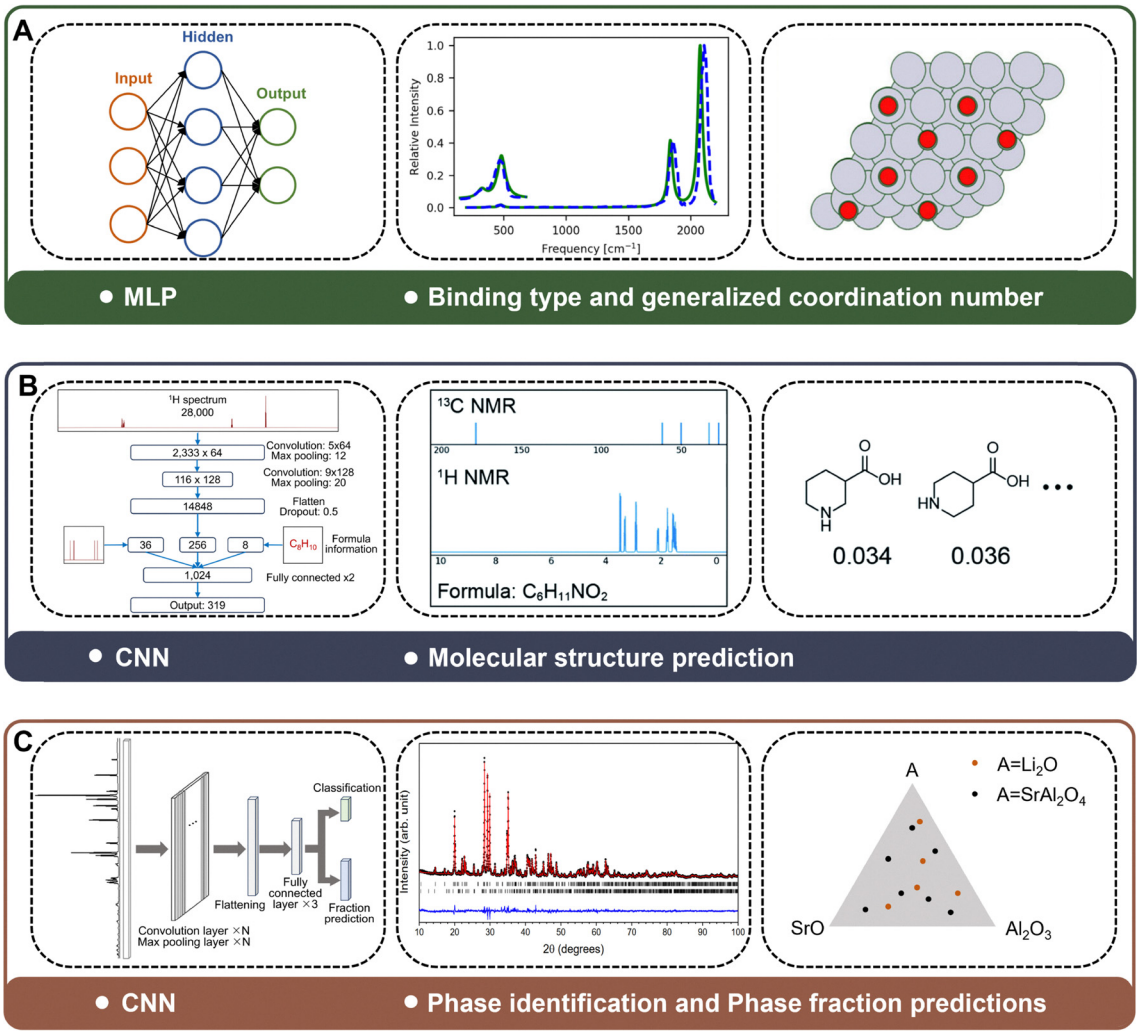


Fig. 4 Representative examples of using ML models to obtain structural information from different types of spectra. (A) Using MLP to predict the binding type and generalized coordination number of the surface of Pt nanoparticles *via* vibrational spectra. Reproduced with permission from ref. 95. (B) Using CNN to predict molecular structure, with ^{1}H NMR, ^{13}C NMR and molecular formula as input. Reproduced with permission from ref. 96. (C) Using CNN for X-ray spectroscopy analysis to identify the phase of a three-compound mixture and predict its phase fraction. Reproduced with permission from ref. 97.

alignment of the patterns learned by the model with the patterns experts used to assign spectra by visual inspection. Moreover, the model also unravels new and non-intuitive patterns learned by the neural network. With the advancement of their ML models, the ability to extract information from IR spectra has expanded significantly. The number of functional groups that CNN can identify has increased to 37 (Fig. 5A),[62] with strong potential for further growth in the next few years.

The application of ML to vibrational spectroscopy has expanded significantly beyond functional group identification.[95,102] In surface science, Lansford and Vlachos studied CO adsorption on Pt nanoparticles. They first calculated the DFT frequencies and intensities at low CO coverage. Then, successive layers of physics-based surrogate models were applied to these data to produce a supplementary dataset of synthetic IR spectra featuring arbitrary combinations of adsorption sites, which were then used to infer structure. The microstructure was described using binding type (atop, bridge, threefold, or fourfold) and the generalised coordination number probability distribution functions. A neural network model is built to map the IR spectra and the microstructure, allowing accurate prediction of the microstructure from both the synthetic and experimental IR spectra. By including the low-frequency range of the spectra, the models could further improve predictions and extend to NO on Pt nanoparticles.

Recent advances in ML have paved the way for innovative molecular identification techniques that can analyse complex mixtures in great detail. A molecular identification platform based on a wavelength-multiplexed hook nanoantenna array (WMHNA) was introduced to collect spectral data (Fig. 5B) of mixed alcoholic solutions.[100] PCA transforms high-dimensional spectral data into a reduced set of orthogonal components that capture maximum variance. This unsupervised technique projects the original data onto a lower-dimensional space defined by principal components, which are linear combinations of the original variables ordered by the

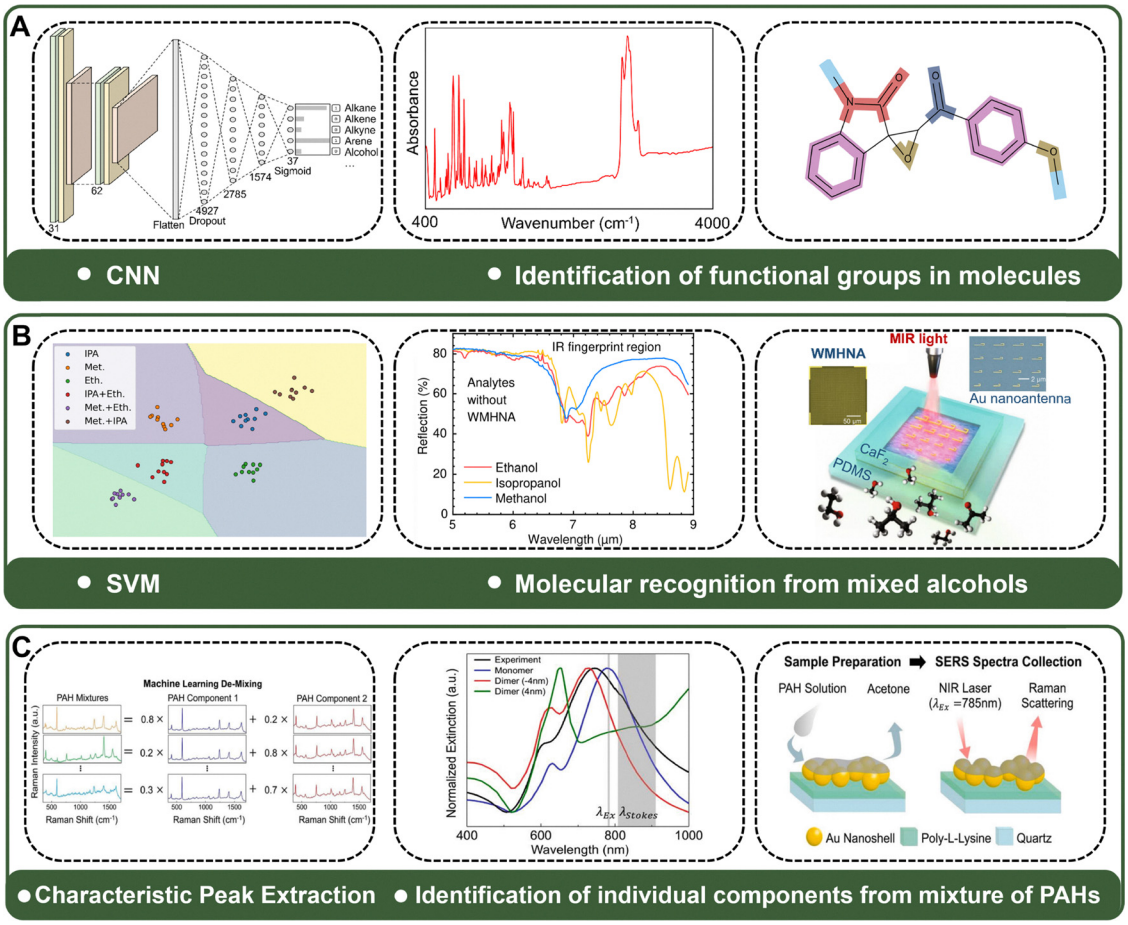


Fig. 5 Representative examples of using ML models to obtain structural information from vibrational spectra. (A) Using CNN to identify functional groups in molecules from IR spectra. Reproduced with permission from ref. 62. (B) Using SVM to recognize compounds from mixed alcohols *via* IR spectroscopy. Reproduced with permission from ref. 100. (C) Developing ML algorithms for the identification of individual components from a mixture of PAHs *via* Raman spectroscopy. Reproduced with permission from ref. 101.

amount of variance they explain. SVMs identify optimal hyperplanes that separate different classes within the feature space, with a margin maximization objective that enhances generalization capabilities. When combined with SVMs, PCA creates a robust framework for molecular identification in complex mixtures. The proof-of-concept test for recognizing molecules in a mixture of alcohols, namely methanol, ethanol, and isopropanol, achieved 100% identification accuracy by using PCA alongside an SVM. The WMHNA platform has been demonstrated to exhibit remarkable sensing capabilities, which are derived from loss engineering and wavelength multiplexing. These capabilities establish a pathway for ultra-sensitive on-chip molecular identification in mixtures of diverse species. This advancement is significant for the frontiers of chemical and biomolecular analysis. Another advancement in ML-assisted molecular identification[103] surpasses traditional ML approaches reliant on peak matching or single functional group identification. By combining IR spectra and MS, Fine *et al.* devised a multi-class, multi-label classification approach that can predict the various functional groups of a molecule without the need for a database. The proposal of a framework for the evaluation of the performance of the multi-label neural network was further pursued, with metrics such as the molecular $F1$ score and the molecular perfection rate being defined for this purpose. This framework and methodology may catalyse further development in functional group identification, leading to accurate and autonomous molecular structure elucidation.

Recent advances in spectroscopic analysis have seen researchers successfully apply generative models to not only resolve complete molecular structures from vibrational spectra, but more importantly, generate novel molecules. Alberts *et al.* successfully applied the transformer model, which is widely used in natural language processing, to infrared spectroscopy.[104] They successfully used the transformer model to parse IR spectra by encoding the spectral curves as integer values and inputting them as sequences, as well as using SMILES as the structure descriptor. The model's prediction accuracy reached a top-1 performance of 45.33%, top-5 of 72.21%, and top-10 of 78.50%. The transformer framework can be used to generate molecular structures in multiple ways. Kanakala *et al.* proposed a model architecture that combines

contrast learning and transformers, enabling the direct generation of molecular structures.[105] Rather than using spectra and structures directly as input and output, they encoded both into a hidden space using an encoder. They then realised the alignment of both using contrast learning in hidden space and trained the decoder using a masked transformer, with spectra as the initial token and molecular structures as subsequent tokens, thus achieving highly accurate structure generation. Lu *et al.* employed a similar approach to align vibrational spectra and molecular structures in the hidden space, utilising spectroscopy to ascertain the similarity of spectra and structures, and generate novel structures.[106] They tested the prediction accuracy of IR, Raman, and two-spectrum coupling simultaneously. The results showed that Raman outperformed IR, while two-spectrum coupling slightly improved the performance of the Raman spectrum. Given the development of multimodal technology, using NMR, MS, and other spectra for coupling is expected to further improve model performance, which could be an important future direction for spectroscopic analysis.

Beyond its extensive application in interpreting IR spectra, ML approaches can also aid in analysing Raman spectra, especially the widely used surface-enhanced Raman spectra (SERS). In a recent work[107] by Luo *et al.* aiming to achieve highly selective and sensitive SERS analysis, a novel framework named Vis-CAD, which combines visual RF, characteristic amplifier, and data augmentation, was developed. RFs comprise multiple decision trees trained on distinct subsets of the data and features. Each tree independently classifies the input, and the final prediction is determined by majority voting (for classification) or averaging (for regression). This ensemble approach reduces overfitting and improves generalization compared to single decision trees. This framework has the advantage of lowering the demand for mass data for training and offers high interpretability. It was exemplified by the SERS trace analysis of polycyclic aromatic hydrocarbons (PAHs). Initially, they used data augmentation techniques to create a large collection of simulated SERS spectra that represent a wide variety of mixtures. Then, these simulated SERS spectra were used to build an RF model that incorporates a characteristic amplifier. This model achieved an accuracy of no less than 99% in qualitative analysis of each PAH in different mixtures and notably improved the sensitivity of SERS detection by at least one order of magnitude. Furthermore, the model's interpretability is highly reliable due to its accurate capture of the characteristic peaks (CPs) associated with the target. Aside from RF, a series of ML methods have also been applied to identify individual components from the SERS spectra.[101,108] These methods, called Characteristic Peak Extraction (CaPE) and Characteristic Peak Extraction (CaPSim), can extract characteristic peaks from SERS spectra based on the number of counts at locations where peaks of the mixture are detected (Fig. 5C). In this way, CaPE is not too sensitive to the specific locations of high-intensity peaks, making it robust to local frequency shifts of CPs, leading to its capability to handle multiple compounds. Specifically, this algorithm identifies individual components and their relative concentrations from a mixture of 5 different PAHs. To extend the method to more general studies of molecule identification, researchers developed CaPSim, which combines SERS with a Raman library. This enables the calculation of the similarity between a query SERS and each spectrum in the Raman library. Furthermore, CaPSim is capable of successfully identifying unknown chemicals based on their SERS spectrum.[108]

With the ability of substance identification, ML plays a vital role in heavy metal ion detection, which involves determining whether the number of target heavy metal ions has exceeded a regulatory threshold. This challenge can be approached as a binary classification problem, making supervised binary classification models particularly appropriate for tasks involving SERS measurements.[109] In particular, Park *et al.* evaluated the efficacy of various preprocessing techniques and ML methodologies for heavy metal ion detection *via* SERS.[109] Among these, the most effective preprocessing approach involved utilising the Baseline Correction (BC) technique to eliminate low-frequency components, which were identified as background noise. The Radial Basis Function (RBF) kernel, in conjunction with the SVM classifier, exhibited superior performance in the detection of $Pb(NO_3)_2$ molecules in comparison to other ML models. Transfer learning models pre-trained on a standard Raman spectral database can also significantly improve the accuracy of Raman spectral interpretation using limited data.[85]

In contrast to traditional spectral classification that assigns spectra to predefined categories, some spectroscopic analysis approaches focus on assessing similarity between spectral pairs to identify related molecular structures.[85,110–113] A typical model, which is based on a Siamese neural network,[114] organises the spectra into pairs and computes a similarity score between them.[110] Siamese neural networks represent a specialized architecture for comparing the similarity between two inputs. These networks consist of two identical subnetworks with shared weights, processing two inputs in parallel and computing a similarity metric between their encoded representations. Siamese networks are particularly valuable for spectroscopic applications with limited training examples per class, as they learn generalizable similarity metrics rather than specific class features. After training the model, a new spectrum can be classified by computing its representation and similarity score with reference spectra. The efficacy of the model was evaluated on disparate datasets, including two Raman and one SERS. The methodology does not necessitate intricate preprocessing techniques, such as baseline correction, smoothing, or normalisation. Additionally, the model is capable of accurately quantifying the uncertainty associated with its predictions, which can serve as an indicator of potential inaccuracy.

Advanced ML architectures can revolutionise Raman spectroscopy by reducing noise effectively and isolating subtle spectral features, thus enhancing data interpretation. The combination of ML with image processing techniques offers a potent approach for the analysis of the spatial and temporal characteristics inherent in diverse SERS datasets. Poppe *et al.* demonstrated the efficacy of a one-dimensional convolutional autoencoder (CAE) in the reconstruction and elimination of the

constant background of 'nanocavity' spectra.[86] CAEs combine convolutional layers with an encoder–decoder architecture. The encoder compresses the input data into a lower-dimensional latent representation, while the decoder reconstructs the original input from this compressed form. This architecture excels at unsupervised feature learning and noise reduction. Subsequently, an iterative threshold detection procedure is applied to the residual spectra with the objective of isolating the subtle picocavity peaks. Finally, the extracted single-molecule spectra are clustered based on their similarities, which helps to uncover recurrent interaction patterns and achieve atomic-level precision in pinpointing the formation sites of adatoms relative to the studied molecule. This study provides a unique insight into the formation behaviour and the coordination geometries of adatoms on metal surfaces. ML approaches can also enhance the Raman signal-to-noise ratio, which is often vulnerable to the intrinsic noise from the instrument. He *et al.* used a modified 1-D deep convolutional neural network, Attention U-net, to learn the instrumental noise by fitting the mapping function between high and low signal-to-noise ratio spectra.[115] The U-net architecture features a contracting path that captures context and a symmetric expanding path for precise localization, while the attention mechanism focuses the network on the most relevant spectral regions. This structure is particularly effective for learning the statistically stable, instrument-specific noise patterns in the frequency domain, allowing the network to differentiate between genuine spectral features and noise without requiring extensive sample-specific training data. By removing the predicted instrumental noise, this approach enhances the signal-to-noise ratio of Raman spectroscopy by about 10 folds and suppresses the mean-square error by almost 150 folds.

In summary, the introduction of AI significantly enhanced the interpretation of IR and Raman spectra. The FCNN model allows for accurately predicting functional groups, such as hydroxyl and carbonyl, from IR spectra. By combining information from both IR and Raman spectra, these models improved recognition accuracy, even mimicking expert-level analysis. The evolution of CNNs advanced this field by automatically learning important spectral features, some previously unrecognised by human experts. ML techniques have also expanded the range of detectable functional groups and provided deeper insights into molecular surface microstructures, as demonstrated in studies involving CO adsorption on Pt nanoparticles. Additionally, the combination of ML and novel sensing platforms, such as WMHNA, has enabled the precise identification of molecules in mixed solutions, including complex alcohol mixtures. Using RF and SVM as classification models, researchers can also efficiently identify individual components from SERS. This capability can be further extended to the identification of unknown chemicals by comparing the similarity of SERS with spectra in a Raman library. Additionally, ML techniques have the capacity to enhance the signal-to-noise ratio of Raman spectra, with Attention U-nets being particularly effective in reducing instrumental noise and improving spectral quality.

3.2 UV-Vis spectra

Compared to the vibrational spectra, UV-vis spectra have fewer features, making it challenging for humans to determine chemical compositions accurately. Thus, instead of using UV to obtain structural information about a completely unknown sample, researchers often analyse the variability of its components and structure in a specific system. However, ML methods have shown the ability to establish the correlations between the UV-vis spectra and the chemical compositions even if the spectra are "featureless."[116] A single UV-vis absorption spectrum of a nanocluster (NC) sample may contain hundreds of features, beyond the capability of human conceivability to analyse. A one-dimensional convolutional neural network (1D CNN) was trained using 454 UV-vis absorption spectra of $[Au_n(SR)_m]^q$ NC samples with their corresponding experimentally derived composition information. The predictions of the composition of NC have a mean absolute error (MAE) of 0.0053 in the test set, indicating good agreement with the experiment. By using these predictions, strong correlations are shown between the UV-vis absorption profiles of metal NC mixtures and their individual compositions. This means that the chemical makeup of metal NCs can be found quickly and accurately by looking at their UV-vis spectra. Researchers have also successfully facilitated the prediction of structural motifs in multi-functional intermediates based on Vacuum UltraViolet (VUV) absorption spectroscopy.[117] Various ML techniques, including PCA, Partial Least Squares Discriminant Analysis (PLS-DA), and decision tree classifiers, have identified different molecular structures. PCA performs well in initial dimensionality reduction for spectral data, while PLS-DA excels by maximizing the covariance between spectra and structural classes, making it ideal for supervised spectral classification tasks. The target is to perform a simple molecular classification comprising five distinct categories: alkane, conjugation with oxygen, non-conjugated alkene, oxygen-containing, or cyclic. PCA revealed that unsupervised clustering methods can only distinguish non-conjugated alkenes from other categories. Supervised clustering methods, such as PLS-DA and decision trees, can classify the molecules with an accuracy higher than 75% for all five categories. This study demonstrates that VUV absorption spectra can differentiate molecular structures, such as constitutional isomers and methylated aromatics, showing the effectiveness of combining VUV spectra with ML for species identification.

Compared to the use of ML to extract information about the composition of substances in UV spectra, recent studies on the UV spectra of protein have proven that ML protocols can even monitor the dynamic structural change of protein *via* Far Ultra Violet (FUV) spectroscopy. In the work by Zhang *et al.*,[118] the prediction of FUV spectra for various proteins was first made *via* ML methods, which lower the computational cost by 3–4 orders of magnitude compared to DFT. Using the ML protocol, they harvested 1000 conformations *via* molecular dynamics (MD) simulation and computed their FUV spectra. The averaged spectra are in good agreement with the experiment. In order to facilitate real-time observation of protein dynamics by

means of time-resolved spectroscopy, the present study integrated MD simulations with a ML-driven simulation of FUV spectra. This approach enabled the uncovering of the temporal evolution of the mini Trp-cage protein's FUV spectra throughout its folding process. Four molecular descriptors were tested in predicting peptide and residue transitions ($n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ at $\sim$220 nm and $\sim$190 nm): internal coordinates, Coulomb matrix (CM), bag of bonds (BOB), and atom-centred symmetry functions (ACSFs). The ML models for the excitation energy ε_0, transition dipole moments μ_T of peptide bonds, and the ground state dipole moment μ_G of residues were applied to predict these parameters for new proteins, construct exciton Hamiltonians, and obtain FUV spectra. For 12 proteins randomly selected from the RCSB Protein Data Bank, the ML-based approach strongly agreed with DFT-based predictions regarding peak positions and line shapes, as evidenced by Spearman's rank correlation coefficients above 0.80. Extending the analysis to 230 proteins confirmed the robustness and transferability of the ML model due to a diverse training dataset, carefully selected molecular descriptors, and optimised hyperparameters. Furthermore, FUV spectroscopy, sensitive to various processes and environmental factors, revealed distinct spectra for wild-type γS-crystallin and its cataract-associated variants G106V and G18V, indicating sensitivity to minor structural variations. Combining MD simulations with ML-based FUV spectral simulations, the time-dependent evolution of FUV spectra was tracked for the mini Trp-cage along its folding path, highlighting changes in the secondary structure and spectral features. These results demonstrate the potential of ML-based FUV simulations for real-time monitoring of protein dynamics and structural changes.

The close correlation between protein spectra and their secondary structures enables the construction of intelligent models for the inversion of structures from spectra, thereby achieving direct structural recognition. In recent studies, Ren *et al.* successfully used two-dimensional UV (2DUV) spectra to predict the secondary structure of proteins (Fig. 6).[119] Both one- and two-dimensional UV spectra were tested in this task. The test results indicate that the prediction accuracies of the one-dimensional linear absorption (LA) and CD spectra are not high, whereas the 2DUV spectra can achieve a prediction accuracy close to 100%. This indicates that the exciton coupling information in the cross peaks of 2DUV spectra is crucial in predicting protein secondary structures. The outcomes demonstrate that 2DUV spectroscopy is an excellent descriptor for recognising protein secondary structures, with an accuracy of 97% and 91% in identifying the secondary structure of randomly selected homologous and non-homologous protein fragments, respectively.

Above all, AI technology has significantly advanced the analysis of UV-vis spectra, especially when dealing with "featureless" spectra that are hard to interpret. ML models enable the precise prediction of molecular structures and chemical compositions, such as the rapid identification of metal nanoclusters[116] and real-time tracking of protein structural changes.[118] Compared to traditional quantum chemistry methods, ML drastically reduces computational costs while maintaining high accuracy, particularly in predicting protein secondary structures and long-range molecular interactions. Additionally, ML models demonstrate excellent transferability across various molecular systems, showing potential in material design and catalysis research. It can efficiently extract key spectral information from complex molecular systems, providing powerful tools for understanding reaction mechanisms, discovering functional materials, and studying dynamics of biological systems.

3.3 CD spectra

The resolution of CD spectra has primarily centred on the absolute conformations of small chiral molecules and the conformational changes of macromolecules, such as proteins. It has been demonstrated that these processes can be effectively realised using ML methods. ML techniques have been used to extract previously unknown spectral features from VCD spectra and, in this way, allow for determining the absolute configuration (AC) of new compounds.[120] Various supervised and unsupervised methods have been applied to a data set of substituted α-pinene: PCA, *t*-stochastic neighbour embedding, decision tree, logistic regression, Naive Bayes, SVM, *k*-nearest neighbours (KNN), RF, and MLP. Of the various machine learning approaches used for AC determination, RF and MLP models are the most effective. Under optimal conditions, the RF model achieves a predictive accuracy of 0.940, whereas a shallow MLP can reach an accuracy of 0.995. The RF model can offer better interpretability than MLP by enabling the extraction

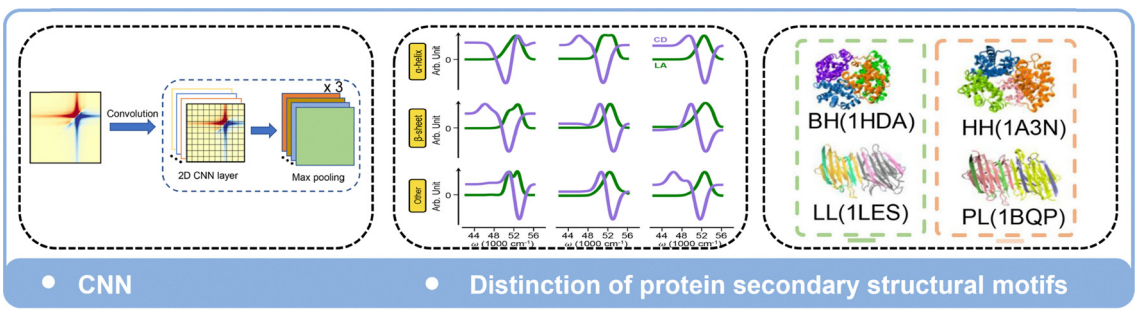


Fig. 6 Representative example of using the CNN model to recognise the secondary structure of a protein from 2D UV-vis spectra. Reproduced with permission from ref. 119.

of key spectral regions that are essential for AC determination. The results also show that changing the level of theory for input spectra does not significantly impact the ability of RF or MLP to establish AC. Although how the ML models extract AC-related information varies, the robustness of the ML approach ensures consistent performance across different levels of theories.

VCD has recently been employed to assign the predominant solution-state conformations simultaneously and to perform a configurational analysis of orbitides.[121] The simulation and subsequent interpretation of the experimental VCD spectra of peptides is notoriously challenging. To overcome the inherent uncertainties in DFT-calculated relative energies, a GA-VCD method has been developed. This method uses a genetic algorithm to fine-tune the Boltzmann populations of the selected conformers within a predetermined uncertainty range. This achieves an optimum match between experimental and simulated spectra. Genetic algorithms are applied for spectral analysis of complex peptides because they can efficiently search vast conformational spaces and optimize multiple parameters simultaneously, mimicking natural selection to find optimal solutions in problems with high uncertainty. This strategy is particularly effective when an accurate estimation of Boltzmann factors is impractical, as is often the case with larger, flexible peptides analysed in polar solvents with hydrogen bonding.

With ML methods, ECD spectroscopy has been used to distinguish intrinsically disordered proteins (IDPs) from ordered proteins.[122] In their model, Micsonai *et al.* used CD data at only three wavelength points ranging from 175 nm to 250 nm to make high-throughput data collection and detection possible. The performance of different methods on a couple of wavelength triplets is examined. For best classification accuracy, the CD of the protein should be measurable in good quality down to 197 nm. The mathematical analysis should use the KNN algorithm with the cosine distance function, independent of the spectral amplitude, *i.e.*, free from concentration determination errors. KNN with cosine distance is especially effective for ECD spectral analysis because it focuses on the pattern similarity rather than absolute magnitude, making it robust against concentration variations and experimental noise that commonly affect CD measurements.

Therefore, AI techniques have enhanced the analysis of CD spectroscopy, particularly VCD and ECD, which are crucial for studying chiral molecules and protein structures. ML models like RF and MLP have demonstrated high accuracy in determining the AC of chiral compounds from VCD spectra, with RF offering interpretability by pinpointing important spectral regions. Furthermore, genetic algorithms combined with VCD address the complexities of peptide conformations, while ECD has been used to distinguish between intrinsically disordered and ordered proteins using minimal spectral data. This approach to high-throughput protein classification represents a significant advancement.

3.4 Fluorescence spectra

Fluorescence spectra are often used as fingerprints of fluorescent molecules due to their multidimensional characteristics, including their characteristic excitation and emission wavelengths, and fluorescence intensities. The applications of this technique include the detection of substances based on fingerprinting features and fluorescence imaging using specific molecules. The ML approach is effective in both types of scenarios to enhance one's resolution of fluorescence spectra.[33,35,36]

In substance detection, ML is widely used to enhance the accuracy of pollutant species and content detection. For instance, Tian *et al.* employed fluorescence spectroscopy of carbon quantum dots for the detection of heavy metal atoms, utilising a series of models including SVM, K-mean, CART, RF, KNN, CNN.[123] Among these models, the optimal model enhanced the detection sensitivity of Cu^{2+} by 4–6 orders of magnitude, achieving a detection level of 100 pM, which is a significant improvement over the conventional method. Similarly, ML has demonstrated remarkable accuracy in the quantitative detection of substances such as hazardous compounds,[124] microplastic particles,[125] and antibiotic concentrations.[126] ML-assisted fluorescence spectroscopy can also be used for substance classification tasks such as identifying olive oil,[127] petroleum[128] and even honey.[129] In a recent study, Förste *et al.* utilized confocal micro-area XRF to train neural networks that could accurately predict the content of 53 elements and simultaneously acquire information on sample density, surface location, and more.[130] These advances show that ML is making fluorescence spectroscopy techniques more accurate and versatile.

In the domain of fluorescence image processing, ML techniques are frequently employed to enhance the resolution of an image, thereby leading to a substantial reduction in the light intensity necessary for molecular detection. For example, Kagan *et al.* used machine learning algorithms to enhance the resolution of monolayers of carbon nanotubes, thereby achieving images of superior clarity.[131] The images predicted by ML algorithms offer a signal-to-noise enhancement that surpasses that of images obtained through super-resolution radial fluctuation algorithm. In a similar manner, Sha *et al.* used a variational modal decomposition method to enhance single-molecule fluorescence images, thereby facilitating the training of a model that enhanced the signal-to-noise ratio.[132] In their method, the enhanced data were first classified using a ResNet to judge whether they are signals or noise. Then the valid signal data were used to build a GAN model, which consist of a UNet-based generator and a CNN-based discriminator. Using this model to enhance the spectrum images, this approach resulted in a 100-fold reduction in the light intensity required for molecular detection.

3.5 NMR spectra

NMR spectra are closely related to the local environments of atoms, and their peaks can be used to determine the number of atoms in the same environment. This makes them useful for structural analysis. However, it has been challenging to determine the complete molecular structure directly using NMR spectroscopy due to the lack of overall structural information about the molecule. Additionally, reducing errors in the spectra

is a concern. The introduction of ML technology has facilitated the resolution of these problems to a certain extent.

ML approaches for analysing NMR have become increasingly prevalent in predicting the complex structure of materials and molecules. A notable example is the development of an ML-assisted analysis framework[133] by Engel *et al.* for NMR crystal structures, which enhances the reliability of identifying experimental structures even when they do not align with traditional confidence intervals. In this work, a total of 3546 and 604 configurations from crystal structures that only contain hydrogen, carbon, nitrogen, oxygen, and sulphur and no more than 200 atoms per unit cell are randomly selected as the training and testing set. Their Bayesian approach can leverage all available information from their dataset, including information that traditional measures may ignore, to quantify the confidence levels in structural identification. Bayesian methods are particularly effective for NMR crystal structure analysis because they inherently account for uncertainties in both experimental measurements and theoretical predictions, allowing for probabilistic reasoning that can incorporate prior knowledge about chemical shift distributions. The research also challenges existing benchmarks by revealing that the uncertainties in ^{13}C chemical shifts are underestimated. The study presents more accurate error estimates for chemical shifts predicted by gauge-including projector-augmented wave (GIPAW)[134,135] DFT calculations, which, when integrated into the analysis, improve the reliability of structure determination. In one case, the use of these corrected estimates resolves ambiguity in determining the structure. Furthermore, the study introduces a visual representation technique that projects the crystal structure landscape into a low-dimensional space to illustrate the similarity between candidate structures or their NMR shifts. This tool aids in diagnosing the reasons behind inconclusive determinations, such as lack of structural diversity or computational shift uncertainties. The integrated Bayesian framework and visual representations provide a systematic approach to identifying the most likely candidate structure matching the experiment, performing comprehensive sanity checks on the candidate pool and predicted NMR shifts, quantifying the confidence level in structural identifications, and analysing factors limiting confidence or resolving power when definitive identification is not feasible. This study can potentially lead to significant advancement in structural biology and materials science, where accurate NMR crystal structure determination is crucial.

Another NMR-based ML tool called "Small Molecule Accurate Recognition Technology" (SMART 2.0) was developed to produce constructed Heteronuclear Single Quantum Coherence (HSQC) spectra from data tables and to predict such spectra from published molecular structures.[136] For its training, SMART 2.0 used 25 434 HSQC spectra derived from natural products present in the JEOL database. Utilizing a CNN model named SqueezeNet,[137] the spectra were mapped into a 180-dimensional embedding space. An additional 27 642 spectra, computed using the ACD Laboratories predictor, were mapped to create HSQC spectra from a selection of mainly marine natural products, including those from NP Atlas[138] and NPASS databases,[139] into the 180-D space. Altogether, these 53 076 HSQC spectra of natural products together represent approximately 15% of all currently identified natural products. Subsequent experiments confirmed the robustness of SMART analyses under various NMR solvent conditions, reinforcing the usefulness of calculated HSQC spectra within the SMART framework. The unique cheminformatic tool was demonstrated to automatically characterize a complex natural product from a cyanobacterial extract mixture. This process culminated in the discovery of a new swinholide class of natural product, named "symplocolide A," and the identification of several known compounds within this structural class. The rapid prediction of structures for major constituents in crude extracts and fractions could significantly aid in prioritising structurally novel or intriguing natural products for further investigation.

In addition to determining the local environment of a substance, some researchers have tried to elucidate molecular structures directly from NMR spectroscopy. A notable study addressing the molecular structure inversion problem from NMR spectra[96,140,141] uses a combination of graph convolution networks and reinforcement learning.[140] In particular, this problem of determining the three-dimensional molecular structure of a compound from its ^{13}C NMR spectra and the molecular formula is formulated as a Markov decision process. To address this, Sridharan *et al.* employed a combination of online Monte Carlo tree search (MCTS) and offline trained graph convolutional neural networks (Fig. 7). This hybrid

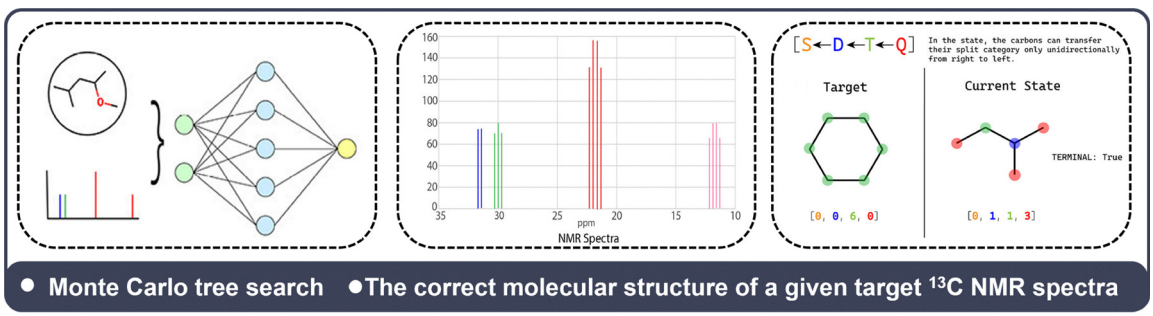


Fig. 7 Representative example of using the Monte Carlo tree search method to construct the structures of molecules with less than 10 non-hydrogen atoms from ^{13}C NMR spectra. Reproduced with permission from ref. 140, Copy © 2022, American Chemical Society.

approach is particularly powerful for NMR-based structural determination because graph convolutional networks inherently capture molecular connectivity and atomic environments, while reinforcement learning effectively explores the vast chemical space by learning from trial-and-error rather than requiring exhaustive enumeration of all possible structures. The framework represents the current molecular state as a graph, with atoms as nodes and potential bonds as actions. The MCTS algorithm explores the chemical space, guided by a policy network that predicts the likelihood of adding each possible bond and a value network that estimates the value of the current state. The forward NMR prediction model defines the reward function, encouraging the agent to explore structures that better match the target NMR spectra. The framework also incorporates chemistry-guided rules to prune the search tree and prevent the exploration of infeasible structures. In a study involving a dataset of more than 2000 organic molecules, each containing fewer than 10 non-hydrogen atoms, the approach demonstrated a 93.8% success rate in accurately identifying the target structure from the agent's predictions. Specifically, the correct structure was ranked as the top guess 57.2% of the time, and the accuracy in predicting the correct structure within the top three predictions was approximately 80%. It also surpasses brute-force database checking, efficiently identifying the correct molecule with minimal use of the forward model. Although the framework is promising, there is still room for improvement, particularly in developing a better scoring function to increase prediction accuracy. The work suggests that integrating additional spectra, such as IR and ^{1}H NMR spectra, could further enhance the framework's ability to determine the structure of unknown molecules, which could significantly bolster drug discovery by enabling rapid and reliable structural verification.

It is worth mentioning that the emergence of generative models provides new ideas for NMR-based *ab initio* generation of molecular structures. Sun *et al.* proposed a cross-modal retrieval method based on a library of focused molecules to realize this process.[142] A bidirectional autoregressive transformer model is first used to generate a library of focused molecules from the chemical shifts of ^{13}C. The library is then used to train an RNN network, generating an expanded set of molecules, and finally the generated molecules are validated against NMR in the latent space to obtain a series of the most probable molecular structures and their scores. The model achieves a top-10 accuracy of 54% in predicting molecular structures on a test set containing about 6500 molecules.

ML techniques have also been applied to 2D NMR spectra to assist in spectral resolution. Compared with one-dimensional spectra, 2D spectra tend to be more informative, but require more time for accurate sampling. This makes the signals in 2D spectra less accurate and more difficult to analyse. However, by utilising ML's ability to resolve intrinsic correlations from complex data, researchers have successfully enhanced 2D NMR spectra and resolved signals. For instance, a recent work combines the matrix completion (MC) algorithm and deep learning (DL) to accelerate the reconstruction of 2D nitrogen-vacancy (NV) spectrum maps.[143] The DL network is trained to learn the complex non-linear mapping from a partially filled spectrum map to its full-resolution map. The traditional MC method is then employed to post-process the DL output map to maintain its low-rank property, thereby alleviating the domain shift problem. This DLMC method has been reported to recover the missing entries from a partially sampled 2D NV spectrum map, significantly enhancing the experimental efficiency. Compared to the standalone MC method and the basic DL method, the proposed DLMC approach can reconstruct the experimental data with a very low sampling ratio and without domain shift. The reconstruction of an efficient 2D NV spectrum map could yield valuable structural information, which could potentially be used to construct the three-dimensional structure of the molecule. In another approach, a system based on deep neural networks has been introduced to interpret chemical shift anisotropy (CSA) in multidimensional separated local field (SLF) solid-state nuclear magnetic resonance (SSNMR) spectra.[144] The method's effectiveness was demonstrated through testing on four samples comprising two synthetic polymers (polyethylene and polyethylene oxide) and two amino acids (acetylvaline and histidine). The SLF SSNMR spectra of the polymers were acquired experimentally, while the spectra of the amino acids were adopted directly from published literature. Spectra based on both ^{13}C and ^{15}N were examined. Training datasets were constructed by simulating all four samples using SIMPSON software. The developed DNN models were then applied to the experimental SLF SSNMR spectra, and the resulting CSA tensor orientations were compared directly with the values reported in the literature. The authors found that all Euler angles of the CSA tensor orientations showed deviations within 5 degrees, demonstrating the effectiveness of the DNN-based approach in accurately interpreting the multidimensional SSNMR spectra. This work represents a significant advancement in the application of ML techniques to assist the interpretation of complex SSNMR data, which provide valuable structural and dynamic information for a wide range of materials, including synthetic and natural polymers.

In conclusion, AI has advanced NMR spectroscopy techniques by enabling more accurate and efficient molecular structure determination. One key innovation is the development of ML frameworks that refine error estimates and integrate more data, improving the reliability of NMR-based crystal structure identification. Tools like SMART 2.0 utilize neural networks to predict HSQC spectra, accelerating the discovery of new natural products by analysing complex mixtures. Additionally, ML models have been applied to interpret multidimensional NMR spectra and reconstruct high-resolution 2D spectrum maps. ML could also improve the spectral signal[145,146] by enhancing the signal-to-noise ratio, assisting the NMR spectroscopy deconvolution process, and even designing ^{1}H and ^{15}N radio frequency pulses for rapid data acquisition. These advancements streamline the NMR analysis process, reducing the need for manual interpretation and enabling faster, more accurate structure identification. This continues the broader theme of using AI to transform various spectroscopic methods

by automating complex tasks and extracting deeper insights from spectral data.

3.6 MS

Mass spectrometry provides information on the charge-to-mass ratio of molecules and molecular fragments. This ratio is closely related to the pathways and probabilities of molecular dissociation and can therefore distinguish structural differences between molecules. Traditional mass spectrometry analysis relies on comparisons with data from standard databases, as well as the calculation of molecular dissociation pathways and mass spectral prediction. However, the former can only be applied to known molecules, while the latter grows exponentially in terms of computation time with the number of molecular dissociation steps and requires a large amount of computation time. The introduction of machine learning methods has drastically reduced the computational requirements for molecular dissociation, thereby improving the efficiency of mass spectral resolution.[72,147,148] An early example of this is the CSI:FingerID method,[72] which uses a molecular fingerprint to describe the molecular structure. This fingerprint can be obtained directly from the molecular structure using a computational process. The method also calculates the molecule's fragment tree based on the mass spectrum and uses machine learning to predict the molecular fingerprint from the mass spectrum and fragment tree. For mass spectra of unknown substances, the molecular fingerprints predicted by the model can be used to identify the most probable molecules by performing similarity calculations with those in the molecular structure library. In another typical work,[147] Ludwig *et al.* introduced a new framework that departs from previous assumptions of molecular property independence, integrating ML predictions as marginal probabilities within a probabilistic framework. The resulting scoring system is designed to be more accurate and comprehensive, considering both deterministic fingerprint dependencies and prediction-related interdependencies. The authors demonstrated that this new scoring method outperforms previous approaches, suggesting a substantial enhancement in the efficiency and reliability of metabolomics analysis. Given the widespread use of CSI: FingerID,[72] this advancement could profoundly impact the scientific community's ability to analyse complex biological samples, thereby contributing to a deeper understanding of biochemical activities and the discovery of new biomarkers or therapeutic targets.

Compound identification *via* MS usually requires a large library of experimentally collected MS/MS spectra for comparison. However, the number of available MS/MS spectra in existing databases is very limited. To address these limitations, Ji *et al.* developed an ML model named DeepMASS.[149] This model first scores the structural similarities between unknown metabolites and those in databases based on their MS/MS spectra. DeepMASS then uses these structural similarity scores to obtain reference metabolites and retrieve structural candidates from public compound databases, ranking them with the assistance of reference metabolites. This integrated approach aims to overcome the challenges faced by previous structural similarity-based methods, which often fail to fully capitalise on the wealth of MS/MS spectra information to identify unknown metabolites. The *de novo* structural construction from standard mass spectrometry databases has also been an important part of researchers' endeavours. The Cheng Lin group has developed a series of algorithms called GlycoDeNovo for resolving glycan structures through tandem mass spectrometry.[150–152] These algorithms do not require a database of glycans; instead, they construct possible glycan molecules from scratch.[150] Since the ions produced by the cleavage of each glycan are theoretically complementary, but not all theoretically present fragments are necessarily observed in experiments, the algorithm first generates complementary peaks for the observed peaks to help reconstruct the topology. Subsequently, an interpretation graph is constructed, where nodes indicate mass spectral peaks and edges indicate how combinations of peaks are interpreted as B or C ions. Based on this graph, the proposed algorithm recursively reconstructs all possible topologies, employing a ML classifier to validate the ion types and score each candidate topology based on the validation results. Furthermore, the efficiency of deconvolution is improved by introducing parallelisation into the algorithm and comparing the spectra of candidates with theoretical spectra.[151]

Inspired by natural language processing algorithms that capture element relatedness for overall similarity assessment of objects, Huber *et al.* also explored calculating spectral similarity scores based on learned embeddings of spectra.[153] In the model called Spec2Vec, a spectrum can be represented by a low-dimensional vector within a continuous abstract space, where vectors representing highly related fragments point in similar directions. The word-embedding approach of Spec2Vec is particularly suitable for MS spectral comparison because it captures the semantic relationships between fragment peaks rather than just their presence or absence, similar to how word embeddings in NLP capture relationships between words in context. Moreover, Spec2Vec takes the relationship between fragments into account rather than relying solely on a binary assessment of each fragment (match/no match). Spec2Vec is also unsupervised and can be trained on any collection of spectra. Dührkop *et al.* also employed deep neural networks to identify unknown compounds,[154] for which neither spectral nor structural reference data are available, as well as to predict classes lacking tandem mass spectrometry training data, going beyond the limitations of traditional approaches that rely on reference databases.

Facing the lack of experimental MS/MS spectra, some researchers turned towards *in silico* methods for MS-based compound identification, including MS/MS spectra prediction and MS/MS spectral fingerprint analysis, often combined with ML methods. An example of these approaches is CFM-ID (Competitive Fragment Modelling Identification, Fig. 8).[155,156] CFM-ID's probabilistic model is well-suited for MS/MS spectra prediction because it can learn the statistical patterns of molecular fragmentation directly from data, capturing both common and rare fragmentation pathways without requiring

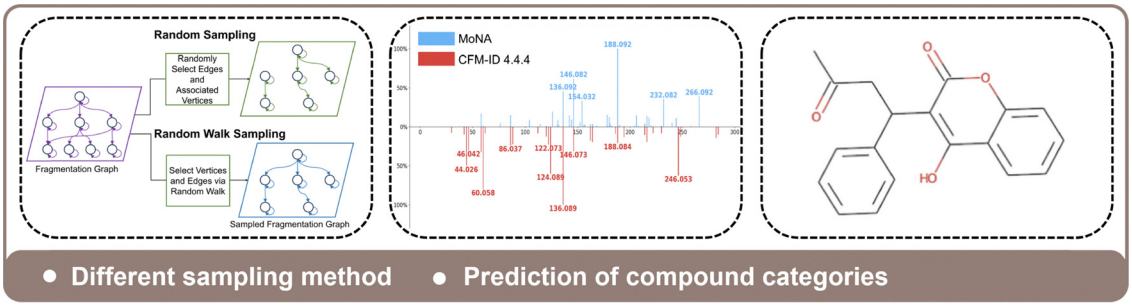


Fig. 8 A representative example of using random sampling and random walk sampling methods to predict the structures from MS spectra. Reproduced with permission from ref. 155 and 156.

explicit encoding of all possible fragmentation rules. Trained on a limited set of experimental MS/MS spectra alongside their corresponding molecular structures, CFM-ID learns the fragmentation behaviour of small molecules when introduced into a quadrupole time-of-flight instrument operating under electron impact or electrospray ionisation with collision-induced dissociation. This learning process enables CFM-ID to generate predicted ESI-MS/MS spectra from a given chemical structure and provide probable fragment ion annotations for every peak in the predicted spectrum. Furthermore, applying CFM-ID to all known small molecule structures, including those with predicted configurations, enables the compilation of an *in silico* MS/MS spectral library that is several times larger than any collection obtained through experimentation. This computational MS/MS spectral library can then be used to identify compounds by matching them with MS/MS spectra measured experimentally against the library.

Despite the fact that the aforementioned methods no longer necessitate large-scale libraries of standard mass spectrometry, the search remains constrained to libraries of known molecular structures. The advent of generative modelling in recent years has enabled mass spectral elucidation to diverge from reliance on existing databases of molecules, thereby facilitating the *de novo* generation of molecular structures. This development holds great promise in terms of enhancing the efficiency of the identification of unknown substances. For instance, Orlova *et al.* developed an automated reaction network generation model, which, when used in conjunction with MS spectroscopy, enabled the determination of the automated oxidation products of oils.[157] Skinnider *et al.* developed an automated method for the analysis of the chemical structures of unknown synthetic drugs using generative models for the detection of unknown synthetic drugs,[158] and were able to analyse the exact chemical structures of the unknown synthetic drugs with an accuracy of 51% and a top-10 accuracy of 86%. It is anticipated that this methodology will be employed in the analysis of other types of small molecules for *de novo* structure generation. Litsa *et al.*, trained a SMILES-to-SMILES auto-encoder network with GRU as both the encoder and decoder, which in turn trained a CNN encoder to embed MS into the hidden space of SMILES, and finally co-located the CNN encoder and GRU decoder for MS-to-structure generation.[159] Stravs *et al.* used molecular fingerprints as an intermediate representation to predict the molecular fingerprint and chemical formula using MS, and then trained the model to generate the SMILES structure from the molecular fingerprint and chemical formula.[160] Bohde *et al.* used a diffusion model to generate molecular structures from scratch, with a pre-training approach that significantly improved accuracy.[161] The distance between the best prediction and the ground truth is only a half of those in previous studies.[159,160]

Unlike its direct use in molecule reconstruction *via* MS, ML has also proven valuable in the signal analysis process. In time-of-flight mass spectrometry (ToF-MS), detected signals originate from not only atomic ions but also from molecular patterns. Wei *et al.* introduced an ML-driven method called 'ML-ToF',[94] which automatically assigns elemental and molecular identities to individual and grouped peaks in ToF-MS spectra. Furthermore, this approach provides uncertainty estimates for these assignments, reflecting the influence of noise levels and morphological features on the observed peak patterns. It is shown that ML-ToF can handle various ToF-MS spectra without prior knowledge of the sample composition. ML-ToF significantly accelerates the peak recognition process, taking only microseconds to produce a labelled spectrum, while human users may require minutes or even hours to achieve the same result.

In short, AI has improved scoring systems for compound identification in MS, integrating molecular property interdependencies for more accurate metabolomics analysis, which has profound implications for biomarker discovery and therapeutic targeting. Additionally, tools like ML-ToF automate peak assignment in ToF-MS spectra, drastically reducing the time required for analysis. Other ML approaches, such as DeepMASS and Spec2Vec, address the scarcity of reference spectra in databases by leveraging structural similarities and embedding techniques to identify unknown metabolites. *In silico* methods, like CFM-ID, predict MS/MS spectra and create large synthetic libraries, enabling more reliable identification of compounds when experimental data are limited.

3.7 X-ray spectra

The substantial amount of standard spectroscopic data, along with the good agreement between theory and experiment,

makes XRD highly suitable for AI applications. ANNs have been used to classify the crystal systems and space groups of inorganic powder specimens based on their XRD patterns. Vecsei *et al.* trained a deep dense neural network on the dataset of over 10^5 theoretically computed powder XRD patterns from various inorganic crystal structure databases,[92] including the Inorganic Crystal Structure Database (ICSD).[84] In particular, the researchers incorporated artificial noise and background signals into theoretical data to mimic experimental conditions. The resulting neural network model was then tested on real experimental data, where its classification accuracy was 54%, which increased to 82% when the network was allowed to refuse the classification of patterns with high uncertainty, leaving half of the data unclassified. A similar research[93] showed that the performance of the ML model on a small experimental dataset can be significantly improved *via* data enhancement. A library of 164 XRD patterns was compiled from the ICSD and then augmented using a physics-informed data augmentation strategy, introducing variations such as peak scaling, elimination, and shifting to mimic real-world experimental conditions. This augmentation generated a more extensive and diverse dataset that retained the original labels of crystal dimensionality and space group, enabling the ML models to train on a broad range of potential experimental outcomes. Among the models tested by the authors, the most successful model, denoted as a-CNN, demonstrated an accuracy of 93% for dimensionality and 89% for space groups in classifying these structural features from the XRD patterns. These approaches significantly enhanced the capability of the ML model to generalise from limited data, thereby improving its utility in material characterisation and novel material discovery. Training CNN models on large datasets, such as nearly one million crystal structures from databases, including the ICSD and the Cambridge Structural Database, can significantly improve their performance.[162] The models are capable of estimating lattice parameters for various crystal systems with a significant reduction in the search space for these parameters, approximately 100- to 1000-fold. This work also showed that challenges remain for ML models in realistic experimental scenarios, such as identifying impurity phases, baseline noise, and peak broadening. Addressing these challenges may involve developing more advanced models. In some cases, such as the Rietveld refinement process,[163] ML has been shown to conduct structure refinement without human intervention.

ML can also be used for efficient phase identification in multiphase inorganic compounds based on XRD patterns.[87,88,97] In a study on the inorganic compounds within the Sr–Li–Al–O quaternary compositional pool,[97] the powder XRD patterns for 170 compounds are simulated and then used to generate an extensive dataset of 1 785 405 synthetic XRD patterns through combinatorically mixing. After being trained on this dataset, CNN models achieved nearly 100% accuracy in phase identification and 86% accuracy in three-step phase-fraction quantification when tested with real experimental data. Despite the good performance of ML models, this work avoided textured data, which are imperative for thin-film diffraction. Later works[164] developed an ML tool called crystallography companion agent (XCA). This tool trained an ensemble of 50 CNNs, which were designed to produce probabilistic classifications rather than absolute ones, addressing the uncertainties inherent in XRD data. The XCA system integrates multimodal data analysis, combining XRD with other techniques like energy-dispersive X-ray spectroscopy (EDX) to enhance analytical accuracy. It is reported that XCA is more accurate in the limiting case of textured phases producing degenerate patterns. A recent work[89] by Szymanski *et al.* bridges model prediction and experimental measurements, enabling automated phase identification and experimental iterative optimisation. Initially, a rapid preliminary XRD scan is performed across a specified 2θ range to gather initial diffraction data, which is then analysed using a deep learning model. The model predicts the likely phases and evaluates its confidence in these predictions. If the confidence level is below a predefined threshold, the algorithm directs the XRD system to either rescan certain 2θ regions at a higher resolution or expand the scan range to capture additional peaks that could confirm or refute the suspected phases. This process utilises class activation maps (CAMs) to pinpoint which features in the diffraction pattern have the greatest influence on the predictions, guiding focused rescans to enhance data quality and prediction certainty. CAMs provide critical interpretability for XRD analysis by highlighting the specific diffraction peaks that contribute most to phase identification, enabling targeted experimental optimization and building trust in the ML model's decision-making process. The cycle of scanning, analysis, and rescanning continues iteratively until the confidence levels improve sufficiently or all analytical avenues are exhausted, thereby optimising the phase identification process for accuracy and efficiency.

By combining theoretical augmentation with robust models, advanced ML techniques are transforming X-ray-based spectroscopic analysis, enabling the rapid identification of both known and unknown material structures. Augmenting the theoretical data by extracting noise from experimental spectra can also empower ML models in structure type identification tasks, such as the types of MOFs. A recent work augmented 1012 theoretical MOF patterns to 72 864 samples and used them to train a CNN model.[90] The optimised model achieved a 96.7% identification accuracy within the top 5 rankings on a test set of 30 hold-out samples. Further analysis using CAMs from the last CNN layer revealed that the model identified MOFs by focusing on the main peaks of the XRD patterns. This interpretable approach not only demonstrated high accuracy but also provided insights into the CNN model's decision-making process. The study's findings highlight the potential of this CNN model for rapid and accurate identification of XRD patterns, with possible extensions to other characterisation techniques such as Raman, NMR, and FTIR spectroscopy. Aside from identifying the types of known materials, a more challenging task is to find unknown compounds based on XRD. This has been done by a model named crystal structure-type identification network (CrySTINet),[91] which can automatically identify

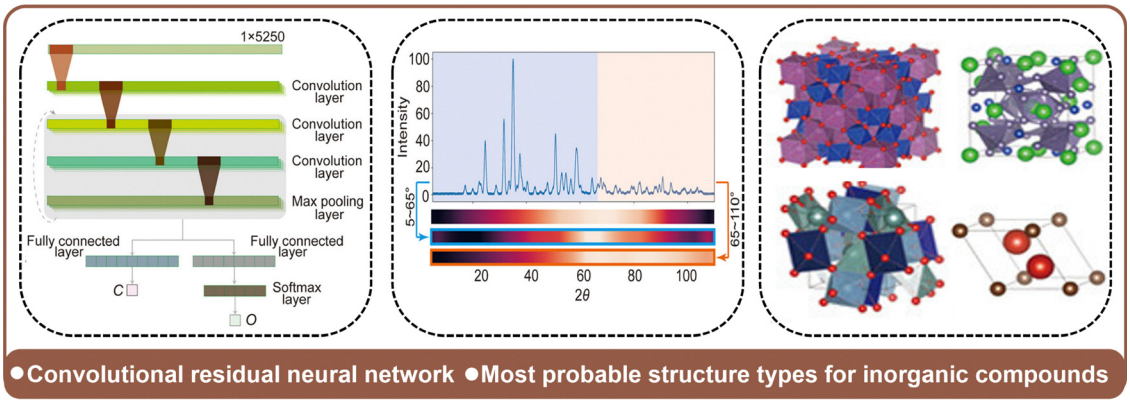


Fig. 9 A representative example of developing a CNN-based model to predict the most probable structure types for inorganic compounds *via* XRD spectra. Reproduced with permission from ref. 91, Copy © 2024, American Chemical Society.

and classify the structural types of unknown compounds based on their XRD patterns (Fig. 9). The applicability and potential of this model are demonstrated on a dataset of 63 963 compounds extracted from ICSD, with all these compounds belonging to the top 100 most popular structure types in the ICSD. CrySTINet reaches a promising accuracy of 80.0% without requiring any prior information about the material compositions. Additionally, the model can be expanded to accommodate various structure types based on specific requirements. Moreover, the application of transferable generative architectures further expands the capacity of ML models in spectroscopic analysis. Chen *et al.* used the transferable vision transformer model to achieve identification of materials from their spectra, including XRD and FTIR.[165] The method first predicts the type of MOF based on XRD, and the model achieves a top-5 prediction accuracy of 95%. The authors then used transfer learning (TL) on this model to a seemingly completely unrelated task: predicting the functional group classification of small molecules based on FTIR. The TL model performed even better than the model trained directly on FTIR data, with a top-5 accuracy of 96.7%. This demonstrates the surprising learning ability and transferability of the Transformer-based models for spectroscopically relevant tasks.

ML approaches have been commonly used for analysing the X-ray absorption near-edge structure (XANES) spectra, which are sensitive to the structure, especially the coordination environment of elements. In a typical study[166] from Timoshenko *et al.*, the coordination numbers (CNs) of nanoparticles have been successfully determined using XANES spectroscopy combined with supervised ML. The study constructed a training set with theoretical XANES spectra for platinum (Pt) nanoparticles of different sizes and shapes, where the sets of corresponding average CNs are known. ANNs were used to refine the three-dimensional geometry of these nanoparticles by correlating spectral features with structural descriptors. The trained ANN can quickly analyse experimental XANES data to determine the CNs of nanoparticles in a sample. It can also effectively reconstruct the average size, shape, and morphology of Pt nanoparticles, aligning with prior experimental results. Using XANES spectra, various studies attempt to predict the CN, local symmetry, and other local properties of materials.[167–173] Trained on a database of over 50 000 simulated XANES spectra from eight first-row transition metal oxides, an adversarial autoencoder model augmented with a rank constraint (RankAAE) is reported to decouple intertwined spectral contributions from multiple structural characteristics.[167] RankAAE creates a continuous and interpretable latent space where each dimension corresponds to an individual structure descriptor, including oxidation and coordination. This approach enables the identification of distinct spectral trends associated with changes in material structures, thus aiding in the discovery of new insights from complex datasets. A recent study[174] by Carbone *et al.* suggests that ''multi-modal'' models have higher potential than traditional ML approaches in predicting molecules from XANES. Combining C, N, and O K-edge XANES of different elements, the prediction accuracy of ML models on small molecules can be significantly improved using single-element K-edge data alone. By incorporating uncertainty quantification within the classifiers, researchers can assess the confidence of predictions, thereby enhancing the model's reliability and applicability.

ML techniques have also been extended to the interpretation of XAS and XPS in various contexts. For materials in amorphous and disordered forms, the spectroscopic signatures of different local atomic environments overlap, making them challenging to resolve using traditional reference data from molecular or crystalline samples. Focusing on carbonaceous materials, ML approaches have been used to resolve the complexities of interpreting XAS and XPS spectra in a series of studies.[175,176] A dataset of 50 carbon structures was used, including bulk and surface samples, and functionalised with hydrogen, oxygen, hydroxyl, and carboxylic acid groups. These structures were clustered according to the atomic environments. Clustering-based ML approaches excel at XAS and XPS analysis of disordered materials because they can identify similar spectral fingerprints across different samples, allowing the model to reduce the dimensionality of the problem while preserving the essential chemical information. The average characteristic

spectrum (XPS and XAS) of each cluster is obtained as the "fingerprint" spectra, representing different chemical environments. These fingerprint spectra were then used to fit experimental XAS and XPS spectra, followed by a Monte Carlo approach to optimise the fitting parameters. The results indicated that this intelligent model successfully deconvoluted overlapping spectral features and yielded quantitative estimates of the presence of various functional groups. The study highlighted that XAS provides more reliable insights into disordered materials' atomic-level structure than XPS. Still, both techniques should be used together for a comprehensive analysis.

In conclusion, the integration of ML with X-ray spectroscopy techniques significantly enhances the accuracy and efficiency of structural analysis. In XRD, ML models such as DNNs and CNNs have been employed to classify crystal systems, space groups, and lattice parameters. These models improve upon traditional methods through data augmentation and realistic experimental noise simulation. ML-based approaches enabled rapid phase identification in multiphase compounds, with tools like XCA combining XRD with other techniques for more accurate analysis. Beyond XRD, ML models have also been applied to XANES spectroscopy to predict coordination numbers and oxidation states, offering new insights into the local structural information of materials. These models have also been extended to interpret complex XPS and XAS spectra, particularly in disordered materials.

4. From structures to spectra

As outlined in Section 2, AI can make end-to-end predictions directly from spectra to chemical structures, thereby significantly simplifying the process of spectral interpretation. However, training such ML models requires substantial data on spectra and their precise structures. In scenarios where the available data is limited, it is still necessary to rely on traditional means for spectral interpretation. Given that the most time-consuming aspect of traditional methods is the quantum mechanical calculation of theoretical spectra, ML algorithms have been introduced to address this challenge. By training models on datasets containing experimental or theoretical structure-spectra pairs, ML algorithms can autonomously extract critical information from molecular structures and predict their spectra within minutes or even seconds, effectively overcoming the efficiency bottleneck of traditional computational methods (Fig. 10).

The primary task of using ML models to predict spectra from molecular structures involves simplifying high-dimensional molecular information into lower-dimensional spectral data while accurately rendering the corresponding molecular spectra. This approach has found broad applications across various research fields and spectral techniques. For instance, ML models have enabled more accurate simulations of IR and Raman spectra with reduced computational costs across various molecular structures, including complex structures like proteins. For UV spectra, ML models can correct systematic errors inherent in low-level quantum chemistry methods and make direct predictions of excited-state energies and transition dipole moments, thereby enhancing the precision and efficiency of spectral calculations. Additionally, MS spectra prediction can be enhanced by applying ML techniques, whereby the molecular topology, fragmentation, and substructure are modelled. Consequently, ML models improve compound identification by reducing the reliance on extensive experimental data. The following sections will provide a detailed overview of these applications, categorised by different types of spectroscopy techniques, and discuss the specific contributions of ML in each research area.

4.1 Vibrational spectra

The pioneering work on predicting vibrational spectra dates back to 2017, focusing on integrating ML techniques with *ab initio* molecular dynamics (AIMD) simulations to predict IR spectra efficiently.[180] Gastegger *et al.* utilised high-dimensional neural network potentials (HDNNPs) to model potential energy surfaces (PES) and employed element-decoupled Kalman filters for precise force calculations. They optimised data point selection through an adaptive sampling scheme and managed large molecular systems using a fragmentation method. A key innovation in their work was the novel molecular dipole moment model based on environment-dependent neural network (NN) charges. This approach significantly reduced the number of required electronic structure calculations. With these approaches, they successfully simulated the IR spectra of methanol, *n*-alkanes containing up to 200 atoms, and protonated alanine tripeptide. Subsequently, numerous studies[177,181–183] have expanded the application of ML techniques to predict vibrational spectra in various systems. Beckmann *et al.* developed an NN model for small systems such as protonated water clusters to accurately predict the IR spectra.[182] This model encompasses both the potential energy surface (NN-PES) and the dipole moment surface (NN-DMS) and is trained on a dataset of over 54 710 configurations, reaching coupled cluster (CC) accuracy for reference dipole moments. By simulating IR spectra through MD that account for anharmonicities and finite-temperature effects, this method drastically improves the simulation accuracy compared to traditional techniques by offering a computationally efficient solution that captures complex MD. Similarly, NN models for predicting PES and dipole moments have also been constructed for the PAH system. Using these trained models with generalized second-order vibrational perturbation theory (GVPT2), IR properties for 34 PAH molecules were predicted.

Moving on to more complex systems, making accurate predictions becomes increasingly challenging. This is due to the increasing degrees of freedom, molecular fluxionality, and intramolecular interactions. Moreover, the computational cost for these systems rises significantly, complicating efforts to achieve high accuracy within a reasonable timeframe. To address these challenges, a computational approach has been developed to simulate the Raman spectra of MXene by

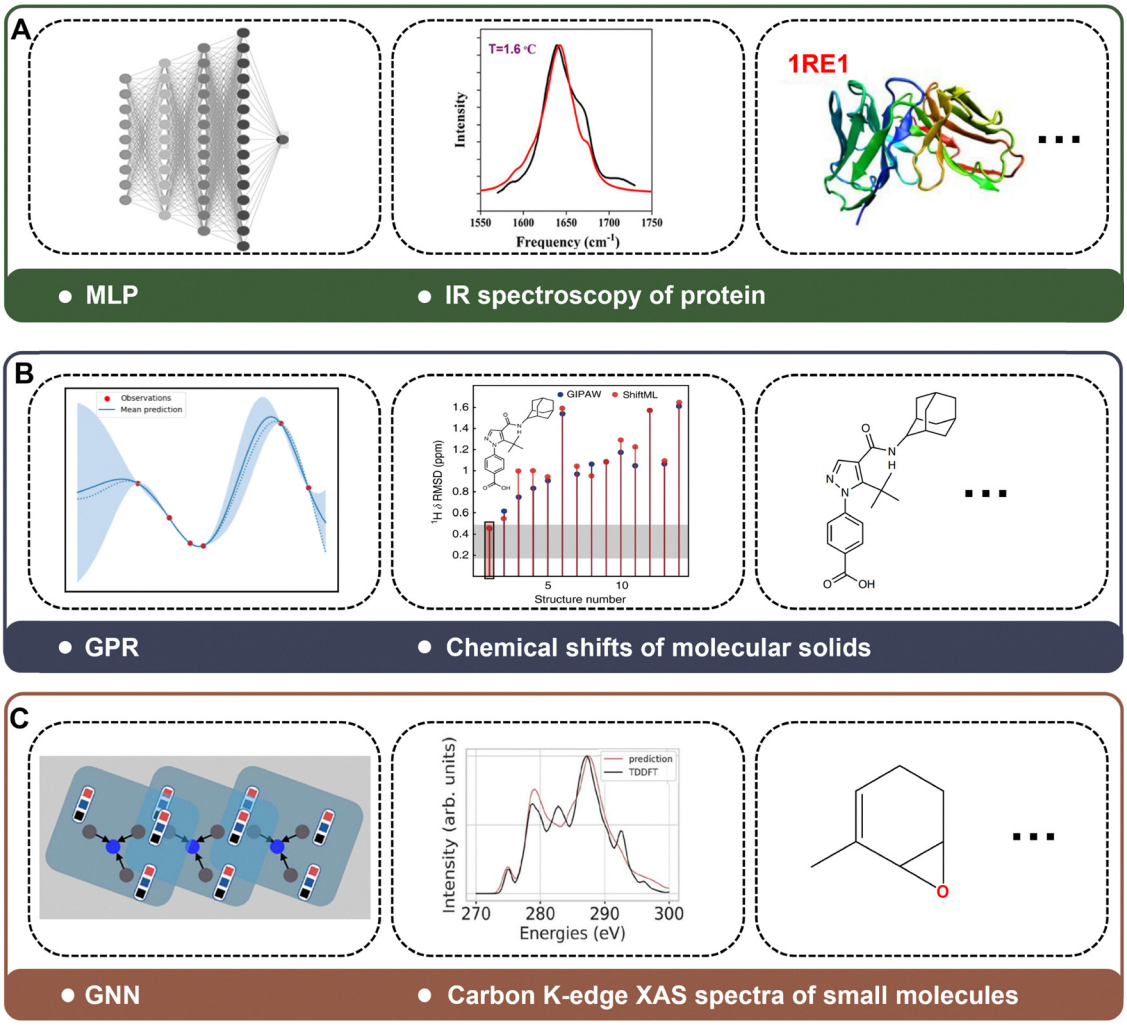


Fig. 10 Representative examples of using ML models to predict spectra from the structure. (A) Using MLP to predict the amide I IR spectra of protein. Reproduced with permission from ref. 177, Copy © 2020, American Chemical Society. (B) Using GPR to predict the chemical shifts of molecular solids containing HCNO. Reproduced with permission from ref. 178. (C) Using GNN to predict the carbon K-edge XAS spectra of small organic molecules with less than 10 non-hydrogen atoms. Reproduced with permission from ref. 179.

combining ML force-field molecular dynamics (MLFF-MD) with a method to reconstruct Raman tensors *via* projection to pristine system modes.[184] This approach accounts for finite temperature effects, mixed surfaces, and structural disorder. When applied to titanium carbide MXene, this technique effectively captures the temperature effect and mixed surface terminations on Raman spectra, leading to better agreement with experimental results. Additionally, NN-based potentials have been developed to simulate the IR spectra of nanosilicate clusters.[185] Specifically, specific NN-based potentials were built to accurately represent the total energy, atomic forces, and dipole moments. An active learning strategy that iteratively improves the model through MD sampling was employed to effectively generate training data. Active learning strategies enable the intelligent sampling of the most informative configurations from the vast conformational space, focusing computational resources on regions that maximize model improvement and ensure robust performance across diverse molecular arrangements. The final ML model accurately predicts the IR spectra by including anharmonic effects for nanosilicate clusters of varying sizes and demonstrates good transferability to high-energy isomers not included in the training data. The ML-assisted IR spectral simulation was also extended to liquid water. Gaussian process regression and ANNs were combined to predict OH-stretch frequencies and transition dipoles of water, utilizing a dataset generated from MD simulations with atom-centred symmetry functions as descriptors.[186] Later, an E(3)-equivariant neural network (e3nn) was developed to fit the atomic polar tensor (APT), enabling accurate IR spectral calculations from MD simulations without requiring extensive AIMD.[187] E(3)-equivariant neural networks are uniquely suited for modelling molecular spectroscopy because they inherently respect the rotational and translational symmetries of physical systems, ensuring that predicted properties, such as dipole moments, transform correctly under rotation and translation operations.

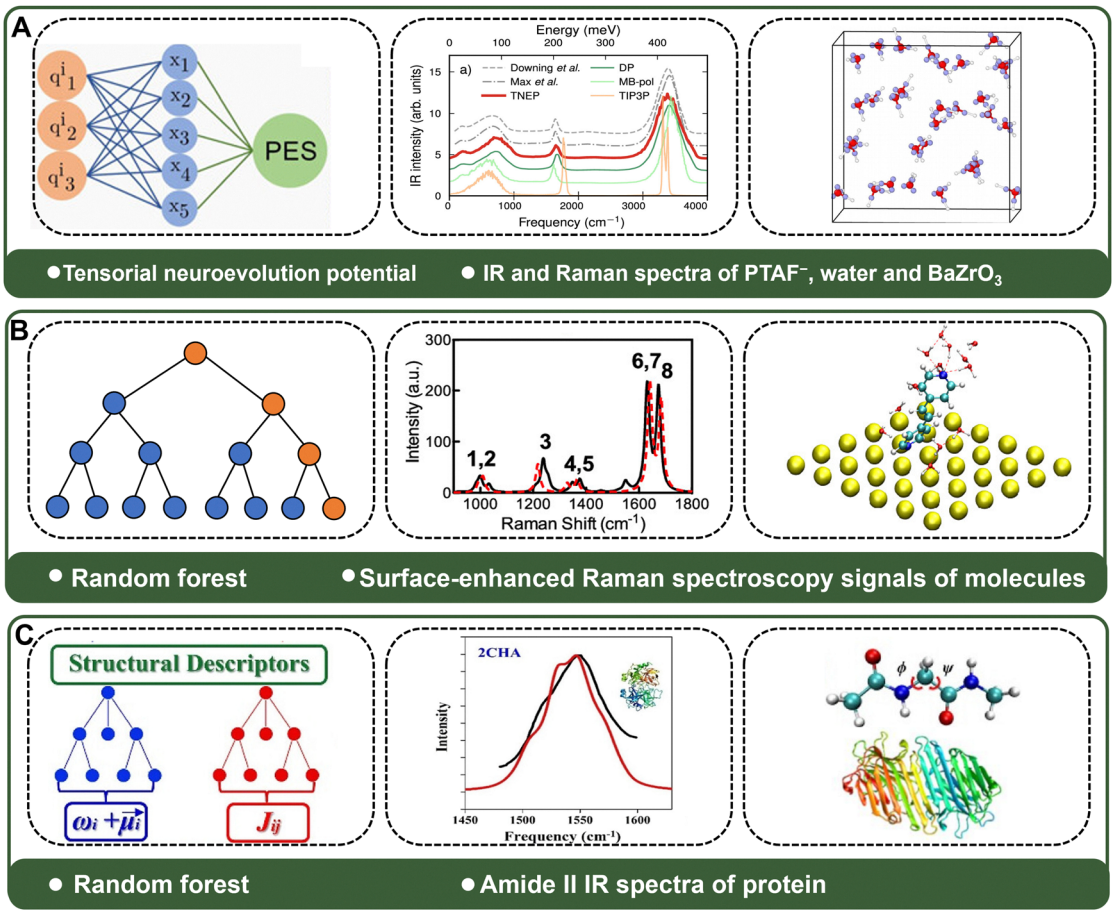


Fig. 11 Representative examples of using ML models to predict vibrational spectra of different types of molecules. (A) Using tensorial neuroevolution potential to predict IR and Raman spectra of a small molecule ($PTAF^-$), a liquid (water), and a solid ($BaZrO_3$). Reproduced with permission from ref. 188. (B) Using RF to predict surface-enhanced Raman spectroscopy of a *trans*-1,2-bis (4-pyridyl) ethylene molecule adsorbed on different metal surfaces. Reproduced with permission from ref. 191, Copy © 2019, American Chemical Society. (C) Using RF algorithm for the amide II IR spectra simulation of different proteins and folding Trp-cage protein. Reproduced with permission from ref. 192, Copy © 2024, American Chemical Society.

In addition to predicting the vibrational spectra of individual systems, some studies aim to predict the vibrational spectra of systems across different scales. A scheme named Tensorial Neuroevolution Potential (TNEP) has been developed for the fast simulation of IR and Raman spectra of molecules at various scales (Fig. 11A).[188] TNEP can simulate tensorial properties such as dipole moment (μ), polarizability (α), and electric susceptibility (χ) directly from the structural data of molecules, liquids, and solids, bypassing the computational overhead of traditional methods. Successfully applied to $PTAF^-$ (a molecule), water (a liquid), and $BaZrO_3$ (a solid), this scheme calculates total energy as the sum of atomic energies derived from local atomic environments. ML models are developed to describe rank-2 virial tensors such as polarizabilities and electric susceptibility. The TNEP approach offers valuable insights into the correlation between molecular properties and the electric susceptibility of extended systems. Additionally, the transferability of these models was examined, focusing on ML-based polarizability models for predicting Raman spectra of large molecular systems, particularly alkanes.[189] Two Neuroevolution Potential (NEP) models were developed using data from alkanes with up to five and nine carbon atoms, respectively. These models accurately predicted polarizabilities and Raman spectra of *n*-undecane, an alkane with eleven carbon atoms. A descriptor space analysis was also introduced to evaluate the overlap between training data and target molecules, enhancing the understanding of model transferability. Extending these multi-scale approaches, recent efforts leveraging generative models have focused on capturing the additional complexity introduced by diverse phase states in complex systems. Extending these multi-scale approaches, recent efforts have successfully used generative models to capture the additional complexity introduced by diverse phase states in complex systems. Na represented the molecular structure as a molecular graph and integrated the phase state of the molecule into a graph neural network.[190] The ensuing results were then transmitted to the transformer's decoder, resulting in the successful generation of infrared spectra for the molecule in various states of matter. The correlation coefficient between the generated spectra and the experimental spectra is close to 0.9, and the error is smaller than that of the time-consuming theoretical calculations. This work introduces the first infrared

spectroscopy generation model capable of producing phase-dependent spectra for structurally complex real molecules and demonstrates the significant potential of generative models in spectra predictions.

In addition to using ML to accelerate IR and Raman simulations by learning vectorial properties, another approach is directly predicting spectral information, such as vibrational frequencies and IR/Raman intensities. For instance, a NN[98] model was developed to efficiently predict stretching frequencies and IR/Raman intensities for hydroxyl (–OH) and carbonyl (C═O) functional groups in over 21k molecules taken from QM9.[193] This ML prediction method is nearly 1000 times faster than traditional first-principles calculations and is capable of mimicking human-like structure recognition. Other than IR, ML methods can also be used to predict SERS. The direct prediction ML model was first applied to predict the SERS signals of molecules adsorbed on gold substrates (Fig. 11B).[191] By employing an RF algorithm using the internal coordinates of the propylene molecules and their relative positions to the gold substrate as feature descriptors, this study achieved prediction accuracies comparable to DFT, while the computation time was only one ten-thousandth. Additionally, this model exhibited excellent transferability under various conditions, including different solvent environments, electric fields, and metal surfaces.

Macromolecules like proteins have more complex conformations than small molecules, resulting in intricate interaction patterns within their spectra. This complexity makes direct spectra prediction challenging. To address this, Ye *et al.* developed machine learning models by representing the protein vibrations as a set of oscillators associated with each peptide bond in its backbone.[194] These models predict the interactions between adjacent amino acids and share them in the exciton Hamiltonian, efficiently constructing the exciton Hamiltonian and diagonalising it to obtain the protein spectra. This approach also allowed accurate prediction of protein IR,[177,192,195] UV,[194] and CD spectra.[196] Notably, by utilising the high sensitivity of IR spectra to structural changes, they analysed and identified a close relationship between protein secondary structures and their IR spectral characteristics (Fig. 10A). For instance, during the binding process of the SARS-CoV-2 spike protein with the hACE2 protein, changes in protein secondary structure, such as increased α-helix and β-sheet content and decreased random coil elements, results in a blue shift of the main peak in the IR spectrum. In contrast, the opposite changes causes a red shift. The ML approach employs a divide-and-conquer strategy to predict amide II vibrational properties, such as frequencies, transition dipole moments, and coupling parameters of adjacent oscillators, by fragmenting proteins into individual peptide bonds and dipeptide units (Fig. 11C). This method significantly reduces computational costs while maintaining accuracy comparable to DFT calculations. By using the *N*-methylacetamide (NMA) molecule as a model for peptides and *N*-acetyl-glycine-*N*′-methylamide (GLDP) for dipeptide coupling, the ML model effectively captures the intricate vibrational interactions within proteins. Validation of the ML protocol was performed on various proteins under different pH conditions, demonstrating its ability to accurately predict hydrogen bond dynamics and structural changes during protein folding. These results are promising for identifying protein secondary structures and monitoring hydrogen bond dynamics during protein folding and under different pH conditions.

In summary, AI enables efficient and accurate simulations of IR and Raman spectra. Early efforts integrated ML with AIMD simulations to reduce computational costs, achieving significant gains by using neural network potentials to model complex systems like MXenes and nanosilicate clusters. Innovations such as TNEP have further extended the scope of these models to predict vibrational properties across a wide range of molecular scales. ML has been employed to predict detailed spectra for more complex structures, such as proteins, accurately capturing intricate molecular interactions. These advances enable faster predictions while maintaining accuracy, offering deeper insights into the relationship between molecular structures and their spectral characteristics, thus pushing the boundaries of vibrational spectroscopy analysis across diverse systems.

4.2 UV-Vis spectra

Efforts have been made to use ML to reduce the computational cost of UV-vis spectra and extend the applicability and transferability of spectral predictions across diverse molecular systems. Early research studies usually predicted the excitation properties first, and then used them to construct UV-vis spectra. The first notable work using ML to predict UV spectra appeared in 2015, exploring a hybrid approach to predict the electronic spectra of small organic molecules.[197] This study presented a Δ-ML model trained to correct the deviations of Time-Dependent Density Functional Theory (TDDFT) predictions from those obtained using the more accurate second-order approximate coupled-cluster singles and doubles (CC2) method. The authors used a dataset of over 20 000 synthetically feasible small organic molecules, training the ML model on the deviations between TDDFT and CC2 results for 10 000 of these molecules. The model successfully reduced the mean absolute error (MAE) of TDDFT predictions to within ±0.1 eV. The study found that the ML model could effectively overcome systematic errors in TDDFT predictions, such as underestimating excitation energies for π-type excitations and overestimating for σ-type excitations. Additionally, the performance of different molecular descriptors, including the Coulomb matrix and the bag-of-bonds, was evaluated, showing similar prediction accuracy for large training sets. Later, in 2020, an ML approach was proposed to model and predict the UV absorption spectra of molecules, focusing on excited states and transition dipole moments.[198] This approach can model these properties in a rotationally covariant manner, allowing for the simultaneous prediction of permanent and transition dipole moments, excited-state energies, and forces. The model was trained on data from methylenimmonium cation ($CH_2NH_2^+$) and ethylene (C_2H_4), accurately predicting UV spectra and electrostatic

potentials. By evaluating the transferability of the models to other molecules (CH_2NH, $CHNH_2$, and $C_2H_5^+$), they showed that ML models trained on multiple molecules perform better than those trained on single molecule. This approach can reduce computational costs compared to traditional quantum chemistry methods, making it feasible to study large molecular systems. Another study introduced a supervised ML approach to efficiently predict the absorption spectra of organic molecules, circumventing the high computational costs associated with traditional *ab initio* methods like Multi-Configurational Self-Consistent Field (MCSCF), Coupled Cluster, and TDDFT.[64] By utilising electronic properties from low-cost theoretical calculations, the researchers have characterized the spectroscopic fingerprints of small molecules. They have trained a CNN model on a database of 21 000 molecules. Ground-state DFT calculations using the LDA XC-functional provided electronic descriptors, while TDDFT calculations with the PBE0 hybrid XC functional were used for validation. The study explores different NN models, including MLP and CNN, optimized through Bayesian optimization. The results demonstrate that the NN can accurately predict the density of excited states, absorption spectra, and charge-transfer character and achieve near chemical accuracy ($\sim$0.1 eV) in excitation energy prediction. As a result, the authors significantly reduce computational costs and enable efficient spectroscopic property predictions.

Rather than predicting excitation energy and oscillator strength, a number of studies have demonstrated the capacity of ML techniques to directly predict UV-vis spectra. A model named UV-adVISor was created using Extended Connectivity Fingerprint (ECFP6) representations and Long-Short Term Memory (LSTM) networks to predict UV spectra from 220 to 400 nm.[199] Absorbance spectra were acquired from internal collections and analysed using HPLC and direct UV-vis spectrophotometry. The datasets were processed, normalised, and prepared for model training, considering different data representations, including ECFP6 fingerprints and tokenised SMILES strings. For medium-sized molecules (20–70 atoms), an ML model named UVvis-MPNN (message-passing neural network) was developed to predict UV-vis absorption spectra (Fig. 12).[200] MPNNs are particularly well suited for spectra prediction of larger molecules because they operate directly on molecular graphs, efficiently capturing both local electronic environments and global molecular connectivity that determine chromophore behaviour, thereby handling the increased complexity of larger systems. An automated workflow was created for structure optimisation and *ab initio* UV-vis spectra prediction, enabling direct comparison with experimental spectra. The ML model was trained on a data set of 1000 molecules, using TD-DFT calculations to obtain excited state energies and oscillator strengths, with solvation effects included. The ML models were validated against experimental spectra, showing potential for high-throughput spectra prediction in drug design.

In addition to accelerating computation and improving structure-based predictions for individual composition, a recent study[201] has focused on enhancing predictive capabilities on the spectra of mixtures using a small training data set. This set includes experimental UV-vis spectra from both substances and mixtures. A two-stage protocol for predicting absorption spectra of complex mixtures was developed, combining the Corrected Group Contribution (CGC) and molecule contribution methods with Bayesian Neural Networks (BNNs). BNNs provide a significant advantage for mixture spectra prediction because they quantify prediction uncertainty in addition to the spectral output, which is crucial when dealing with the inherent complexity and variability of intermolecular interactions in mixtures with limited training data. In the first stage, the CGC method utilises revised group contributions along with electronic and atomic connectivity descriptors to accurately predict the maximum wavelength and full spectra of a substance using a training set of fewer than 100 samples. In the second stage, the molecule contribution method considers intermolecular interactions and employs binary mixture data to train the model, enabling accurate prediction of multi-component mixture spectra.

To summarise, AI has advanced UV spectra prediction by overcoming the limitations of traditional quantum chemistry methods. Early studies applied the Δ-ML model to correct systematic errors in TDDFT predictions, achieving accuracy comparable to higher-level quantum chemistry methods without significant computational expense. Later developments showed ML's ability to directly predict complex properties,

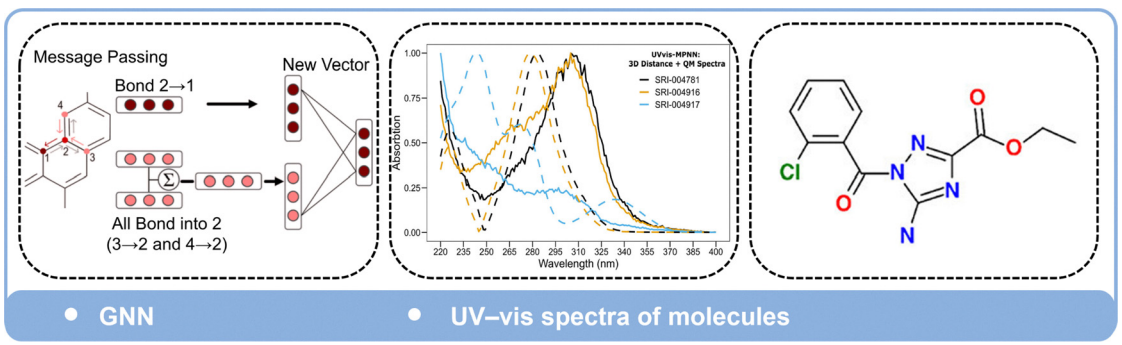


Fig. 12 Representative example of developing the GNN-based model to predict UV-vis spectra for organic molecules of sizes varying from 20 to 70 atoms. Reproduced with permission from ref. 200, Copy © 2023C, American Chemical Society.

such as excited-state energies and dipole moments, offering efficient and accurate alternatives to traditional calculations. Additionally, ML models have been extended to predict absorption spectra for complex mixtures, integrating molecular contributions and Bayesian neural networks, further expanding ML's applicability in spectral analysis across various systems.

4.3 CD spectra

ML has also been applied to predict CD spectra. In a recent study, Vermeyen *et al.* explored the capability of ML models to predict VCD spectra from the geometry of single molecular conformers.[202] They used a MLP trained on molecular geometries and DFT-computed VCD spectra for the tetra-substituted naphthalene framework. This approach allowed them to assess the transferability of their ML models across various substituents in the molecules, enabling control over conformational flexibility and intramolecular interactions. The results showed that ML could accurately predict VCD spectra from the geometry of a conformer, significantly reducing computational costs. However, the ML models were not transferable between different molecules or configurations of the same molecule, which requires further improvement. This research highlights ML's potential in VCD spectroscopy, offering a promising approach to reducing reliance on extensive quantum chemistry calculations.

Very recently, the ECDFormer model has been introduced for efficiently predicting ECD spectra of chiral molecules using deep learning (Fig. 13).[203] This model focuses on learning peak properties such as the number, position, and symbol of peaks. Specifically, their model was trained on a dataset containing ECD spectra for 22 190 chiral molecules. The ECDFormer model comprises four main modules: molecular feature extraction, peak property learning, peak property prediction, and ECD spectra rendering. The model employs GeoGNN,[204] a transformer encoder, for geometric feature encoding and MLPs for peak property predictions. Experimental results demonstrate that ECDFormer outperforms baseline models in accuracy and efficiency, particularly for complex molecules and natural products. Additionally, the model generalises well to multi-chiral-centre molecules with minimal performance decline.

For large molecules, based on a tensorial embedded atom neural network (T-EANN) and embedded density descriptors, Zhao *et al.* developed a model to predict CD spectra of proteins.[196] This model addresses the challenges of accurately predicting electric and magnetic transition dipole moments of peptide bonds. The ML model employs descriptors invariant to translation, rotation, and permutation, combining virtual ML outputs with atomic coordinates to obtain symmetry-conserving tensors. The ML predictions of permanent dipole moments and excitation energies align perfectly with their DFT and TDDFT values, reaching Pearson coefficients over 0.95 and MREs below 0.3%. Regarding electric and magnetic transition dipole moments, the model also achieved high accuracy, with Pearson coefficients over 0.95 and mean relative errors below 1.5% compared to TDDFT values. Besides the its excellent efficiency and accuracy, the ML model was also used to predict a series of CD spectra associated with changes in the protein configuration along its folding path, showing its potential for real-time spectroscopy studies on protein dynamics.

In summary, integrating AI with CD spectroscopy greatly enhances the prediction of CD spectra from molecular structures. Models like ECDFormer have demonstrated the ability to efficiently predict ECD spectra for complex molecules using deep learning techniques supported by large datasets. While early efforts in VCD prediction highlighted challenges with transferability, recent developments in ML-based models have shown promise in overcoming these limitations, especially in predicting dipole moments in proteins. These advancements pave the way for faster, more accurate spectra predictions, expanding the practical use of CD spectroscopy in complex molecular systems and real-time analysis.

4.4 Fluorescence spectra

Conventional computational methods for fluorescence spectroscopy require time-consuming excited-state calculations, and the results of the computations frequently deviate from experimental observations. The introduction of the ML method has successfully reduced the cost of computational simulations and can be used to predict emission wavelengths and intensities, quantum yields, emission lifetimes, and other properties of a substance, in direct comparison with experimental results.

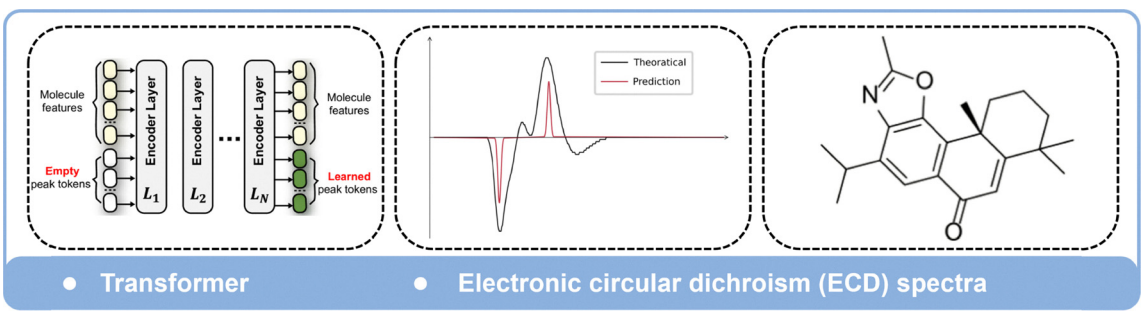


Fig. 13 Representative example of developing a transformer-based model (ECDFormer) to predict the peak properties (peak number, position, and symbol) of ECD spectra of chiral molecules. Reproduced with permission from the preprint version of ref. 203.

It has been demonstrated that various ML methods can be employed to predict fluorescence spectra. A relatively straightforward approach is to use a molecular graph representation combined with a graph neural network to make predictions regarding fluorescence properties. A paradigmatic example is the study of Joung *et al.*[205] In their study, molecular structure and solvent information were input into the graph neural network, enabling the intelligent model to learn the interaction effects of chromophores and solvents. This model subsequently made successful predictions regarding the absorption peak position and bandwidth, extinction coefficient, emission peak position and bandwidth, the photoluminescence quantum yield (PLQY), and the emission lifetime. The model was also employed to perform virtual screening to identify a blue-emitting organic molecule. Another molecular representation that has been widely adopted is molecular fingerprint-based descriptors. These have a fixed input length and are suitable for combining with various machine learning methods, such as SVM, MLP, RF, *etc.* For instance, Ju *et al.* used a range of models to predict the emission wavelength and PLQY of solvated organic fluorescent dyes.[206] A comparison was made between SVM, kernel ridge regression (KRR), MLP, KNN, light gradient boosting machine, and gradient boost regression tree. In the context of the prediction task concerning PLQY, the MAE has been observed to attain a maximum of 0.21. Notably, several studies are analogous to this study, including the prediction of the absorption and emission wavelengths of fluorescent dyes under differing solvent conditions,[207] and the prediction of the major fluorescence peaks of carbon dots.[208]

4.5 NMR spectra

Research on predicting NMR chemical shifts with AI models has been focused on various systems, including organic small molecules,[209–215] carbohydrates,[216] nucleic acids,[217] proteins,[218] and molecular solids.[178,219]

For organic molecules, Kang *et al.* predicted the NMR chemical shifts without any atomic annotation.[209] In particular, they built an MPNN model that employs a loss function invariant to atomic permutations. Moreover, the model is trained using weakly supervised learning to minimise discrepancies between a set of predicted and actual chemical shifts in a molecule rather than individual atomic chemical shifts. As a result, this model allows for predicting atomic-level chemical shifts based on NMR spectra without manual annotations from NMR experts. Another approach by Gao *et al.*[210] accurately predicted ^{13}C and ^{1}H chemical shifts using a DNN model based on a combination of chemical environment descriptors and DFT-calculated isotropic shielding constants. In particular, the chemical environment descriptors do not need any pre-knowledge of molecular structure, including atomic number, total valence, minimum ring size to which the atom belongs, Gasteiger charge, Crippen log$P$ contribution, Crippen molar refractivity contribution, and so on. Training the model with a data set of 476 ^{13}C and 270 ^{1}H experimental chemical shifts, the authors demonstrated that the model is more accurate from bare DFT predictions or simple linear regression based on DFT results. The introduction of DFT-calculated descriptors can effectively improve prediction performance. Han *et al.* proposed inputting DFT-optimised 3D structures into a GNN model and combining them with shielding tensor descriptors computed from DFT to directly predict ^{1}H and ^{13}C chemical shifts.[220] Following training on 26 913 ^{13}C chemical shifts and 12 806 ^{1}H chemical shifts, this model achieved an average absolute error of 0.944 ppm for ^{13}C and 0.185 ppm for ^{1}H.

Obtaining high-precision NMR chemical shifts from electronic structure theories can be challenging. As a result, many ML approaches have been developed to improve the accuracy of NMR chemical shift predictions from quantum chemistry calculations. Unzueta *et al.*[211] proposed a Δ-ML approach for predicting NMR chemical shielding constants, with errors of only one-half to one-third of those from DFT compared to the experimental values. In their model, the authors start with a baseline low-accuracy DFT calculation to obtain an initial estimate of isotropic chemical shielding constants. Then, an NN is applied to improve this estimate to the accuracy of target shielding constants of a higher-level theory. The model used atomic environment vectors (AEVs) as input descriptors to capture the correlation between chemical shift and the local chemical environment of an atom, considering the inexpensive baseline calculation already accounted for the long-range interactions. In this way, the Δ-ML model leads to an accurate prediction of chemical shifts by including these crucial long-range interactions without including large molecules in the training set. Following the idea of Δ-ML, Li *et al.* developed the iShiftML model[214] based on active learning, which reaches the prediction accuracy of chemical shifts to the gold-standard CCSD(T)/CBS level. Similarly, this model used the diamagnetic and paramagnetic shielding tensor elements from DFT calculations and AEV descriptors of the small molecules as the input. It demonstrates remarkable transferability, allowing for precise predictions of chemical shifts across various systems, including previously unseen molecules from the NS372 data set, gas-phase experimental small organic molecules, and the more complex vannusal B.

For different types of substances, it is often difficult for generic models to achieve the best predictive performance. Chen *et al.* showed that the structural representations in existing models are mainly for proteins and small molecules,[204,221] and may not be usable in carbohydrate systems without specific modifications.[216] They constructed a GlycoNMR dataset containing 2609 carbohydrate structures and 211 543 atomically labelled chemical shift data for saccharides. Training 2D GNN models on this dataset enabled the prediction of ^{1}H and ^{13}C chemical shifts with RMSEs of 0.13 ppm and 1.9 ppm, respectively, and 0.09 ppm and 0.51 ppm on five state-of-the-art 3D GNNs. This demonstrates the importance of dataset quality and model representation when solving spectroscopy-related problems using machine learning.

Other than small organic molecules, several advancements have been made in NMR chemical shift prediction for biomolecules, with various models focusing on proteins and nucleic acids.[218,222,223] A noteworthy study is the development

of SHIFTX2[222] by Han *et al.* in 2011, a well-known tool for predicting ^{1}H, ^{13}C, and ^{15}N chemical shifts of proteins. SHIFTX2 can predict a broader range of backbone and side-chain chemical shifts accurately. This performance is achieved by using a large, high-quality database of over 190 training proteins and the incorporation of additional features such as χ_2 and χ_3 angles, solvent accessibility, hydrogen bond geometry, pH, and temperature, and combining both sequence-based and structure-based chemical shift prediction techniques to improve its predictive power further. Yang *et al.*[218] developed a GNN model to predict NMR chemical shifts of proteins directly from 3D structures without feature engineering. The GNN can capture important phenomena like hydrogen-bond-induced downfield chemical shifts between multiple proteins and secondary structure effects, as well as predict chemical shifts of organic molecules. Chandy *et al.*[217] developed an RF model for the prediction of ^{1}H and ^{13}C NMR chemical shifts of nucleic acids. They trained the model using chemical shifts computed *via* a DFT protocol based on their molecules-in-molecules fragment method, which accounts for microsolvation effects. They refined the model on a compact dataset comprising 2080 ^{1}H and 1780 ^{13}C chemical shift values by incorporating both structural and electronic descriptors obtained from low-level semiempirical calculations (GFN2-xTB). It achieved good performance when tested on 8 new nucleic acids, with sizes ranging from 600 to 900 atoms.

Lyndon Emsley and co-workers presented a GPR framework for predicting chemical shifts in molecular solids, named ShiftML, which uses the smooth overlap of atomic positions (SOAP) kernel[224] to represent the local atomic environment around each atom (Fig. 10B).[178] Gaussian process regression with SOAP kernels is particularly effective for NMR prediction because it provides a mathematically elegant way to map the similarity between atomic environments to similarities in chemical shifts, while inherently quantifying prediction uncertainty—critical for crystalline materials where small structural differences can cause significant spectral changes. The predicted chemical shielding for a given atom is calculated as a weighted sum of the SOAP kernel evaluations between the target atomic environment and those of the training configurations, for which the chemical shifts are determined through GIPAW DFT calculations. An updated version of this model, named ShiftML2, was trained on GIPAW DFT chemical shifts for an extensive set of over 14 000 structures.[219] These structures included 12 common elements and comprised both relaxed and thermally perturbed crystal structures. ShiftML2 showed slight improvements over previous versions when applied to DFT-relaxed structures and maintained accuracy for distorted structures, which posed significant challenges for ShiftML1. This version also enabled chemical shift computations for more chemically diverse structures.

In summary, AI has been proven highly effective in predicting NMR chemical shifts across various systems. For small organic molecules, models such as MPNN and DNN can efficiently and accurately predict the chemical shift from the molecular structure using the chemical environment and/or DFT-based descriptors. Δ-ML models facilitate high-accuracy prediction of NMR chemical shifts from computationally efficient quantum chemistry calculations. For more complicated protein and biological systems, models such as SHIFTX2 and GNN predict NMR chemical shifts directly from 3D protein structures, accounting for hydrogen bonding and secondary structure effects. Additionally, ShiftML combines ML with GPR and the SOAP kernel to predict chemical shifts in molecular solids, capturing the correlation between chemical shifts and local atomic environments. These AI advancements enable more accurate, scalable, and automated analysis of complex molecular systems, reducing reliance on experimental data.

4.6 MS

As mentioned, CFM-ID[155,156] is a program designed to accurately predict ESI-MS/MS spectra for a given compound structure, enabling compound identification. Key innovations in CFM-ID version 4.0 include a refined method where parameters are learned directly from the molecular topology, which improves the prediction accuracy of the model. Additionally, a novel approach has been introduced to model ring cleavage as a sequence of simple chemical bond dissociations, enhancing the accuracy of fragmentation simulations, particularly for compounds with ring structures. Furthermore, the rule-based predictor has been expanded to cover a broader range of chemical classes, including acylcarnitines, acylcholines, flavonols, flavones, flavanones, and flavonoid glycosides, greatly increasing the diversity of compounds the model can handle. Wei *et al.*[225] proposed Neural Electron–Ionization Mass Spectrometry (NEIMS), an NN model designed to predict the Electron Ionisation Mass Spectrometry (EIMS) spectra for small molecules. In this model, the prediction is approached by a multidimensional regression model, where the output is a vector representing the intensity at each m/z bin. The model uses ECFPs to represent molecules, which captures molecular subgraphs and records the frequency of occurrence for each subgroup rather than just presence. These fingerprints are processed through an MLP. NEIMS employs a bidirectional prediction mode, where the forward prediction leverages ECFP to predict the intensities of small fragment peaks directly, and the reverse prediction accounts for larger fragments resulting from neutral losses by considering the mass changes due to the removal of residual groups from the molecule. A coordinate-wise gating mechanism optimises the weighted combination of the forward and reverse predictions. By utilising this bidirectional prediction mode, the model can handle peaks generated by both small molecular fragments and cleavage events, thereby improving overall prediction performance. Zhu *et al.*[226] proposed rapid approximate subset-based spectra prediction (RASSP) for predicting EIMS spectra of small molecules. This model combines physically plausible substructure enumeration and deep learning. It utilises two models: FormulaNet generates a probability distribution over chemical subformula, while SubsetNet determines probability distributions across vertex subsets within molecular graphs. Collectively, these models demonstrate comparable predictive accuracies and

excel in generalisation under scenarios characterised by high resolution but limited data.

Zhou *et al.*[227] used a Supporting Vector Regression (SVR) algorithm to predict Collision Cross-Section (CCS) values, which are critical for metabolite identification in Ion Mobility-Mass Spectrometry (IM-MS). The SVR model was trained using 14 molecular descriptors, including accurate mass, formal and physiological charge, octanol/water partition coefficient, aqueous solubility, acid and base dissociation constants, acceptor and donor atom sums, polar atom surface area, rotatable bonds, molar refractivity, and molecular polarizability. Using the developed model, researchers assembled an extensive CCS database encompassing 35 203 metabolites from the Human Metabolome Database (HMDB). This new database includes predictions for five distinct ion adducts operating in both positive and negative ion modes, culminating in a total of 176 015 predicted CCS values.

In summary, AI is increasingly applied in mass spectrometry to improve MS spectra prediction. Programs like CFM-ID and NEIMS use machine learning to predict ESI-MS/MS and EIMS spectra by modelling molecular topology, fragmentation, and substructure. RASSP combines substructure enumeration with deep learning for high-accuracy predictions. Additionally, AI models such as SVR can predict CCS values in IM-MS, aiding metabolite identification. These innovations demonstrate how AI can significantly enhance MS spectra prediction, further improving compound identification by reducing reliance on extensive experimental data and enabling faster, more accurate predictions, even in complex or low-data scenarios.

4.7 X-Ray spectra

Some proof-of-concept studies were first conducted to predict X-ray spectra of small organic molecules using ML models. An MPNN was first proposed by Carbone *et al.* to predict XANES spectra for small organic molecules.[228] The model was trained on DFT-calculated XANES spectra of oxygen and nitrogen K-edges from the QM9 database, which contains approximately 134 000 small organic molecules. The model achieved high accuracy, predicting peak locations within 1 eV of the ground truth for 90% of cases. Furthermore, it concludes that the functional group is a key descriptor of XANES spectra and emphasises the importance of the local chemical environment in predicting XANES spectra. Following this idea, Kotobi *et al.* investigated[179] the interpretability of GNN models when predicting XAS (Fig. 10C). Specifically, GCN,[229] GraphNet,[230] and GATv2[231] models are trained using carbon K-edge XAS spectra of 65 000 molecules from the QM9 set. GATv2's attention mechanism provides an advantage for XAS prediction by dynamically weighting the importance of different atomic neighbours during message passing, effectively modelling the varying contributions of different atomic contributions for the absorption. The results show that the GATv2 model has a slightly lower average RSE value than GraphNet and GCN, indicating its superiority for XAS predictions. More importantly, this study investigates the interpretability of the GNN structure *via* quantitative analysis of atomic contributions to each peak in the spectra, demonstrating that the degree of explainability in different architectures of GNN models differentiates their predictions. GNN models that can effectively capture both the local and more global chemical environments of an atom within a molecule not only predict better spectra but also provide explanations that align well with the quantum mechanical interpretations of XAS.

Penfold and co-workers conducted a series of studies on predicting X-ray absorption spectra for transition metal materials.[232–237] They initially began by exploring the impact of different structural representations on the performance of a DNN for predicting Fe K-edge XANES spectra.[232] They compared the CM and radial distribution curve (RDC) representations, finding that RDC provided better performance, with lower mean squared error and faster convergence. Then, they trained a DNN on more than 9000 Fe K-edge XANES spectra sourced from the materials project database, with only the local geometry around the absorbing atom as input.[233] This initial implementation accurately predicts peak positions and intensities. The model's performance is demonstrated through its application to the structural refinement of tris(bipyridine)iron(II) and nitrosylmyoglobin. They then optimised a DNN to predict Co K-edge XANES spectra instantaneously. They applied it to analyse T-jump pump/X-ray probe data of Co^{2+} in chlorinated aqueous solution, revealing structural changes primarily driven by sample heating and an increased Debye–Waller factor.[234] In 2022, they developed XANESNET, a deep neural network for predicting the lineshape of first-row transition metal K-edge XANES spectra using local coordination geometry encoded in atom-centred symmetry functions (ACSFs).[235] ACSFs provide an ideal descriptor for XANES prediction because they efficiently encode the local three-dimensional arrangement of atoms around the absorbing centre in a way that preserves rotational and translational invariance, while capturing the critical coordination geometry that determines XANES spectral features. XANESNET achieves an average error of 2%–4% and matches prominent peak positions with over 90% accuracy within sub-eV precision. This model was then extended to predict $L_{2/3}$-edge spectra,[236] and valence-to-core X-ray emission spectra (Fig. 14).[237] In addition, an artificial intelligence *ab initio* (AI-ai) framework was proposed to predict the XPS spectra of the solid electrolyte interphase of lithium metal batteries by combining hybrid *ab initio* and reactive molecular dynamics and ML models.[238]

The integration of AI represents a transformative advancement in predicting X-ray spectra, particularly XANES. Models like MPNNs and XANESNET provide rapid and accurate spectra predictions, significantly reducing computational costs compared to traditional methods. Key studies, especially those focused on transition metal materials, have refined ML approaches to handle complex spectra and revealed dynamic structural changes. Additionally, ML has been applied to molecular excitation and photoemission spectra, with neural networks like DTNN and SchNet capturing intricate spectral features.

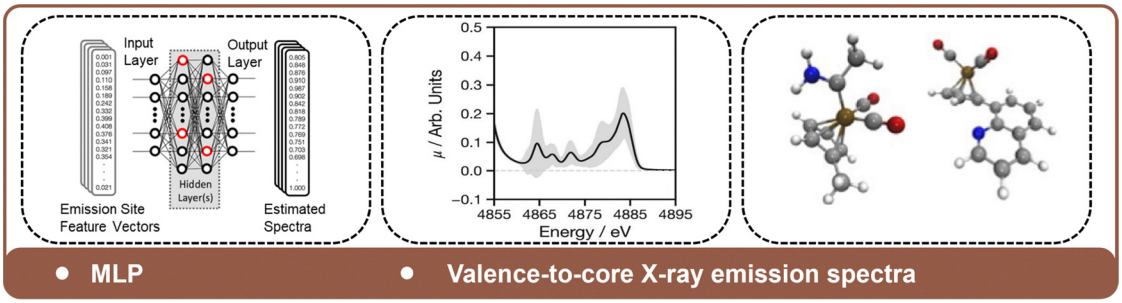


Fig. 14 Representative example of using the MLP model to predict the first-row transition metal K-edge valence-to-core X-ray emission spectra (VtC-XES) for metal-containing complexes. Reproduced with permission from ref. 237.

4.8 Multiple spectra

In addition to predicting a single type of spectra, efforts have also been made to develop ML models to predict multiple spectra simultaneously. Zhang *et al.* simulated vibrational and electronic spectra (Raman and UV-vis) in molecular and condensed phase systems using symmetry-preserving neural network models named T-EANN.[239] In this model, the tensorial response and transition properties, such as transition dipole moments, are learned in a way as the multiplication of their NN output and atomic coordinates, which maintain their rotationally covariant symmetry. This strategy allows the tensorial NN models to be as efficient as their scalar counterparts, allowing effective prediction of the response and transition properties of water oligomers, liquid water, and protein units. For vibrational and NMR spectra, Gastegger *et al.* proposed FieldSchNet, a DNN that models molecular interactions with arbitrary external fields in another study.[183] The FieldSchNet model allows for the simulation of molecular spectra such as IR, Raman, and NMR spectra and the evaluation of molecule–environment interaction, such as the solvent effects within implicit or even explicit solvent models. The model is applied to simulate liquid ethanol IR spectra, where it accurately describes solvent effects and hydrogen bonding interactions. In addition, the authors utilised FieldSchNet to investigate the Claisen rearrangement reaction. This work highlights its potential as a tool for inverse chemical design, as it successfully demonstrates how to create an external environment that significantly lowers the activation barrier of the rearrangement.

Recently, a novel network named Deep Equivariant Tensor Attention Network (DetaNet) has utilised equivariant MPNN and attention mechanisms to simulate IR, Raman, UV spectra, and $^{1}H/^{13}C$ NMR spectra (Fig. 15).[240] DetaNet's equivariant architecture is uniquely suited for multi-spectral prediction because it preserves physical symmetries across different tensor orders (scalars, vectors, and higher-order tensors), allowing it to simultaneously model diverse spectroscopic properties that transform differently under rotation while sharing representational power across related physical phenomena. This network requires only molecular coordinates as input to effectively predict various molecular properties, including scalars (energy, charge), vectors (atomic forces, dipole moment), second-order tensors (quadrupole moment, polarizability), and third-order tensors (first hyperpolarizability, octupole moment). When tested on the QM9 and QM7-X[241] datasets, DetaNet achieved an R^2 greater than 99.9% for all properties. In terms of computational efficiency, DetaNet improves by three to five orders of magnitude compared to DFT calculations. DetaNet provides a foundational knowledge base and powerful software tools for molecular identification, nanostructure characterisation, reaction mechanism analysis, and elucidation of structure–activity relationships.

5. From spectra to functional properties

Intelligent models for predicting material properties are often constructed from structural information. However, the

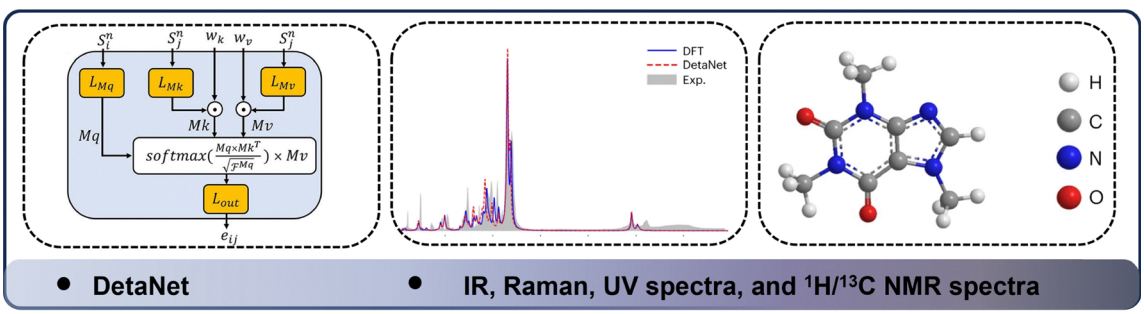


Fig. 15 Representative example of developing the DetaNet model to predict multiple types of spectra, including the IR, Raman, UV spectra, and $^{1}H/^{13}C$ NMR spectra. Reproduced with permission from ref. 240.

correlation between high-dimensional structures and lower-dimensional properties is often highly complex. Also, it is almost impossible to accurately determine their equilibrium structures for some complex systems, such as high-entropy alloys and polymers. It is, therefore, necessary to identify more general descriptors than those derived from structural analysis.

Spectra are an outstanding alternative to molecular structure-related descriptors in ML for property predictions. Unlike the molecular structure, spectra are not only closely related to the chemical structure but also carry information about the electronic structure, which is also crucial for understanding matter properties. Moreover, spectra represent projections of the high-dimensional physical world onto lower-dimensional spaces, making themselves a lower dimension than complex, high-dimensional structures. Hence, using spectra as descriptors could solve the cases where building complex structural models of matter is impossible, and it also lowers the number of parameters needed for ML methods. In addition, spectra can be regarded as the latent space between structures and properties, facilitating the learning of their correlation with properties by ML models. They are also experimentally measurable and intrinsically physically meaningful, enhancing the interpretability of the model. Therefore, a growing number of studies have explored the use of spectra as descriptors for predicting chemical properties.

This section outlines some of the recent advances in the prediction of substance properties from spectra using ML models. It has been demonstrated that intelligent models can accurately predict a range of chemical properties from spectra, including molecular adsorption properties on the surfaces of different materials, chemical bonding properties of molecules, and the electronic state topology of materials. As research progresses, it is anticipated that an increasing number of spectral-property correlations will be identified.

5.1 Vibrational spectra

The Jiang group conducted a series of studies and found that IR and Raman spectra can effectively predict the adsorption properties. Using the example of CO adsorption on metal surfaces, Wang *et al.* built using RF and CNN models, which accurately predict interaction properties from the IR and Raman spectral features. These properties include CO adsorption energy, charge transfer between surface and absorbance, CO bond energy, and d-band center.[63] Furthermore, the authors explored the interpretability of the models using the sure independence screening and sparsifying operator (SISSO) method,[242] where they construct quantitative mathematical formulas between the ML-predicted interaction properties and the spectral features. These formulas show excellent universality and predictive power, with typical r values in predictions exceeding 0.8 and many surpassing 0.9. Using spectroscopy-based formulas allows one to decouple the effects of the substrate and the adsorbate. In these formulas, the adsorbate's spectral signals, which are readily measurable at a macroscopic scale, serve as the variables, while the parameters reflect the substrate's inherent characteristics. The models also showed the ability to transfer predictions to different molecules (*e.g.*, NO) on various surfaces (gold, silver, and gold–silver alloys) and to predict the properties of multi-molecule adsorption states. Chong *et al.* also predicted the adsorption energy and charge transfer of CO@CuBTC, a copper-based MOF, using a similar approach of constructing highly transferable and interpretable mathematical formulas between interaction properties and spectral features (Fig. 16).[243] These formulas demonstrated greater robustness and fault tolerance than NN models, even with only 20 samples to calibrate across different adsorption systems and exhibiting higher accuracy in detecting suspicious data. This study shows that mathematical formulas better describe general physical principles than the traditional neural network approach, particularly when data is scarce or of low quality. Aside from predicting the properties at the lowest energy configuration of adsorption system, they have also used ML models to predict statistic adsorption properties *via* IR and Raman spectra. Since actual adsorption systems arise from a statistical distribution of molecular configurations, with conformational probabilities adhering to an energy-based distribution, they construct an ML model for the prediction of the average values of important adsorption properties from conformationally averaged spectra, including adsorption energy, charge transfer between CO and alloy surface, and CO bond energy.[244] The results show that, even when dealing with a large and undefined number of surface molecules, the method demonstrates exceptional predictive accuracy. Furthermore, the quantitative relationships

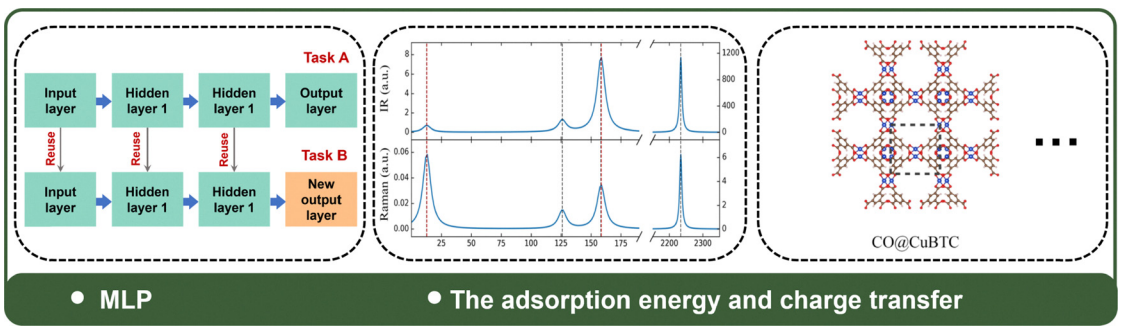


Fig. 16 Performing MLP to predict the adsorption energy and charge transfer of CO adsorbed on a copper-based MOF. Reproduced with permission from ref. 243.

based on averaged spectra provide a theoretical basis for isolating essential interaction characteristics observed in real chemical settings.

Wang *et al.* also explored using vibrational spectral features to explore catalytic properties by employing unsupervised ML on vibrational spectra of 300 zinc-based metal–organic frameworks (Zn-MOFs).[245] Their work demonstrates how DFT-calculated IR and Raman spectral features can be translated into insights about the electronic and catalytic attributes of MOFs, closely mimicking the behaviour of natural enzyme carbonic anhydrase (CA). By harnessing the high-dimensional chemical information inherent in vibrational spectra, this research not only identifies MOFs with potential as efficient CO_2 converters but also expands the utility of spectroscopic tools in designing biomimetic catalysts. This innovative ML application to spectroscopic data enriches the ability to predict and optimise the catalytic performance of the field, setting a new standard for enzyme-mimicking material discovery. Beyond molecular adsorption on solid surfaces, spectra can also effectively predict molecular adsorption properties at different solid–liquid interfaces.[246] Du *et al.* developed a hierarchical knowledge extraction multi-expert neural network (HMNN) and trained it in two steps: first, learning the relationship between infrared and Raman spectra and CO adsorption properties and then learning the differences in these relationships across different solvent systems. The hierarchical multi-expert neural network architecture is uniquely suited for complex interfacial systems because it can first learn general spectra–property relationships common across environments, then specialize to the unique aspects of specific interfaces, enabling accurate zero-shot predictions in new solvent environments with minimal additional training. They conducted a series of experiments on three previously unseen solvent systems: isopropanol (IPA), dimethyl sulfoxide (DMSO), and ethylene carbonate (EC). In these solvent systems, the model stably achieved zero-shot predictions of adsorption properties at chemical accuracy levels. The results demonstrate that using spectra as descriptors, neural networks can be trained to acquire and integrate knowledge from diverse and related domains and then fine-tune with small amounts of data to enhance generalisation capabilities for universal property predictions.

In addition to molecular adsorption interactions, spectroscopy can also be used to predict more complex interactions, such as protein–ligand binding. In a recent study, Chen *et al.* proposed spectra as a descriptor to construct a ''spectra–properties'' relationship with the nature of protein–ligand interactions, which would ultimately drive the discovery of effective ligand molecules for target proteins.[247] A fragment integral spectrum (FIS) descriptor was constructed based on the IR spectra of protein and ligand molecules in order to represent the molecule itself, rather than structural descriptors. A neural network was then trained to predict whether the ligand can bind to the protein directly from the spectra. The model demonstrated excellent transferability and successfully recommended binding ligands for unknown target proteins. It was validated on target proteins related to Alzheimer's disease and SARS-CoV-2 virus. Further analysis demonstrated that the model based on FIS descriptors exhibited superior performance in comparison to models employing molecular graph or chemical information descriptors. This may be due to the fact that the structural descriptors principally contain atomic position information, which is not directly related to interactions, while the spectra have embedded molecular electronic and vibrational state information, which is directly related to molecular interactions, and thus have better performance in predicting interaction features. This study underscores the immense potential of spectroscopic descriptors in the realm of drug discovery and offers a novel perspective on their function.

In general, AI models shed new light on vibrational spectra analysis, particularly in predicting surface adsorption properties. Utilising IR and Raman spectral features as input, AI models such as CNNs and decision trees accurately predict key properties like adsorption energies, charge transfer, and bond geometries in systems like metal surfaces and MOFs. Techniques like SISSO enhance the robustness and transferability of these models, generating interpretable formulas with minimal data input. Additionally, AI-driven approaches extend the applicability of spectroscopic predictions to more complex environments, such as solid–liquid interfaces, and provide novel insights into catalytic mechanisms, paving the way for more efficient material design.

5.2 NMR spectra

NMR spectra carry information about the atomic chemical environment and are well-suited as descriptors for predicting molecular properties. NMR spectra were used to predict molecular properties using a stereo molecular graph bidirectional encoder representations from transformers (SMG-BERT) model (Fig. 17).[248] The transformer-based SMG-BERT architecture is particularly effective for NMR-based property prediction because its self-attention mechanism can capture long-range dependencies between different parts of a molecule that influence NMR shifts, while its bidirectional nature enables it to contextualize each atom's environment based on the entire molecular structure. This model integrates three types of data: 3D geometric parameters, 2D connectivity information and 1D SMILES representations. The SMG-BERT model was trained using the PubChem dataset and incorporates NMR chemical shifts and bond dissociation energies (BDEs) as chemical descriptors to improve interpretability. The model uses MPNNs and transformer encoder layers to generate accurate chemical representations. The pre-training involves tasks like atomic and NMR reconstruction, bond energy prediction, and 3D information reconstruction, ensuring rich information in the atomic representation. Based on benchmark results across 12 molecular datasets, SMG-BERT demonstrates the significance of incorporating NMR spectroscopy data in predicting molecular structures and properties. Additionally, it shows that NMR data enhances the interpretability of the model, aligning it more closely with chemical logic.

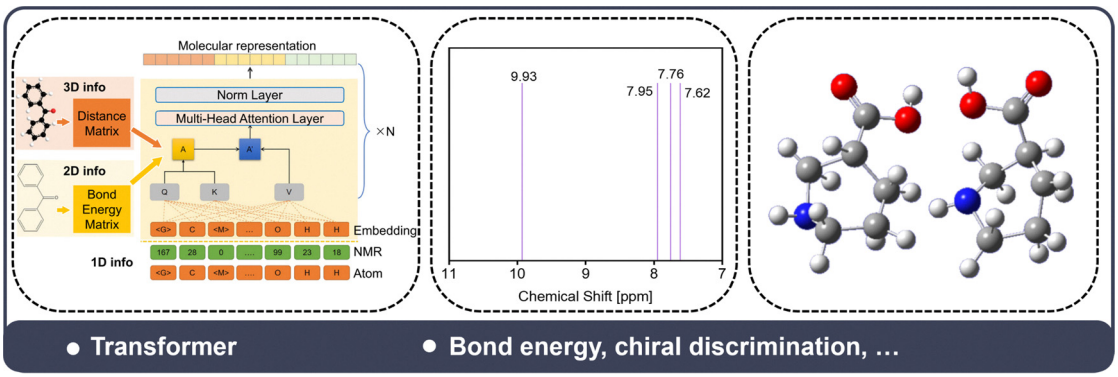


Fig. 17 Developing a transformer-based model to predict molecular properties, NMR chemical shifts, and BDEs as chemical descriptors. Reproduced with permission from ref. 248.

Different types of spectra characterise various dimensions of information about molecules, and their combined analysis can significantly enhance the prediction of molecular properties. This was demonstrated in a study by Guo *et al.* predicting the chemical properties of hydroxyl groups, including BDE, bond length, and α-C connectivity, by fusing spectral information from different types of spectra.[249] The results showed that using a single type of spectrum, such as infrared, Raman, or NMR spectroscopy, did not achieve high prediction accuracy. However, integrating vibrational spectra and NMR spectroscopy features greatly improved the model's performance. The model also shows strong transferability when applied to different molecules or solution environments.

Overall, AI empowers the NMR spectroscopy to predict molecular properties directly. State-of-the-art ML models such as SMG-BERT combine NMR data with structural features, improving predictions of chemical properties. Integrating NMR with other types of spectra, such as IR and Raman, in an AI model further enhances predictive accuracy, particularly for complex systems. This highlights the importance of using diverse spectral information to capture the multifaceted nature of molecular interactions and properties.

5.3 X-ray spectra

Using ML technology, researchers can predict the properties of materials directly from X-ray spectra, avoiding the high costs of theoretical calculations or experimental measurements. For instance, Andrejevic *et al.* built an ML model to classify topological and trivial materials based on their XANES spectra (Fig. 18).[250] The model was trained on computed XANES spectra of over 10 000 inorganic materials, which are from either a published database[251] or from materials project, with their XAS computed using the FEFF9[82] program. The materials were labelled based on the Topological Quantum Chemistry (TQC) framework, resulting in a dataset of 13 151 materials. The model demonstrated an overall accuracy of approximately 90% on the test set, showing high effectiveness in identifying topological materials. This method offers a promising pathway for high-throughput screening of candidate topological materials. It could be applied to study materials under various conditions, such as electric, magnetic, or strain fields, and even in disordered or amorphous states.

Based on XRD spectra, Zhang *et al.* introduced an unsupervised learning approach to discover solid-state lithium-ion conductors (SSLCs).[252] This method overcomes the scarcity of property data on labelled materials by using unsupervised learning to identify potential SSLC candidates without requiring extensive conductivity measurements. The researchers utilised a dataset of 2986 lithium-containing compounds from ICSD[84] and represented their structures using modified X-ray diffraction (mXRD) patterns focused on the anion lattice. They performed clustering to group materials with similar mXRD representations, leading to the identification of groups enriched with known SSLCs. AIMD simulations were then used

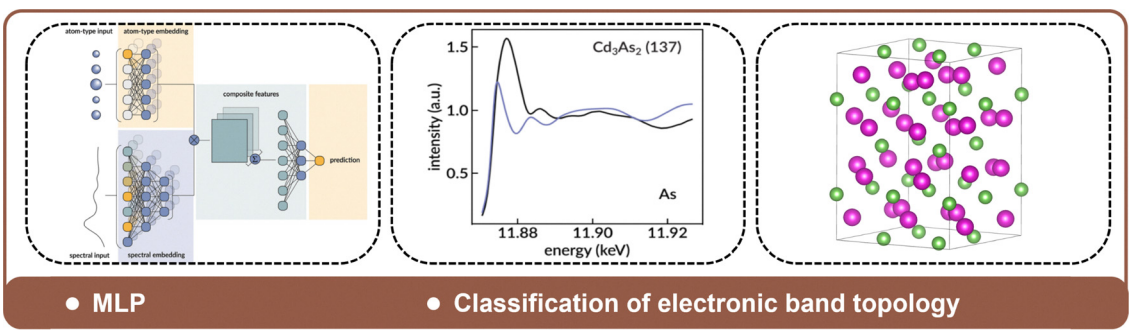


Fig. 18 Using MLP models for the classification of electronic band topology of materials *via* their XAS spectra. Reproduced with permission from ref. 250.

to validate the conductivity of the predicted candidates. This approach successfully identified 16 new SSLCs with room-temperature conductivities higher than 10^{-4} S cm^{-1}, including three with conductivities exceeding 10^{-2} S cm^{-1}. The study demonstrates that unsupervised learning can effectively guide the discovery of new materials with high ionic conductivities, significantly reducing the computational cost and effort associated with high-throughput screening.

AI techniques have transformed X-ray spectroscopy, such as XANES and XRD, into a powerful tool for analysing and discovering materials with specific properties. In XANES-based studies, ML models have been able to accurately classify topological materials, a key advancement for identifying materials with unique electronic structures. This capability not only accelerates the screening of topological materials but also provides a novel tool for studying complex electronic properties. On the other hand, using unsupervised learning in XRD data allows for identifying SSLCs based on structural patterns without requiring extensive experimental measurements. This technique successfully reveals candidates with high ionic conductivity, emphasising ML's efficiency in material discovery.

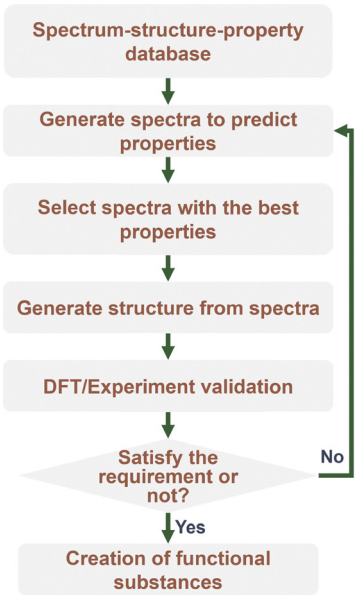


Fig. 19 Flowchart for the on-demand creation of new substances based on the spectrum–structure–property relationship.

6. From spectra to structures and properties: AIGC inverse design

As outlined in Sections 3–5, with ML methods, spectroscopy can serve as a bridge to construct quantitative relationships between structure and properties, allowing for the discovery of spectroscopic characteristics of materials with specific properties. Furthermore, the correlation between spectra and structures strengthens method development for structure generation in spectroscopic space. The use of spectra to link structure and properties is expected to facilitate the on-demand design of functional substances *via* Artificial Intelligence Generative Content (AIGC) methods (Fig. 19). This design process can be described below. First, theoretical datasets linking spectra, structures, and properties are assembled, providing the basis for generating initial (possibly random) spectra. In the next step, we can predict the properties of matter, allowing the spectra corresponding to the best properties to be selected. These selected spectra are then converted into candidate structures, whose properties are verified *via* DFT calculations and experimental measurements. If the experimental results deviate from predictions, they are fed back into the ML model, prompting the generation of new spectra. Through repeated cycles of theoretical modelling and experimental validation, the ML model continually improves, ultimately enabling the creation of matter that satisfies the desired target conditions.

Recently, Yang *et al.* explored the generation of catalyst structures based on spectra.[253] The quantitative relationship between the adsorption state (including catalytic properties and CO structure information) and the spectral features of CO, a key intermediate molecule in the carbon dioxide reduction reaction, was investigated using a single-atom catalyst (SAC) with metal atoms dispersed on a metal oxide carrier as an illustrative example (Fig. 20A). Two spectroscopic-based ML models were constructed: one for predicting the properties within adsorption energy E_{ads} and charge transfer Δq, and the other for inverting the structure of the adsorbed molecule CO. The second model predicts six structural parameters, including bond lengths, bond angles, and dihedral angles, using infrared spectral signals of CO. These parameters enable the precise determination of the position of the small molecule relative to the monatomic catalyst, which in turn uniquely determines the spatial relative coordinates of the CO molecule, allowing for structure inversion. Based on the above two ML models, they developed an artificial intelligence generation workflow for catalytic structure design (Fig. 20B). First, many spectra are randomly generated to predict the E_{ads} rapidly. The predicted E_{ads} values are compared with the desired adsorption energy to identify the spectrum corresponding to the desired property. Then, the structure of the CO molecule was constructed from this spectrum, and the properties were verified through DFT calculations. This work not only quantifies the spectrum–structure–property relationship but also paves the way for catalytic tailoring *via in situ* ultrafast spectroscopy.

For retrosynthesis planning problems, Zhang *et al.* explored the potential of embedding spectral descriptors in graph neural networks to assist in reaction path generation.[254] Five ML models were designed and combined with a Monte Carlo tree search to construct a reaction path inverse planning algorithm. This algorithm, for any small organic molecule, can provide the synthesis steps from a commercially available feedstock molecule to the target molecule, the reasonableness scores for each step, and information on the reaction conditions, such as catalysts and solvents. The 460 000 spectroscopic data and 150 000 bond dissociation energy data obtained from the

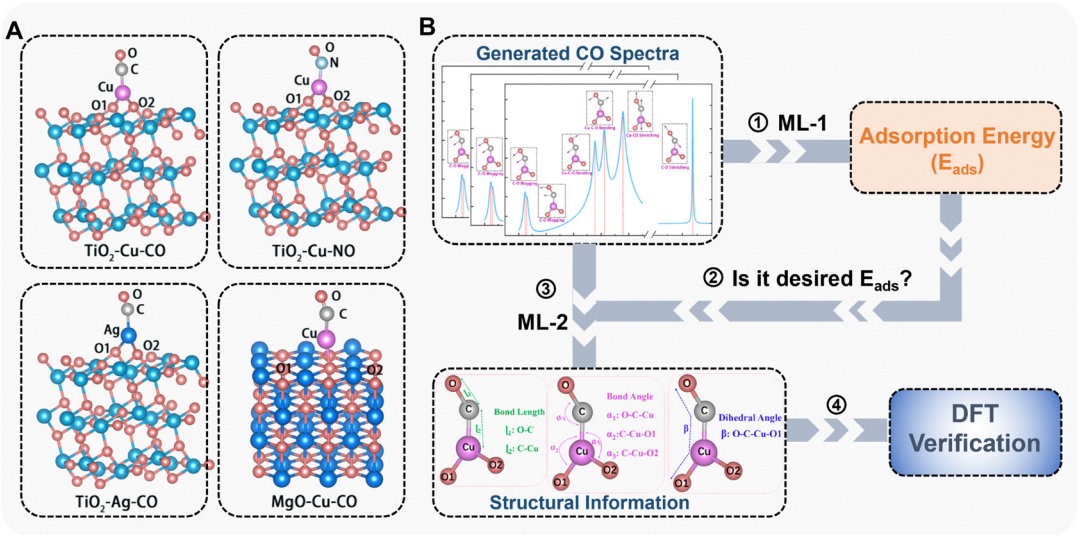


Fig. 20 Example of generating absorption structures on-demand based on the spectrum–structure–property relationship. Reproduced with permission from ref. 253. (A) Typical absorption structures of a CO or NO molecule absorbed on different substrates. (B) Flowchart of designing absorption structure that meets the target adsorption energy requirement.

ab initio calculations were used to introduce information on spectroscopic features, bond dissociation energies, and reaction conditions in chemical reactions. This information was integrated into the molecular graph. A chemistry-informed molecular graph (CIMG) descriptor (Fig. 21A), which integrates spectroscopic and chemical information, was designed and applied to five ML models for inverse reaction prediction, catalyst prediction, solvent prediction, reaction reasonableness

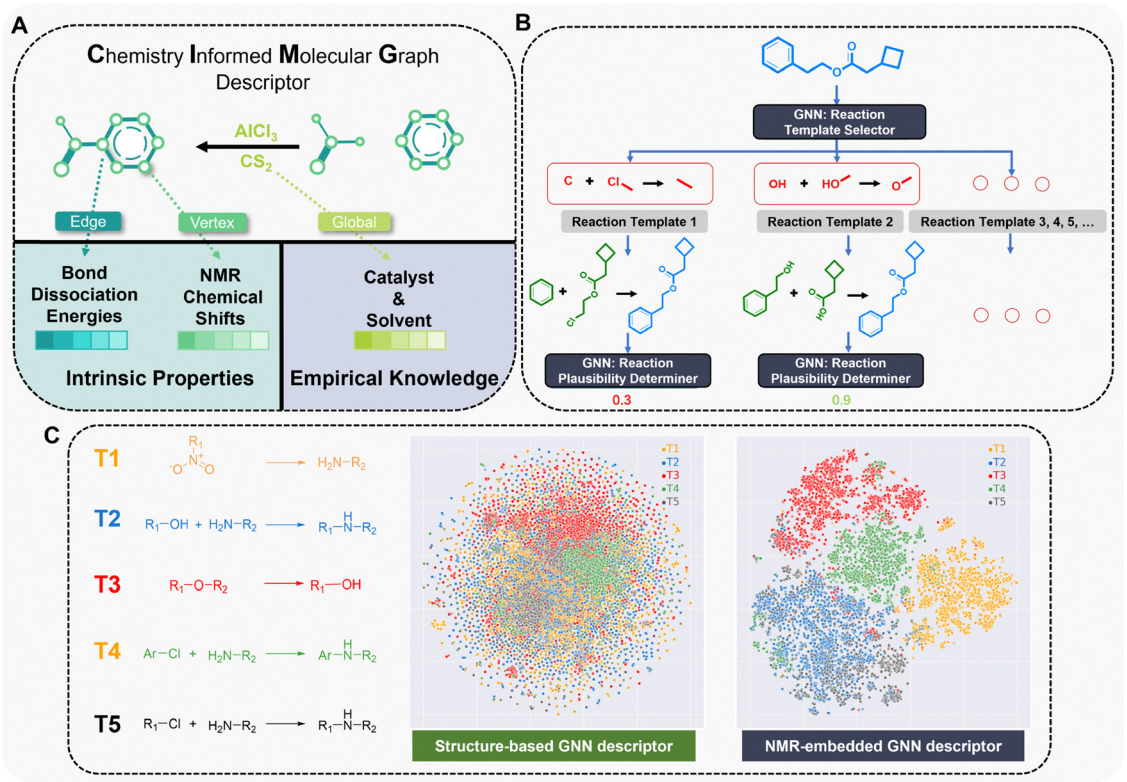


Fig. 21 Example of retrosynthesis planning based on the spectral descriptor. Reproduced with permission from ref. 254. (A) Beyond structural information, the chemistry-informed molecular graph (CIMG) descriptor also embeds NMR chemical shifts, bond dissociation energies, and information on catalyst and solvent. (B) Flowchart of designing a synthetic route for a target molecule. (C) Comparison of the PCA clustering performance based on molecular fingerprints and CIMG for 5 most frequently occurring reaction templates.

judgment, and multi-step path planning, among other things. This approach combines deep chemical knowledge with advanced ML techniques to improve the accuracy and efficiency of inverse synthesis planning. This has led to significant advances in computational chemistry and has the potential to simplify new chemical syntheses. Tests have demonstrated that CIMG descriptors incorporating spectroscopic features can help ML models decipher high-dimensional correlations of conformational relationships and enhance the chemical knowledge of ML models (Fig. 21C).

The UV-vis spectrum of a substance can directly reflect its light absorption characteristics. For this reason, many studies have attempted to use ultraviolet spectra to customise the design of molecules with specific properties. In a study on the relationship between features of the UV-vis absorption spectrum (''photoactivity'') and phototoxicity,[255] ML algorithms are used to predict features of the UV-vis spectra from molecular structures, offering a valuable strategy for reducing experimental tests in phototoxicity assessment. The RF algorithm was trained using 72 787 organic molecules encoded with Morgan circular fingerprints. It successfully classifies molecules based on UV-vis spectral features linked to photoreactivity. On an independent test set comprising 998 molecules, it achieved an accuracy of 0.89, a sensitivity of 0.90 and a specificity of 0.88. These results highlight the potential of ML algorithms for classifying molecules according to the UV-vis absorption spectrum to assist in photosafety evaluation. For designing substance with target absorption properties, Jung *et al.* introduced a methodology that trains two distinct machine learning workflows in parallel to predict the peak wavelength of optical absorption, λ_{max}.[256,257] Traditionally, this is determined through experimental means using UV-vis spectroscopy. One workflow employs deep residual convolutional neural networks to extract rich, multi-level feature representations of dye molecules and solvents, while the other uses a gradient boosting algorithm accompanied by additional procedures to select a minimal yet highly relevant subset of descriptor features from an extensive list, thereby reducing feature redundancy. The potential applications of the ML models include corroborating experimental measurements of λ_{max} values, serving as alternatives to the computationally intensive DFT computations, and identifying new optically active chemicals. In a similar work by Choudhury *et al.* that aims to engineer biomimetic materials with desired spectroscopic properties,[258] *i.e.*, strong absorption at the selected wavelength, ML models are used to identify the key chromophoric components of dihydroxyindole carboxylic acid melanin and their associated 3D structures, which influence absorption across different wavelength ranges. Combining ML with TDDFT, this approach predicts the structure of the most efficient absorbers in each UV-vis wavelength range. As a result, their inverse design approach interprets the contributions of dominant chromophoric units in various wavelength domains of the melanin spectrum.

For fluorescence spectroscopy, researchers have constructed a closed-loop research method of ''generation-validation-feedback'' based on generative models, thus realizing the *de novo* generation of fluorescent molecules.[259] In the work, Sumita *et al.* constructed a model called ChemTS, which can utilize the RNN submodel to generate SMILES, and then screen the validity and fluorescence properties of the molecules by using RDkit and DFT calculations, and feedback the results to the generative model. Through iterative optimisation, six out of eight molecules designed using this scheme were successfully validated through experimental means, and the obtained molecules could produce fluorescence visible to the naked eye.

7. From spectra to multi-modal models

Building on the insights gained from mapping spectra to structures, properties, and inverse design, we now explore how integrating multiple spectral modalities can synergistically enhance our understanding of chemical systems. Analysing the spectral characteristics of chemical matter has enabled scientists to gain significant insights into molecular structures, the types of chemical bonds, the laws governing interactions, and the mechanisms underlying chemical reactions. Spectroscopic descriptors encompass not only the numerous dimensions of chemical composition but also the intricate interactions between chemical compounds and the environment. This provides sufficient information to predict the structure and properties of matter. In light of the aforementioned considerations, there should be a robust correlation between different types of spectra of the same chemical compound. The application of ML techniques enables the utilisation of one spectrum to compensate for the deficiencies of other spectra. Moreover, recently developed large models have shown great ability in multi-modal transition, and this technique has also been used to advance the cross-modal prediction of structures, spectra, and functional properties.

A study of galaxy spectra has demonstrated that spectral information in different frequency bands is correlated, thus enabling the completion of missing wavelength ranges.[260] In astronomical observation, mid-infrared spectra contain a wealth of information regarding compounds; however, obtaining mid-infrared spectra of most galaxies is challenging. The present study endeavours to reconstruct mid-infrared spectra within the 5–35 micrometre range by means of restricted multiband photometry of approximately 20 bands ranging from ultraviolet to submillimetre waves, as available in the Galaxy Survey. By constructing simulated spectra of 10 000 galaxies, they trained a model using the Generative Latent Optimization (GLO) framework and achieved mid-infrared spectral reconstruction with a 60% success rate. As the spectral bands in this study span four orders of magnitude, the sources of the different frequency spectra vary. However, intrinsic, high-dimensional correlations still provide an important basis for multimodal transformations between different types of spectra.

In a recent endeavour, Yang *et al.* initiated preliminary explorations in the field of IR and Raman spectroscopic studies

of small molecule adsorption systems.[261] The low-frequency region of IR and Raman spectra, which is below 1000 cm^{-1}, is often referred to as the "fingerprint" region. Bands located in this region are highly sensitive to structural changes due to rotations and vibrations. They are also susceptible to distortion from environmental factors such as instrumental Gaussian noise, false peaks caused by sample contamination, and signal redshift due to environmental variations. In contrast, high-frequency signals at frequencies higher than 1000 cm^{-1} are more readily measurable due to their greater specificity and reduced likelihood of significant peak overlapping and disruption by noise. Their study attempted to use high-frequency data to improve the quality of the low-frequency spectra (Fig. 22A and B). The ML model used in their work, a combination of the U-net network with attention, demonstrated the ability to extract features from noisy low-frequency signals, which are commonly encountered in IR and Raman spectroscopic measurements. To illustrate the efficacy of this approach, they applied it to the low-frequency vibrational spectra of *trans*-1,2-bis(4-pyridyl)ethylene (BPE) adsorbed on an Ag surface. The model established a high-dimensional correlation between high-quality data in high-frequency modes and low-quality data in low-frequency modes. It refined the fingerprint by aligning poor signals with good ones, demonstrates that the spectrum contains ample information.

With the technology of large models, they have also explored the cross-modal prediction of spectral and structural descriptors.[262] Based on the BERT model framework, they achieved intermodal fusion based on pre-training for modes such as intra-structural coordinates, IR spectroscopy, and Raman spectroscopy for the CO/NO adsorption system by complementing the occlusion data (Fig. 22C–E). BERT-based models excel at cross-modal spectroscopic predictions because their bidirectional encoding architecture enables the learning of contextual relationships in all directions across different spectral types. Additionally, their masked learning approach naturally suits the task of predicting missing information in one modality from available information in others. The model can use the other two modes to predict the other missing mode, resulting in accurate cross-modal prediction of IR spectra, Raman spectra, and internal coordinates. This result demonstrates the feasibility of cross-modal prediction between different chemical properties using a large model, in particular, the possibility of constructing inter-modal spectral and structural correlations to predict other properties of matter.

Recent work has also shown that the spectra of molecules contain information to predict the key properties that usually can only obtained through structure.[263,264] The molecular assembly index (MA) is designed to evaluate molecular complexity by quantifying the difficulty of constructing a molecule from its building blocks and identifying its shortest construction route. With ML technologies, researchers demonstrated that the MA can be experimentally measured using three independent spectra: NMR, MS, and IR. The MA of an unknown molecule can be reliably estimated by analysing the IR absorbance peaks, the carbon resonance signals from NMR, or the molecular fragments revealed through tandem MS. Furthermore, combining these three spectra together as input provides a more robust prediction. This work indicates that multi-modal models can facilitate a more profound comprehension of the spectrum–structure–property relationship than other ML models. In addition to property prediction, it is also feasible to accurately generate suitable structures from multiple types of spectra or to predict other types of spectra. A large model, constructed based on spectroscopic space, may be capable of suggesting potential spectra and structures in accordance with the desired functional properties, thereby ushering in a new era of on-demand intelligent design.

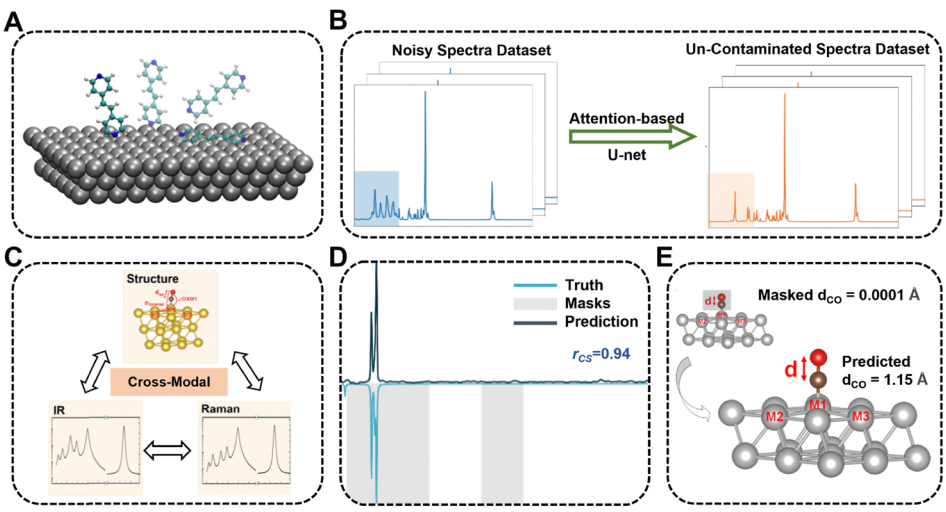


Fig. 22 Example of retrosynthesis planning based on the spectral descriptor. (A) and (B) Are reproduced with permission from ref. 261, and (C)–(E) are reproduced with permission from ref. 262. (A) Different BPE absorption configurations on the Ag surface. (B) Repairing the noisy low-frequency region of vibrational spectra based on high-frequency spectra *via* an attention-based U-net. (C) Cross-modal prediction of absorption structure, IR and Raman spectra. (D) Prediction of masked Raman spectrum using IR and structural information. (E) Prediction of masked structural information using IR and Raman spectra.

8. Model interpretability

In spectroscopy, interpretability represents a critical departure from conventional machine learning applications, as predictive accuracy alone is insufficient—scientists require an in-depth understanding to validate physical insights and trust model outputs. Several complementary approaches have emerged to address this challenge across spectroscopic techniques. Model-agnostic methods such as feature importance analysis and SHAP provide global insights into which spectral regions drive predictions,[179] while CAMs reveal the specific discriminative regions (peaks and peak series) that neural networks use for classification in both X-ray diffraction and absorption spectroscopy.[90,93,179] In X-ray absorption spectroscopy, CAM attributions successfully identify atomic contributions to specific peaks with high agreement with quantum mechanical results, particularly when using graph neural networks with attention mechanisms (GATv2) and global graph context (GraphNet) over simple graph convolutional networks.[179]

Technique-specific adaptations have proven particularly valuable across the spectroscopic landscape. In IR spectroscopy, a two-step approach to explainability reveals how convolutional neural networks learn characteristic group frequencies while also discovering non-intuitive patterns, such as overtones and combination bands.[99] Notably, these networks also learn to use the absence of peaks to distinguish functional groups when characteristic regions are crowded with overlapping bands.[99] In NMR spectroscopy, attention mechanisms effectively highlight chemical environment effects and chemical shifts, with the integration of chemical information making attention matrix results more aligned with established chemical knowledge.[248]

More sophisticated approaches include physics-constrained models and symbolic methods that achieve unprecedented interpretability. The RankAAE model creates explicit one-to-one mappings between latent variables and interpretable structural descriptors, achieving high quantitative correlations and enabling visualization of how individual structure descriptors affect XANES spectra through distinct spectral trends.[167] At the end of interpretability, symbolic regression methods, such as SISSO,[242] generate explicit mathematical formulas that relate vibrational spectral features to catalytic properties using experimentally measurable descriptors with clear chemical meaning.[243] These physics-based formulas demonstrate remarkable robustness, maintaining performance with as few as 20 training samples and handling up to 50% questionable labels due to their alignment with underlying physical principles.[243]

The validation of these interpretability methods relies heavily on agreement with established knowledge, such as quantum chemistry calculations, known structure–spectrum relationships, or expert spectroscopist practices, ensuring that models learn meaningful chemical patterns rather than spurious correlations.[99,179] CAM studies demonstrate that interpretability mechanisms successfully disclose how CNN models make decisions, shedding new light on interpretable deep learning for materials characterization.[90] While challenges remain, particularly regarding feature correlation and the need for multi-scale explanations, these interpretability advances are essential for building trust and enabling scientific discovery in ML-driven spectroscopic analysis.

9. Summary and outlook

In conclusion, this review examines the potential of AI techniques to efficiently address a range of challenges in spectroscopic studies. For conventional problems in spectral analysis and prediction, AI can effectively establish correlations between spectra and structures, facilitating the extraction of crucial structural information and rapidly predicting spectra based on structural data. This markedly enhances research efficiency. Furthermore, AI is capable of establishing correlations between spectra and properties, thereby enabling the direct prediction of functional properties from spectra. By modelling relationships between spectra, structures, and properties, AI offers a novel perspective on spectroscopic research and extends beyond traditional methods. Researchers can design spectra to meet target property requirements and generate appropriate molecular structures, thereby achieving on-demand design and creation of matter. Furthermore, machine learning models can identify intrinsic correlations between different species' spectra and between spectra of various frequencies, thereby complementing unknown spectral information. The utilisation of AI for deeper studies of the ''spectrum–structure–property'' relationship has the potential to expand the scope of spectral-based intelligent design across a broader range of disciplines.

Looking forward, the rapid development of AI techniques is expected to engender further progress in the field of spectroscopic research. Large language models (LLMs), such as ChatGPT-4, have achieved significant success and are increasingly being integrated into scientific research. Their advanced natural language understanding and generation capabilities facilitate data analysis, streamline communication, and support innovative discoveries across diverse fields of science. As discussed in this review, spectroscopy bridges macroscopic properties and microscopic structure of matter, thereby serving as a natural language for the quantum-mechanical microscopic world. Integrating these existing large-scale theoretical simulations, experimental results, and successful AI models gives rise to a new type of AI model, similar to LLMs but focused on matter. This model, referred to as a large spectroscopy model, centres on spectroscopic descriptors. It aims to establish comprehensive connections between various types of spectra, molecular structures, and numerous possible moments. Moreover, based on self-attention algorithms, this model will enable the monitoring of matter evolvement during chemical reactions through spectroscopy. This capability will facilitate the direct design of molecules based on their spectral characteristics, allowing for customised matter with specific functionalities. Additionally, this approach will support the development of new chemical theories grounded in measurable continuous

quantities, create innovative methods for predicting structures within spectroscopic space, and drive the advancement of large scientific models.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (22025304, 22033007, 22303088, and 22393892), the Innovation Program for Quantum Science and Technology (2021ZD0303303), the CAS Project for Young Scientists in Basic Research (YSBR-005), and the Fundamental Research Funds for the Central Universities (WK9990000129). We thank Yang Wang, Qinyu Qiao, Shijie Tao, and all the colleagues who have contributed to this work.

References

1 N. Bohr, *London, Edinburgh Dublin Philos. Mag. J. Sci.*, 1913, **26**, 1–25.
2 N. Bohr, *Nature*, 1913, **92**, 231–232.
3 P. A. M. Dirac, *Proc. R. Soc. London, Ser. A*, 1928, **117**, 610–624.
4 G. B. Arfken, D. F. Griffing, D. C. Kelly and J. Priest, in *International Edition University Physics*, ed. G. B. Arfken, D. F. Griffing, D. C. Kelly and J. Priest, Academic Press, 1984, pp. 818–840.
5 C. R. Baiz, B. Błasiak, J. Bredenbeck, M. Cho, J.-H. Choi, S. A. Corcelli, A. G. Dijkstra, C.-J. Feng, S. Garrett-Roe, N.-H. Ge, M. W. D. Hanson-Heine, J. D. Hirst, T. L. C. Jansen, K. Kwac, K. J. Kubarych, C. H. Londergan, H. Maekawa, M. Reppert, S. Saito, S. Roy, J. L. Skinner, G. Stock, J. E. Straub, M. C. Thielges, K. Tominaga, A. Tokmakoff, H. Torii, L. Wang, L. J. Webb and M. T. Zanni, *Chem. Rev.*, 2020, **120**, 7152–7218.
6 J. Kozuch, K. Ataka and J. Heberle, *Nat. Rev. Methods Primers*, 2023, **3**, 1–19.
7 S.-Y. Ding, E.-M. You, Z.-Q. Tian and M. Moskovits, *Chem. Soc. Rev.*, 2017, **46**, 4042–4076.
8 X. Wang, S.-C. Huang, S. Hu, S. Yan and B. Ren, *Nat. Rev. Phys.*, 2020, **2**, 253–271.
9 X. X. Han, R. S. Rodriguez, C. L. Haynes, Y. Ozaki and B. Zhao, *Nat. Rev. Methods Primers*, 2022, **1**, 1–17.
10 C. Höppener, J. Aizpurua, H. Chen, S. Gräfe, A. Jorio, S. Kupfer, Z. Zhang and V. Deckert, *Nat. Rev. Methods Primers*, 2024, **4**, 1–20.
11 M. L. Bols, J. Ma, F. Rammal, D. Plessers, X. Wu, S. Navarro-Jaén, A. J. Heyer, B. F. Sels, E. I. Solomon and R. A. Schoonheydt, *Chem. Rev.*, 2024, **124**, 2352–2418.
12 A. J. Miles, R. W. Janes and B. A. Wallace, *Chem. Soc. Rev.*, 2021, **50**, 8400–8413.
13 A. Romani, C. Clementi, C. Miliani and G. Favaro, *Acc. Chem. Res.*, 2010, **43**, 837–846.
14 B. Reif, S. E. Ashbrook, L. Emsley and M. Hong, *Nat. Rev. Methods Primers*, 2021, **1**, 1–23.
15 K. De Bruycker, A. Welle, S. Hirth, S. J. Blanksby and C. Barner-Kowollik, *Nat. Rev. Chem.*, 2020, **4**, 257–268.
16 G. R. D. Prabhu, E. R. Williams, M. Wilm and P. L. Urban, *Nat. Rev. Methods Primers*, 2023, **3**, 1–22.
17 N. P. Lockyer, S. Aoyagi, J. S. Fletcher, I. S. Gilmore, P. A. W. van der Heide, K. L. Moore, B. J. Tyler and L.-T. Weng, *Nat. Rev. Methods Primers*, 2024, **4**, 1–21.
18 F. de Groot, *Chem. Rev.*, 2001, **101**, 1779–1808.
19 M. Chergui, M. Beye, S. Mukamel, C. Svetina and C. Masciovecchio, *Nat. Rev. Phys.*, 2023, **5**, 578–596.
20 C. T. Chantler, G. Bunker, P. D'Angelo and S. Diaz-Moreno, *Nat. Rev. Methods Primers*, 2024, **4**, 1–25.
21 S. B. Kotsiantis, I. D. Zaharakis and P. E. Pintelas, *Artif. Intell. Rev.*, 2006, **26**, 159–190.
22 Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436–444.
23 M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255–260.
24 J. Westermayr and P. Marquetand, *Chem. Rev.*, 2021, **121**, 9873–9926.
25 W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng and Z. Lin, *Sensors*, 2022, **22**, 9764.
26 J. Yi, E.-M. You, G.-K. Liu and Z.-Q. Tian, *Nat. Nanotechnol.*, 2024, 1–5.
27 R. Han, R. Ketkaew and S. Luber, *J. Phys. Chem. A*, 2022, **126**, 801–812.
28 P. J. Linstrom and W. G. Mallard, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg MD, 20899.
29 V. L. Orkin, V. G. Khamaganov, L. E. Martynova and M. J. Kurylo, *J. Phys. Chem. A*, 2011, **115**, 8656–8668.
30 Gas Phase Core Excitation Data Base, **https://unicorn.chemistry.mcmaster.ca/corex/cedb-title.html**.
31 AIST: Spectral Database for Organic Compounds, SDBS, **https://sdbs.db.aist.go.jp/**.
32 V. Mannam, Y. Zhang, X. Yuan, C. Ravasio and S. S. Howard, *J. Phys. Photonics*, 2020, **2**, 042005.
33 M. Mousavizadegan, A. Firoozbakhtian, M. Hosseini and H. Ju, *TrAC, Trends Anal. Chem.*, 2023, **167**, 117216.
34 P. Li, W. Li, Y. Zhang, P. Zhang, X. Wang, C. Yin and R. Chen, *ACS Mater. Lett.*, 2024, **6**, 1746–1768.
35 S. Han, J. Y. You, M. Eom, S. Ahn, E. Cho and Y. Yoon, *Adv. Photonics Res.*, 2024, **5**, 2300308.
36 W. Cai, C. Ye, F. Ao, Z. Xu and W. Chu, *Water Res.*, 2025, **277**, 123281.
37 C. Cobas, *Magn. Reson. Chem.*, 2020, **58**, 512–519.

38 E. Jonas, S. Kuhn and N. Schlörer, *Magn. Reson. Chem.*, 2022, **60**, 1021–1031.
39 M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos and E. Tuzhilina, *Nat. Rev. Methods Primers*, 2022, **2**, 100.
40 S. Marukatat, *Artif. Intell. Rev.*, 2023, **56**, 5445–5477.
41 P. Contreras and F. Murtagh, *Handbook of Cluster Analysis*, Chapman and Hall/CRC, 2015.
42 X. Ran, Y. Xi, Y. Lu, X. Wang and Z. Lu, *Artif. Intell. Rev.*, 2023, **56**, 8219–8264.
43 M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, *IEEE Intell. Syst. Appl.*, 1998, **13**, 18–28.
44 C. Williams and C. Rasmussen, *Advances in Neural Information Processing Systems*, MIT Press, 1995, vol. 8.
45 V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, *Chem. Rev.*, 2021, **121**, 10073–10141.
46 A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody and S. D. Brown, *J. Chemom.*, 2004, **18**, 275–285.
47 Y. Song and Y. Lu, *Shanghai Arch. Psychiatry*, 2015, **27**, 130–135.
48 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
49 J. H. Friedman, *Ann. Stat.*, 2001, **29**, 1189–1232.
50 J. H. Friedman, *Comput. Stat. Data Anal.*, 2002, **38**, 367–378.
51 T. Chen and C. Guestrin, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794.
52 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, vol. 30.
53 L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018, vol. 31.
54 C. Bentéjac, A. Csörgő and G. Martínez-Muñoz, *Artif. Intell. Rev.*, 2021, **54**, 1937–1967.
55 G. P. Zhang, *IEEE Trans. Syst. Man Cybern. Pt. C (Appl. Rev.)*, 2000, **30**, 451–462.
56 H. Taud and J. F. Mas, in *Geomatic Approaches for Modeling Land Change Scenarios*, ed. M. T. Camacho Olmedo, M. Paegelow, J.-F. Mas and F. Escobar, Springer International Publishing, Cham, 2018, pp. 451–455.
57 K. O'Shea and R. Nash, *arXiv*, 2015, preprint, arXiv:1511.08458, DOI: **10.48550/arXiv.1511.08458**.
58 Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, *IEEE Trans. Neural Network Learn. Syst.*, 2022, **33**, 6999–7019.
59 F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, *IEEE Trans. Neural Network*, 2009, **20**, 61–80.
60 Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu, *IEEE Trans. Neural Network Learn. Syst.*, 2021, **32**, 4–24.
61 D. Bank, N. Koenigstein and R. Giryes, in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, ed. L. Rokach, O. Maimon and E. Shmueli, Springer International Publishing, Cham, 2023, pp. 353–374.
62 G. Jung, S. G. Jung and J. M. Cole, *Chem. Sci.*, 2023, **14**, 3600–3609.
63 X. Wang, S. Jiang, W. Hu, S. Ye, T. Wang, F. Wu, L. Yang, X. Li, G. Zhang, X. Chen, J. Jiang and Y. Luo, *J. Am. Chem. Soc.*, 2022, **144**, 16069–16076.
64 C. M. de Armas-Morejón, L. A. Montero-Cabrera, A. Rubio and J. Jornet-Somoza, *J. Chem. Theory Comput.*, 2023, **19**, 1818–1826.
65 SpectraBase, **https://spectrabase.com/**.
66 Database of Raman spectroscopy, X-ray diffraction and chemistry of minerals, **https://rruff.info/**.
67 S. G. Ramalli, A. J. Miles, R. W. Janes and B. A. Wallace, *J. Mol. Biol.*, 2022, **434**, 167441.
68 S. Kuhn, H. Kolshorn, C. Steinbeck and N. Schlörer, *Magn. Reson. Chem.*, 2024, **62**, 74–83.
69 D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V. W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H. B. Schiöth, R. Greiner and V. Gautam, *Nucleic Acids Res.*, 2022, **50**, D622–D631.
70 H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *J. Mass Spectrom.*, 2010, **45**, 703–714.
71 Critical Assessment of Small Molecule Identification, **https://casmi-contest.org/2022/index.shtml**.s.
72 K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12580–12585.
73 METLIN | Scripps Research, **https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage**.
74 LIPID MAPS, **https://www.lipidmaps.org/**.
75 ICDD Database Search -, **https://www.icdd.com/pdfsearch/**.
76 Crystallography Open Database, **https://www.crystallography.net/cod/**.
77 American Mineralogist Crystal Structure Database, **https://rruff.geo.arizona.edu/AMS/amcsd.php**.
78 A. Y. Lee, C. J. Powell, J. M. Gorham, A. Morey, J. H. J. Scott and R. J. Hanisch, *Data Sci. J.*, 2024, **23**, 45.
79 ChemDraw | Revvity Signals Software, **https://revvitysignals.com/products/research/chemdraw**.
80 Mnova Software Suite – Mestrelab, **https://mestrelab.com/main-product/mnova**.
81 NMR Prediction | 1H, 13C, 15N, 19F, 31P NMR Predictor, **https://www.acdlabs.com/products/spectrus-platform/nmr-predictors/**.
82 J. J. Rehr, J. J. Kas, F. D. Vila, M. P. Prange and K. Jorissen, *Phys. Chem. Chem. Phys.*, 2010, **12**, 5503–5513.

83 The Largest Curated Crystal Structure Database | CCDC, https://www.ccdc.cam.ac.uk/solutions/software/csd/.
84 M. Hellenbrandt, *Crystallogr. Rev.*, 2004, **10**, 17–22.
85 R. Zhang, H. Xie, S. Cai, Y. Hu, G. Liu, W. Hong and Z. Tian, *J. Raman Spectrosc.*, 2020, **51**, 176–186.
86 A. Poppe, J. Griffiths, S. Hu, J. J. Baumberg, M. Osadchy, S. Gibson and B. de Nijs, *J. Phys. Chem. Lett.*, 2023, **14**, 7603–7610.
87 N. J. Szymanski, C. J. Bartel, Y. Zeng, Q. Tu and G. Ceder, *Chem. Mater.*, 2021, **33**, 4204–4215.
88 N. Q. Le, M. Pekala, A. New, E. B. Gienger, C. Chung, T. J. Montalbano, E. A. Pogue, J. Domenico and C. D. Stiles, *J. Phys. Chem. C*, 2023, **127**, 21758–21767.
89 N. J. Szymanski, C. J. Bartel, Y. Zeng, M. Diallo, H. Kim and G. Ceder, *npj Comput. Mater.*, 2023, **9**, 31.
90 H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin and J. Lin, *J. Chem. Inf. Model.*, 2020, **60**, 2004–2011.
91 L. Chen, B. Wang, W. Zhang, S. Zheng, Z. Chen, M. Zhang, C. Dong, F. Pan and S. Li, *J. Am. Chem. Soc.*, 2024, **146**, 8098–8109.
92 P. M. Vecsei, K. Choo, J. Chang and T. Neupert, *Phys. Rev. B*, 2019, **99**, 245120.
93 F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne and T. Buonassisi, *npj Comput. Mater.*, 2019, **5**, 60.
94 Y. Wei, R. S. Varanasi, T. Schwarz, L. Gomell, H. Zhao, D. J. Larson, B. Sun, G. Liu, H. Chen, D. Raabe and B. Gault, *Patterns*, 2021, **2**, 100192.
95 J. L. Lansford and D. G. Vlachos, *Nat. Commun.*, 2020, **11**, 1513.
96 Z. Huang, M. S. Chen, C. P. Woroch, T. E. Markland and M. W. Kanan, *Chem. Sci.*, 2021, **12**, 15329–15338.
97 J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh and K.-S. Sohn, *Nat. Commun.*, 2020, **11**, 86.
98 H. Ren, H. Li, Q. Zhang, L. Liang, W. Guo, F. Huang, Y. Luo and J. Jiang, *Fundam. Res.*, 2021, **1**, 488–494.
99 L. H. Rieger, M. Wilson, T. Vegge and E. Flores, *Digital Discovery*, 2023, **2**, 1957–1968.
100 Z. Ren, Z. Zhang, J. Wei, B. Dong and C. Lee, *Nat. Commun.*, 2022, **13**, 3859.
101 M. M. Bajomo, Y. Ju, J. Zhou, S. Elefterescu, C. Farr, Y. Zhao, O. Neumann, P. Nordlander, A. Patel and N. J. Halas, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2211406119.
102 O. Usoltsev, A. Tereshchenko, A. Skorynina, E. Kozyr, A. Soldatov, O. Safonova, A. H. Clark, D. Ferri, M. Nachtegaal and A. Bugaev, *Small Methods*, 2024, **8**, 2301397.
103 J. A. Fine, A. A. Rajasekar, K. P. Jethava and G. Chopra, *Chem. Sci.*, 2020, **11**, 4618–4630.
104 M. Alberts, T. Laino and A. C. Vaucher, *ChemRxiv*, 2023, preprint, DOI: 10.26434/chemrxiv-2023-5v27f.
105 G. C. Kanakala, B. Sridharan and U. D. Priyakumar, *Digital Discovery*, 2024, **3**, 2417–2423.
106 X. Lu, H. Ma, H. Li, J. Li, T. Zhu, G. Liu and B. Ren, *arXiv*, 2025, preprint, arXiv:2503.07014, DOI: 10.48550/arXiv.2503.07014.
107 S. Luo, W. Wang, Z. Zhou, Y. Xie, B. Ren, G. Liu and Z. Tian, *Anal. Chem.*, 2022, **94**, 10151–10158.
108 Y. Ju, O. Neumann, M. Bajomo, Y. Zhao, P. Nordlander, N. J. Halas and A. Patel, *ACS Nano*, 2023, **17**, 21251–21261.
109 S. Park, J. Lee, S. Khan, A. Wahab and M. Kim, *Sensors*, 2022, **22**, 596.
110 B. Li, M. N. Schmidt and T. S. Alstrøm, *Analyst*, 2022, **147**, 2238–2246.
111 W. Zhou, Y. Tang, Z. Qian, J. Wang and H. Guo, *RSC Adv.*, 2022, **12**, 5053–5061.
112 J. C. Martinez, J. R. Guzmán-Sepúlveda, G. R. Bolaños Evia, T. Córdova and R. Guzmán-Cabrera, *Int. J. Thermophys.*, 2018, **39**, 79.
113 L.-W. Shang, Y.-L. Bao, J.-L. Tang, D.-Y. Ma, J.-J. Fu, Y. Zhao, X. Wang and J.-H. Yin, *J. Raman Spectrosc.*, 2022, **53**, 237–246.
114 G. R. Koch, R. Zemel and R. Salakhutdinov, *ICML deep learning workshop*, 2015, vol. 2.
115 H. He, M. Cao, Y. Gao, P. Zheng, S. Yan, J.-H. Zhong, L. Wang, D. Jin and B. Ren, *Nat. Commun.*, 2024, **15**, 754.
116 T. Chen, J. Li, P. Cai, Q. Yao, Z. Ren, Y. Zhu, S. Khan, J. Xie and X. Wang, *Nano Res.*, 2023, **16**, 4188–4196.
117 A. C. Doner, H. A. Moran, A. R. Webb, M. G. Christianson, A. L. Koritzke, N. S. Dewey, S. W. Hartness and B. Rotavera, *J. Quant. Spectrosc. Radiat. Transfer*, 2023, **297**, 108438.
118 J. Zhang, S. Ye, K. Zhong, Y. Zhang, Y. Chong, L. Zhao, H. Zhou, S. Guo, G. Zhang, B. Jiang, S. Mukamel and J. Jiang, *J. Phys. Chem. B*, 2021, **125**, 6171–6178.
119 H. Ren, Q. Zhang, Z. Wang, G. Zhang, H. Liu, W. Guo, S. Mukamel and J. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2202713119.
120 T. Vermeyen, J. Brence, R. V. Echelpoel, R. Aerts, G. Acke, P. Bultinck and W. Herrebout, *Phys. Chem. Chem. Phys.*, 2021, **23**, 19781–19789.
121 J. M. Batista and V. P. Nicu, *Phys. Chem. Chem. Phys.*, 2023, **25**, 22111–22116.
122 A. Micsonai, É. Moussong, N. Murvai, Á. Tantos, O. Tőke, M. Réfrégiers, F. Wien and J. Kardos, *Front. Mol. Biosci.*, 2022, **9**, 863141.
123 C. Tian, Y. Lee, Y. Song, M. R. Elmasry, M. Yoon, D.-H. Kim and S.-Y. Cho, *ACS Appl. Nano Mater.*, 2024, **7**, 5576–5586.
124 M. C. R. Remolina, Z. Li and N. M. Peleato, *J. Hazard. Mater.*, 2022, **430**, 128491.
125 C. Post, S. Brülisauer, K. Waldschläger, W. Hug, L. Grüneis, N. Heyden, S. Schmor, A. Förderer, R. Reid, M. Reid, R. Bhartia, Q. Nguyen, H. Schüttrumpf and F. Amann, *Sensors*, 2021, **21**, 3911.
126 Z. Xu, K. Wang, M. Zhang, T. Wang, X. Du, Z. Gao, S. Hu, X. Ren and H. Feng, *Sens. Actuators, B*, 2022, **359**, 131590.
127 S. Chen, X. Du, W. Zhao, P. Guo, H. Chen, Y. Jiang and H. Wu, *Spectrochim. Acta, Part A*, 2022, **279**, 121418.
128 M. Xie, L. Xie, Y. Li and B. Han, *Spectrochim. Acta, Part A*, 2023, **302**, 123059.
129 S. Ji, S. Hao, J. Yuan and H. Xuan, *Spectrochim. Acta, Part A*, 2025, **327**, 125418.

130 F. Förste, L. Bauer, Y. Wagener, F. Hilgerdenaar, F. Möller, B. Kanngießer and I. Mantouvalou, *Anal. Chem.*, 2025, **97**, 7177–7185.
131 B. Kagan, A. Hendler-Neumark, V. Wulf, D. Kamber, R. Ehrlich and G. Bisker, *Adv. Photonics Res.*, 2022, **3**, 2200244.
132 H. Sha, H. Li, Y. Zhang and S. Hou, *APL Photonics*, 2023, **8**, 096102.
133 E. A. Engel, A. Anelli, A. Hofstetter, F. Paruzzo, L. Emsley and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2019, **21**, 23385–23400.
134 C. J. Pickard and F. Mauri, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2001, **63**, 245101.
135 J. R. Yates, C. J. Pickard and F. Mauri, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2007, **76**, 024401.
136 R. Reher, H. W. Kim, C. Zhang, H. H. Mao, M. Wang, L.-F. Nothias, A. M. Caraballo-Rodriguez, E. Glukhov, B. Teke, T. Leao, K. L. Alexander, B. M. Duggan, E. L. Van Everbroeck, P. C. Dorrestein, G. W. Cottrell and W. H. Gerwick, *J. Am. Chem. Soc.*, 2020, **142**, 4114–4120.
137 F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally and K. Keutzer, *arXiv*, 2016, preprint, arXiv:1602.07360, DOI: **10.48550/arXiv.1602.07360**.
138 J. A. Van Santen, G. Jacob, A. L. Singh, V. Aniebok, M. J. Balunas, D. Bunsko, F. C. Neto, L. Castaño-Espriu, C. Chang, T. N. Clark, J. L. Cleary Little, D. A. Delgadillo, P. C. Dorrestein, K. R. Duncan, J. M. Egan, M. M. Galey, F. P. J. Haeckl, A. Hua, A. H. Hughes, D. Iskakova, A. Khadilkar, J.-H. Lee, S. Lee, N. LeGrow, D. Y. Liu, J. M. Macho, C. S. McCaughey, M. H. Medema, R. P. Neupane, T. J. O'Donnell, J. S. Paula, L. M. Sanchez, A. F. Shaikh, S. Soldatou, B. R. Terlouw, T. A. Tran, M. Valentine, J. J. J. Van Der Hooft, D. A. Vo, M. Wang, D. Wilson, K. E. Zink and R. G. Linington, *ACS Cent. Sci.*, 2019, **5**, 1824–1833.
139 X. Zeng, P. Zhang, W. He, C. Qin, S. Chen, L. Tao, Y. Wang, Y. Tan, D. Gao, B. Wang, Z. Chen, W. Chen, Y. Y. Jiang and Y. Z. Chen, *Nucleic Acids Res.*, 2018, **46**, D1217–D1222.
140 B. Sridharan, S. Mehta, Y. Pathak and U. D. Priyakumar, *J. Phys. Chem. Lett.*, 2022, **13**, 4924–4933.
141 M. Hohenner, S. Wachsmuth and G. Sagerer, *Knowl.-Based Syst.*, 2005, **18**, 207–215.
142 H. Sun, X. Xue, X. Liu, H.-Y. Hu, Y. Deng and X. Wang, *Anal. Chem.*, 2024, **96**, 5763–5770.
143 X. Kong, L. Zhou, Z. Li, Z. Yang, B. Qiu, X. Wu, F. Shi and J. Du, *npj Quantum Inf.*, 2020, **6**, 79.
144 W. Tao, W. Yu, X. Zou and W. Chen, *J. Magn. Reson.*, 2023, **353**, 107492.
145 N. Schmid, S. Bruderer, F. Paruzzo, G. Fischetti, G. Toscano, D. Graf, M. Fey, A. Henrici, V. Ziebart, B. Heitmann, H. Grabner, J. D. Wegner, R. K. O. Sigel and D. Wilhelm, *J. Magn. Reson.*, 2023, **347**, 107357.
146 V. S. Manu, C. Olivieri and G. Veglia, *Nat. Commun.*, 2023, **14**, 4144.
147 M. Ludwig, K. Dührkop and S. Böcker, *Bioinformatics*, 2018, **34**, i333–i340.
148 S. Gao, H. Y. K. Chau, K. Wang, H. Ao, R. S. Varghese and H. W. Ressom, *Metabolites*, 2022, **12**, 605.
149 H. Ji, Y. Xu, H. Lu and Z. Zhang, *Anal. Chem.*, 2019, **91**, 5629–5637.
150 P. Hong, H. Sun, L. Sha, Y. Pu, K. Khatri, X. Yu, Y. Tang and C. Lin, *J. Am. Soc. Mass Spectrom.*, 2017, **28**, 2288–2301.
151 Z. Chen, J. Wei, Y. Tang, C. Lin, C. E. Costello and P. Hong, *J. Am. Soc. Mass Spectrom.*, 2022, **33**, 436–445.
152 J. Wei, D. Papanastasiou, M. Kosmopoulou, A. Smyrnakis, P. Hong, N. Tursumamat, J. A. Klein, C. Xia, Y. Tang, J. Zaia, C. E. Costello and C. Lin, *Chem. Sci.*, 2023, **14**, 6695–6704.
153 F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers and J. J. J. van der Hooft, *PLoS Comput. Biol.*, 2021, **17**, e1008724.
154 K. Dührkop, L.-F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein and S. Böcker, *Nat. Biotechnol.*, 2021, **39**, 462–471.
155 F. Wang, D. Allen, S. Tian, E. Oler, V. Gautam, R. Greiner, T. O. Metz and D. S. Wishart, *Nucleic Acids Res.*, 2022, **50**, W165–W174.
156 F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner and D. S. Wishart, *Anal. Chem.*, 2021, **93**, 11692–11700.
157 Y. Orlova, A. A. Gambardella, I. Kryven, K. Keune and P. D. Iedema, *J. Chem. Inf. Model.*, 2021, **61**, 1457–1469.
158 M. A. Skinnider, F. Wang, D. Pasin, R. Greiner, L. J. Foster, P. W. Dalsgaard and D. S. Wishart, *Nat. Mach. Intell.*, 2021, **3**, 973–984.
159 E. E. Litsa, V. Chenthamarakshan, P. Das and L. E. Kavraki, *Commun. Chem.*, 2023, **6**, 132.
160 M. A. Stravs, K. Dührkop, S. Böcker and N. Zamboni, *Nat. Methods*, 2022, **19**, 865–870.
161 M. Bohde, M. Manjrekar, R. Wang, S. Ji and C. W. Coley, *arXiv*, 2025, preprint, arXiv:2502.09571, DOI: **10.48550/arXiv.2502.09571**.
162 S. R. Chitturi, D. Ratner, R. C. Walroth, V. Thampy, E. J. Reed, M. Dunne, C. J. Tassone and K. H. Stone, *J. Appl. Crystallogr.*, 2021, **54**, 1799–1810.
163 Z. Feng, Q. Hou, Y. Zheng, W. Ren, J.-Y. Ge, T. Li, C. Cheng, W. Lu, S. Cao, J. Zhang and T. Zhang, *Comput. Mater. Sci.*, 2019, **156**, 310–314.
164 P. M. Maffettone, L. Banko, P. Cui, Y. Lysogorskiy, M. A. Little, D. Olds, A. Ludwig and A. I. Cooper, *Nat. Comput. Sci.*, 2021, **1**, 290–297.
165 Z. Chen, Y. Xie, Y. Wu, Y. Lin, S. Tomiya and J. Lin, *Digital Discovery*, 2024, **3**, 369–380.
166 J. Timoshenko, D. Lu, Y. Lin and A. I. Frenkel, *J. Phys. Chem. Lett.*, 2017, **8**, 5091–5098.
167 Z. Liang, M. R. Carbone, W. Chen, F. Meng, E. Stavitski, D. Lu, M. S. Hybertsen and X. Qu, *Phys. Rev. Mater.*, 2023, **7**, 053802.
168 J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans and A. I. Frenkel, *Phys. Rev. Lett.*, 2018, **120**, 225502.
169 A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot,

J. A. van Bokhoven, C. Copéret and A. V. Soldatov, *npj Comput. Mater.*, 2021, **7**, 203.
170 A. Martini, S. A. Guda, A. A. Guda, G. Smolentsev, A. Algasov, O. Usoltsev, M. A. Soldatov, A. Bugaev, Y. Rusalev, C. Lamberti and A. V. Soldatov, *Comput. Phys. Commun.*, 2020, **250**, 107064.
171 S. Xiang, P. Huang, J. Li, Y. Liu, N. Marcella, P. K. Routh, G. Li and A. I. Frenkel, *Phys. Chem. Chem. Phys.*, 2022, **24**, 5116–5124.
172 Y. Liu, N. Marcella, J. Timoshenko, A. Halder, B. Yang, L. Kolipaka, M. J. Pellin, S. Seifert, S. Vajda, P. Liu and A. I. Frenkel, *J. Chem. Phys.*, 2019, **151**, 164201.
173 J. Li, B. Mei, H. Liu, X. Li, S. Gu, J. Ma, X. Yu and Z. Jiang, *J. Phys. Chem. C*, 2021, **125**, 18979–18987.
174 M. R. Carbone, P. M. Maffettone, X. Qu, S. Yoo and D. Lu, *J. Phys. Chem. A*, 2024, **128**, 1948–1957.
175 A. Aarva, V. L. Deringer, S. Sainio, T. Laurila and M. A. Caro, *Chem. Mater.*, 2019, **31**, 9243–9255.
176 A. Aarva, V. L. Deringer, S. Sainio, T. Laurila and M. A. Caro, *Chem. Mater.*, 2019, **31**, 9256–9267.
177 S. Ye, K. Zhong, J. Zhang, W. Hu, J. D. Hirst, G. Zhang, S. Mukamel and J. Jiang, *J. Am. Chem. Soc.*, 2020, **142**, 19071–19077.
178 F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, *Nat. Commun.*, 2018, **9**, 4501.
179 A. Kotobi, K. Singh, D. Höche, S. Bari, R. H. Meißner and A. Bande, *J. Am. Chem. Soc.*, 2023, **145**, 22584–22598.
180 M. Gastegger, J. Behler and P. Marquetand, *Chem. Sci.*, 2017, **8**, 6924–6935.
181 Y. Zhang and B. Jiang, *Nat. Commun.*, 2023, **14**, 6424.
182 R. Beckmann, F. Brieuc, C. Schran and D. Marx, *J. Chem. Theory Comput.*, 2022, **18**, 5492–5501.
183 M. Gastegger, K. T. Schütt and K.-R. Müller, *Chem. Sci.*, 2021, **12**, 11473–11483.
184 E. Berger, Z.-P. Lv and H.-P. Komsa, *J. Mater. Chem. C*, 2023, **11**, 1311–1319.
185 Z. Tang, S. T. Bromley and B. Hammer, *J. Chem. Phys.*, 2023, **158**, 224108.
186 A. A. Kananenka, K. Yao, S. A. Corcelli and J. L. Skinner, *J. Chem. Theory Comput.*, 2019, **15**, 6850–6858.
187 P. Schienbein, *J. Chem. Theory Comput.*, 2023, **19**, 705–712.
188 N. Xu, P. Rosander, C. Schäfer, E. Lindgren, N. Österbacka, M. Fang, W. Chen, Y. He, Z. Fan and P. Erhart, *J. Chem. Theory Comput.*, 2024, **20**, 3273–3284.
189 M. Fang, S. Tang, Z. Fan, Y. Shi, N. Xu and Y. He, *J. Phys. Chem. A*, 2024, **128**, 2286–2294.
190 G. S. Na, *Anal. Chem.*, 2024, **96**, 19659–19669.
191 W. Hu, S. Ye, Y. Zhang, T. Li, G. Zhang, Y. Luo, S. Mukamel and J. Jiang, *J. Phys. Chem. Lett.*, 2019, **10**, 6026–6031.
192 S. Ye, K. Zhong, Y. Huang, G. Zhang, C. Sun and J. Jiang, *J. Am. Chem. Soc.*, 2024, **146**, 2663–2672.
193 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
194 S. Ye, W. Hu, X. Li, J. Zhang, K. Zhong, G. Zhang, Y. Luo, S. Mukamel and J. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 11612–11617.
195 S. Ye, G. Zhang and J. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2025879118.
196 L. Zhao, J. Zhang, Y. Zhang, S. Ye, G. Zhang, X. Chen, B. Jiang and J. Jiang, *JACS Au*, 2021, **1**, 2377–2384.
197 R. Ramakrishnan, M. Hartmann, E. Tapavicza and O. A. von Lilienfeld, *J. Chem. Phys.*, 2015, **143**, 084111.
198 J. Westermayr and P. Marquetand, *J. Chem. Phys.*, 2020, **153**, 154112.
199 F. Urbina, K. Batra, K. J. Luebke, J. D. White, D. Matsiev, L. L. Olson, J. P. Malerich, M. A. Z. Hupcey, P. B. Madrid and S. Ekins, *Anal. Chem.*, 2021, **93**, 16076–16085.
200 A. D. McNaughton, R. P. Joshi, C. R. Knutson, A. Fnu, K. J. Luebke, J. P. Malerich, P. B. Madrid and N. Kumar, *J. Chem. Inf. Model.*, 2023, **63**, 1462–1471.
201 J. Fan, C. Qian and S. Zhou, *Research*, 2023, **6**, 0115.
202 T. Vermeyen, A. Cunha, P. Bultinck and W. Herrebout, *Commun. Chem.*, 2023, **6**, 1–9.
203 H. Li, D. Long, L. Yuan, Y. Wang, Y. Tian, X. Wang and F. Mo, *Nat. Comput. Sci.*, 2025, **5**, 234–244.
204 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nat. Mach. Intell.*, 2022, **4**, 127–134.
205 J. F. Joung, M. Han, J. Hwang, M. Jeong, D. H. Choi and S. Park, *JACS Au*, 2021, **1**, 427–438.
206 C.-W. Ju, H. Bai, B. Li and R. Liu, *J. Chem. Inf. Model.*, 2021, **61**, 1053–1065.
207 C.-H. Chen, K. Tanaka and K. Funatsu, *J. Fluoresc.*, 2018, **28**, 695–706.
208 C. Xing, G. Chen, X. Zhu, J. An, J. Bao, X. Wang, X. Zhou, X. Du and X. Xu, *Nano Res.*, 2024, **17**, 1984–1989.
209 S. Kang, Y. Kwon, D. Lee and Y.-S. Choi, *J. Chem. Inf. Model.*, 2020, **60**, 3765–3769.
210 P. Gao, J. Zhang, Q. Peng, J. Zhang and V.-A. Glezakou, *J. Chem. Inf. Model.*, 2020, **60**, 3746–3754.
211 P. A. Unzueta, C. S. Greenwell and G. J. O. Beran, *J. Chem. Theory Comput.*, 2021, **17**, 826–840.
212 J. Fang, L. Hu, J. Dong, H. Li, H. Wang, H. Zhao, Y. Zhang and M. Liu, *Sci. Rep.*, 2021, **11**, 18686.
213 P. Gao, Z. Liu, J. Zhang, J.-A. Wang and G. Henkelman, *Crystals*, 2022, **12**, 1740.
214 J. Li, J. Liang, Z. Wang, A. L. Ptaszek, X. Liu, B. Ganoe, M. Head-Gordon and T. Head-Gordon, *J. Chem. Theory Comput.*, 2024, **20**, 2152–2166.
215 Y. Guan, S. V. S. Sowndarya, L. C. Gallegos, P. C. S. John and R. S. Paton, *Chem. Sci.*, 2021, **12**, 12012–12026.
216 Z. Chen, R. P. Badman, B. L. Foley, R. J. Woods and P. Hong, *J. Data-centric Mach. Learn. Res.*, 2024, **1**, 1–37.
217 S. K. Chandy and K. Raghavachari, *J. Chem. Theory Comput.*, 2023, **19**, 6632–6642.
218 Z. Yang, M. Chakraborty and A. D. White, *Chem. Sci.*, 2021, **12**, 10802–10809.
219 M. Cordova, E. A. Engel, A. Stefaniuk, F. Paruzzo, A. Hofstetter, M. Ceriotti and L. Emsley, *J. Phys. Chem. C*, 2022, **126**, 16710–16720.
220 C. Han, D. Zhang, S. Xia and Y. Zhang, *J. Chem. Theory Comput.*, 2024, **20**, 5250–5258.

221 Y. Guan, S. V. Shree Sowndarya, L. C. Gallegos, P. C. St. John and R. S. Paton, *Chem. Sci.*, 2021, **12**, 12012–12026.
222 B. Han, Y. Liu, S. W. Ginzinger and D. S. Wishart, *J. Biomol. NMR*, 2011, **50**, 43–57.
223 D.-W. Li, A. L. Hansen, C. Yuan, L. Bruschweiler-Li and R. Brüschweiler, *Nat. Commun.*, 2021, **12**, 5229.
224 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
225 J. N. Wei, D. Belanger, R. P. Adams and D. Sculley, *ACS Cent. Sci.*, 2019, **5**, 700–708.
226 R. L. Zhu and E. Jonas, *Anal. Chem.*, 2023, **95**, 2653–2663.
227 Z. Zhou, X. Shen, J. Tu and Z.-J. Zhu, *Anal. Chem.*, 2016, **88**, 11084–11091.
228 M. R. Carbone, M. Topsakal, D. Lu and S. Yoo, *Phys. Rev. Lett.*, 2020, **124**, 156401.
229 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2015, vol. 28.
230 P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li and R. Pascanu, *arXiv*, 2018, preprint, arXiv:1806.01261, DOI: 10.48550/arXiv.1806.01261.
231 S. Brody, U. Alon and E. Yahav, *arXiv*, 2022, preprint, arXiv:2105.14491, DOI: 10.48550/arXiv.2105.14491.
232 M. M. M. Madkhali, C. D. Rankine and T. J. Penfold, *Molecules*, 2020, **25**, 2715.
233 C. D. Rankine, M. M. M. Madkhali and T. J. Penfold, *J. Phys. Chem. A*, 2020, **124**, 4263–4270.
234 M. M. M. Madkhali, C. D. Rankine and T. J. Penfold, *Phys. Chem. Chem. Phys.*, 2021, **23**, 9259–9269.
235 C. D. Rankine and T. J. Penfold, *J. Chem. Phys.*, 2022, **156**, 164102.
236 L. Watson, C. D. Rankine and T. J. Penfold, *Phys. Chem. Chem. Phys.*, 2022, **24**, 9156–9167.
237 T. J. Penfold and C. D. Rankine, *Mol. Phys.*, 2023, **121**, e2123406.
238 Q. Sun, Y. Xiang, Y. Liu, L. Xu, T. Leng, Y. Ye, A. Fortunelli, W. A. I. Goddard and T. Cheng, *J. Phys. Chem. Lett.*, 2022, **13**, 8047–8054.
239 Y. Zhang, S. Ye, J. Zhang, C. Hu, J. Jiang and B. Jiang, *J. Phys. Chem. B*, 2020, **124**, 7284–7290.
240 Z. Zou, Y. Zhang, L. Liang, M. Wei, J. Leng, J. Jiang, Y. Luo and W. Hu, *Nat. Comput. Sci.*, 2023, **3**, 957–964.
241 J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr. and A. Tkatchenko, *Sci. Data*, 2021, **8**, 43.
242 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 083802.
243 Y. Chong, Y. Huo, S. Jiang, X. Wang, B. Zhang, T. Liu, X. Chen, T. Han, P. E. S. Smith, S. Wang and J. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2220789120.
244 S. Jiang, X. Wang, Y. Chong, Y. Huang, W. Hu, P. E. S. Smith, J. Jiang and S. Feng, *J. Phys. Chem. Lett.*, 2024, **15**, 2400–2404.
245 Y. Wang, Y. Huang, X. Wang and J. Jiang, *J. Phys. Chem. Lett.*, 2024, 6654–6661.
246 W. Du, F. Ma, B. Zhang, J. Zhang, D. Wu, E. Sharman, J. Jiang and Y. Wang, *J. Am. Chem. Soc.*, 2024, **146**, 811–823.
247 C. Chen, L. Wang, Y. Feng, W. Yao, J. Liu, Z. Jiang, L. Zhao, L. Zhang, J. Jiang and S. Feng, *Chem. Sci.*, 2025, **16**, 6355–6365.
248 J. Zhang, W. Du, X. Yang, D. Wu, J. Li, K. Wang and Y. Wang, *Front. Mol. Biosci.*, 2023, **10**, 1216765.
249 S. Guo, J. Jiang, H. Ren and S. Wang, *J. Phys. Chem. Lett.*, 2023, **14**, 7461–7468.
250 N. Andrejevic, J. Andrejevic, B. A. Bernevig, N. Regnault, F. Han, G. Fabbris, T. Nguyen, N. C. Drucker, C. H. Rycroft and M. Li, *Adv. Mater.*, 2022, **34**, 2204113.
251 K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong and K. A. Persson, *Sci. Data*, 2018, **5**, 180151.
252 Y. Zhang, X. He, Z. Chen, Q. Bai, A. M. Nolan, C. A. Roberts, D. Banerjee, T. Matsunaga, Y. Mo and C. Ling, *Nat. Commun.*, 2019, **10**, 5260.
253 T. Yang, D. Zhou, S. Ye, X. Li, H. Li, Y. Feng, Z. Jiang, L. Yang, K. Ye, Y. Shen, S. Jiang, S. Feng, G. Zhang, Y. Huang, S. Wang and J. Jiang, *J. Am. Chem. Soc.*, 2023, **145**, 26817–26823.
254 B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang and Y. Luo, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2212711119.
255 R. Mamede, F. Pereira and J. Aires-de-Sousa, *Sci. Rep.*, 2021, **11**, 23720.
256 S. G. Jung, G. Jung and J. M. Cole, *J. Chem. Inf. Model.*, 2024, **64**, 1486–1501.
257 J. Shao, Y. Liu, J. Yan, Z.-Y. Yan, Y. Wu, Z. Ru, J.-Y. Liao, X. Miao and L. Qian, *J. Chem. Inf. Model.*, 2022, **62**, 1368–1375.
258 A. Choudhury, R. Ramakrishnan and D. Ghosh, *Chem. Commun.*, 2024, **60**, 2613–2616.
259 M. Sumita, K. Terayama, N. Suzuki, S. Ishihara, R. Tamura, M. K. Chahal, D. T. Payne, K. Yoshizoe and K. Tsuda, *Sci. Adv.*, 2022, **8**, eabj3906.
260 A. Rissaki, O. Pavlou, D. Fotakis, V. Papadopoulou Lesta and A. Efstathiou, *Astron. Comput.*, 2024, **47**, 100823.
261 G. Yang, H. Xiao, H. Gao, B. Zhang, W. Hu, C. Chen, Q. Qiao, G. Zhang, S. Feng, D. Liu, Y. Wang, J. Jiang and Y. Luo, *J. Am. Chem. Soc.*, 2024, **146**, 28491–28499.
262 G. Yang, S. Jiang, Y. Luo, S. Wang and J. Jiang, *J. Phys. Chem. Lett.*, 2024, **15**, 8766–8772.
263 A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, S. I. Walker and L. Cronin, *Nature*, 2023, **622**, 321–328.
264 M. Jirasek, A. Sharma, J. R. Bame, S. H. M. Mehr, N. Bell, S. M. Marshall, C. Mathis, A. MacLeod, G. J. T. Cooper, M. Swart, R. Mollfulleda and L. Cronin, *ACS Cent. Sci.*, 2024, **10**, 1054–1064.

