Review
doi.org/10.1002/ansa.202000162
Chemistry Europe
European Chemical Societies Publishing

Received: 30 November 2020
Revised: 18 December 2020
Accepted: 21 December 2020

Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data

Rola Houhou[1,2] | **Thomas Bocklitz**[1,2]

[1] Institute of Physical Chemistry, Friedrich-Schiller-University Jena, Jena, Germany

[2] Department of Photonic Data Science, Member of Leibniz Research Alliance "Leibniz-Health Technologies", Leibniz Institute of Photonic Technologies, Jena, Germany

Correspondence
Thomas Bocklitz, Leibniz Institute of Photonic Technologies, Member of Leibniz Research Alliance "Leibniz-Health Technologies", 07745 Jena, Germany.
Email: thomas.bocklitz@uni-jena.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: CRC1076AquaDiva

Abstract

Artificial intelligence-based methods such as chemometrics, machine learning, and deep learning are promising tools that lead to a clearer and better understanding of data. Only with these tools, data can be used to its full extent, and the gained knowledge on processes, interactions, and characteristics of the sample is maximized. Therefore, scientists are developing data science tools mentioned above to automatically and accurately extract information from data and increase the application possibilities of the respective data in various fields. Accordingly, AI-based techniques were utilized for chemical data since the 1970s and this review paper focuses on the recent trends of chemometrics, machine learning, and deep learning for chemical and spectroscopic data in 2020. In this regard, inverse modeling, preprocessing methods, and data modeling applied to spectra and image data for various measurement techniques are discussed.

KEYWORDS
2D chromatography, atomic force microscope, chemometrics, deep learning, electron microscope, inverse problem, machine learning, mass spectroscopy, nuclear magnetic resonance, preprocessing, vibrational spectroscopy, X-ray spectroscopy

1 | INTRODUCTION

A variety of measurement techniques to explore hidden aspects of a sample and to measure specific characteristics of a sample exists. Each of these measurement techniques exhibits its properties and is employed to measure a particular attribute of the sample, for example, the molecule structure at the atomic levels, the mass of particles or molecules, the isotopic signature of a sample, the absorbance, or the vibrational modes of molecules. However, the generated data from these measurements is often not directly utilized. Instead, chemometrics, machine learning, and deep learning tools are employed to extract underlying information from the data and link the data to specific applications. This review summarizes recent trends in the development of chemometrics, machine learning, and deep learning methods for a set of chemical important measurement methods like nuclear magnetic resonance (NMR), mass spectroscopy (MS), vibrational spectroscopy, X-ray spectroscopy, atomic force microscope (AFM), electron microscope (EM), and two-dimensional (2D) chromatography. Generally, the analysis of the respective data can be divided into two steps: data enhancement, for example, inverse modeling and preprocessing, and data modeling. A workflow showing the different steps involved in the analysis of chemical data is illustrated in Figure 1 and these steps are further discussed below.

The first step, data enhancement, is necessary because artifacts distort the generated data from all discussed measurement techniques. However, the enhancement can be achieved through either an inverse problem or a forward problem. A variety of methods to

doi.org/10.1002/ansa.202000162
Chemistry Europe
European Chemical Societies Publishing

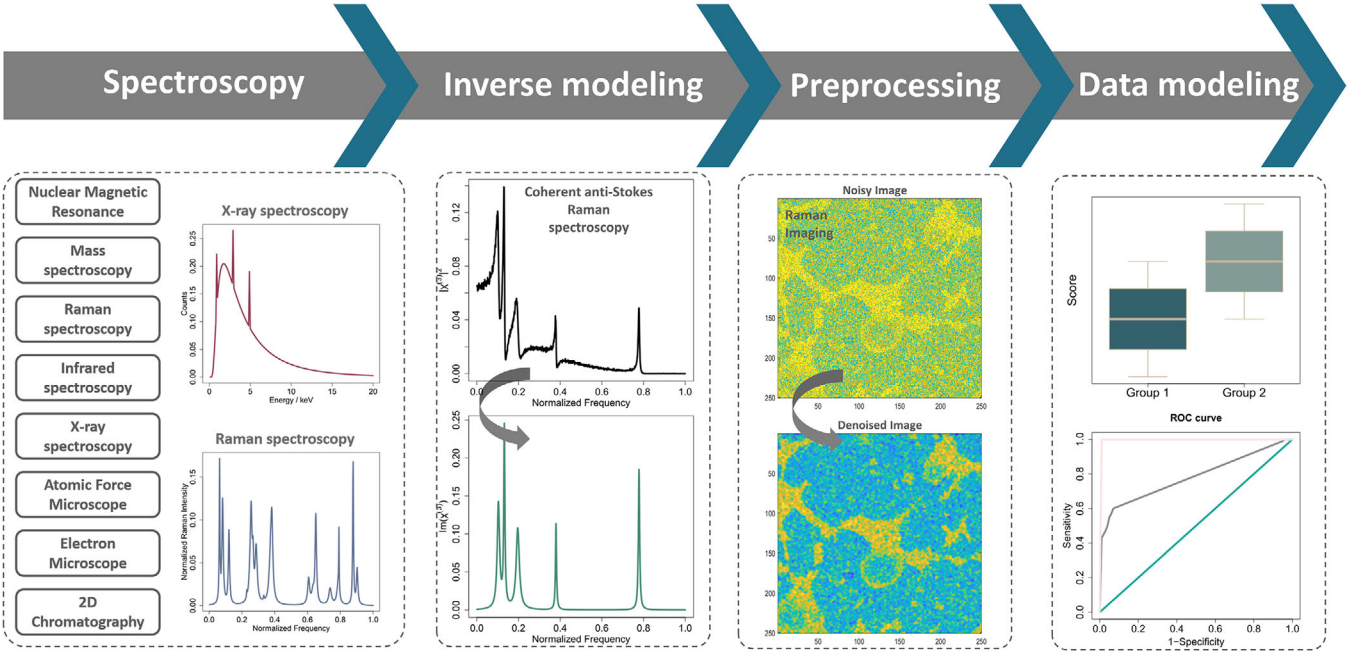

FIGURE 1 General workflow of spectroscopic data analysis from data enhancement to data modeling. Inverse modeling, preprocessing, and data modeling represent key steps in this workflow.

solve or use both problems have been investigated. The preprocessing methods belong to the forward model type methods and are applied to remove different artifacts, for example, denoising,[1,2] baseline correction.[3–5] Moreover, the inverse problem is implemented using the observed data to estimate the original parameters of the chemical or physical system.[6,7] Learning from data via data modeling is the next step after the improvement of data is completed. In this regard, a wide range of mathematical and statistical techniques have been developed to extract important and relevant insights. Feature extraction[8,9] or selection,[10,11] classification,[12,13] and regression[14,15] are considered the essential categories of data modeling. Beside these classical techniques, deep learning gained popularity to solve spectroscopic and chemical applications.[16,17] Furthermore, the importance of chemometric methods increases to deal with the increasing size of spectroscopic datasets. Additionally, chemometric methods are needed to correct artifacts and shortcomings of specific spectroscopic technique. Therefore, the trends in 2020 and the limitations of applying chemometrics, machine learning, and deep learning methods on the aforementioned spectroscopic measurements are reviewed.

This review paper is divided into three main sections. The recent trends in applying chemometrics, machine learning, and deep learning methods to enhance spectroscopic data, including preprocessing methods and inverse modeling, are mentioned in section 1. Data modeling techniques for spectral data, including NMR, MS, vibrational spectroscopy, and X-ray, are covered in section 2. Finally, the advancements of chemometrics and artificial intelligence-based methods applied to imaging data involving AFM, EM, and 2D chromatography are discussed in section 3.

2 | CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING METHODS FOR ENHANCEMENT OF CHEMICAL DATA

In the last years, a growing interest in data analysis methods such as artificial intelligence-based methods can be recognized. For chemical applications, data generation is more beneficial if it is coupled with optimal techniques for the analysis of the generated data. The data analysis methods for the acquired chemical data are sample and task dependent. However, pre-analysis techniques are vital in evaluating chemical data regardless of the task to be solved. While chemical data are retrieved using various measurement techniques, the distortion of these data is very common. These distortions are called artifacts and result from the measurement devices, from the measurement or corrupting processes, and from the nature of the samples itself. The artifact removal or suppression leads to a data enhancement and is crucial for the data analysis to produce meaningful results. This enhancement of the chemical data can be categorized into two main types: the forward and the inverse problems. In the forward problem, the objective is to remove measurement artifacts and errors to determine the composition of the samples and the underlying structures of the chemical information. Concerning this problem, preprocessing techniques are being developed to remove the unwanted variations that limit the extraction of the underlying chemical-relevant structures. Instead, the inverse problem aims to reconstruct the missing information of the chemical/physical system, which was introduced through the measurement process. Recent trends in the application of the forward problems are referred to as preprocessing problems and the inverse problems are shown in the two following subsections.

2.1 | Preprocessing

A detailed examination of preprocessing methods for a given data set is critical as these methods can also remove relevant chemical information. Therefore, the search for the best preprocessing method is vital, considering its impact on the subsequently performed data analysis and its outcome. These preprocessing methods can be employed to either remove noise contributions, replace missing values, interpret or remove baselines, or even a combination of these targets.[18] Depending on the studied problem, either a single preprocessing method or a combination of methods is applied to remove the underlying measurement errors and artifacts.

Nevertheless, retrieving the best preprocessing method or the best combination of techniques to remove all artifacts in the data at hand is challenging. Moreover, the preprocessing choice is the result of an exhaustive and long trial and error process. Consequently, a scientific effort is made to select appropriate preprocessing techniques. Recent papers concentrated on solving the trial and error problem by implementing automated software solutions that compare different preprocessing methods. An open-source python module "nippy" was developed by Torniainen et al[19] for semi-automatic comparisons of preprocessing techniques for near-infrared spectroscopy (NIRS). The presence of noise in the NIRS measurements significantly affects the multivariate methods needed for further analysis. Therefore, preprocessing categories are presented, including clipping, scatter correction, smoothing, derivatives, trimming, and resampling. This python-module was tested using two examples, and it resulted in a fast selection of different preprocessing combinations. The authors recommended following a specific order to apply these methods, which is crucial in NIR spectral data analysis. Further work to search for the best preprocessing strategies was presented in Martyna et al.[20] The authors proposed a novel concept to assess the preprocessing strategy using the ratio of between-groups to within-groups variance. This ratio was calculated on the first latent variable derived from the regularized multivariate analysis of variance (MANOVA). It is used to select the best preprocessing strategy that optimally highlights the differences between groups for highly multidimensional data. In addition, the search for the best preprocessing strategy was carried out using a genetic algorithm. Furthermore, the performance of this novel concept was verified by utilizing two forensic Raman spectral data sets. By assessing both problems in terms of discrimination, the authors successfully point out the sequence in which the preprocessing steps should be performed and extracted their most appropriate parameters. Since applying a preprocessing method might not entirely remove all artifacts, combining multiple methods is promising, and some of the recent publications deal with this combination. Roger et al.[21] developed a new approach to combine several pre-treatment techniques using sequential and orthogonalized partial least squares (SPORT). In their method, the authors applied different pre-treatment techniques to the same NIR transmission spectra including raw data, first and second derivatives, standard normal variate, and variable sorting for normalization. The sequential and orthogonalized partial least squares (SO-PLS) approach is used afterward to combine the resulting blocks of data. The application of SPORT showed good calibration performance in comparison with the existing stacking approach. With the booming of artificial intelligence, researchers utilized deep learning networks for various preprocessing tasks, for example, denoising. Zhang et al[22] tested two deep learning networks via the denoising autoencoder (DAE-1) and the stacked autoencoder (DAE-2) on NIR spectra. The results were compared to other denoising methods, including moving average smoothing, Savitzky-Golay smoothing (SGS), wavelet transform (WT), and empirical mode decomposition (EMD). In this regard, artificial and real NIR spectra were used for the model evaluation. The DAE-1 and DAE-2 applied on both simulated and real NIR spectra showed a better performance than the other methods. An additional study by Raulf et al[23] investigated the removal of disturbing scattering components from an infrared spectrum. The authors proposed deep learning via a convolutional neural network (CNN) as a preprocessing tool to remove the (resonant) Mie scattering. Their results showed that the deep learning approach is faster and can be used for a strong generalization across different tissue types. Their approach also overcame the trade-off between computation time and the bias of the corrected spectrum towards a reference spectrum. Moreover, Wahl et al[24] used a CNN as one single-step preprocessing for Raman spectra. In this paper, the CNN was trained using simulated data to handle three preprocessing steps, for example, cosmic ray removal, smoothing, and baseline subtraction. The preprocessing results were generally of higher quality than what was achieved using reference methods including second-difference, asymmetric least squares, and cross-validation. Additionally, the authors showed reliable results on measured Raman spectra from polyethylene, paraffin, and ethanol with background contamination from polystyrene. In conclusion, the authors proved deep learning as a promising tool for the automated preprocessing of Raman spectra. A drawback of the study was that the tested data basis was limited.

2.2 | Inverse modeling

Inverse modeling is the process of reconstructing missing information from observed measurements to identify its source or the corresponding model parameters. Inverse modeling tries to infer knowledge from given measurement data in the observation space Y to the underlying unknown state X of the sample or to a parameter function in the state space of X. General solutions for this problem do not exist or depends in an unstable way on the measurements, which is related to the ill-posed problem characteristics of inverse modeling.[25,26] A diversity of algorithms exists, and recent developments on this subject are listed below.

Yuan et al.[27] developed a new inverse modeling algorithm of a series of X-ray intensity measurements. Their objective was to recover the structure and composition of two-dimensional (2D) heterogeneous materials measured using X-ray spectroscopy by varying the beam's energy and position. Their method involves an iterative process of forward modeling, based on Monte-Carlo simulation, to determine the optimal structure to minimize the relative differences between the simulated and experimental characteristic X-ray intensities. In

conclusion, the authors proved the feasibility of their approach in analyzing 2D heterogeneous materials for quantitative electron-induced X-ray. However, the input parameters such as beam positions, beam energies, and voxel size must be chosen appropriately. Another study by Hong et al[28] suggested using inverse modeling by combining X-ray computed tomography (XCT) testing and the finite element method to acquire the rust volume reduction coefficient. Conventional numerical models for corrosion expansion ignore the penetration of rust into the concrete matrix along the longitudinal direction. Their approach showed that the obtained reduction coefficient in the rust volume is linked to the rust expansion coefficient. The corrosion level obtained by XCT testing was significantly higher than what they found using the conventional corrosion model. Takeda et al[29] introduced a fundamental methodology of an automated system for material design. The authors built at the modeling stage a regression and a classification model to predict material properties and attributes. At the design stage, the trained predictive model is inverted to change and tune material structures. As a result, the two stages can achieve material design with user-demanded requirements. Also, the authors were able to inverse-design new molecular structures that satisfy the targeted LUMO energies. An inverse model was applied via long short-term memory (LSTM), a recurrent neural network-based, to retrieve the Raman-like shape from broadband coherent anti-Stokes Raman scattering (CARS)[29] by Houhou et al. The authors compared deep learning to other phase retrieval methods (maximum entropy method and Kramers – Kronig relation). The LSTM network outperformed these methods using artificial and experimental broadband CARS data. Additionally, the authors proved the stability of the deep learning method regarding the non-resonant background within CARS spectra. Guo et al[31] implemented a deep learning method in an inverse problem manner to remove artifacts from infrared spectra, which are caused by optical effects. Subsequently, the model could extract the pure absorption of the sample from the infrared measurements. The authors proposed an artifact removal approach based on a 1-dimensional U-Net shaped CNN using Poly (methyl methacrylate) as materials. The pure absorbance was successfully retrieved even when the absorbance is entirely overwhelmed by extensive artifacts. For the same objective, a different deep learning network was implemented by Magnussen et al[32] to recover the pure absorbance from infrared spectra. Initially, the Mie extinction extended multiplicative signal correction (ME-EMSC) algorithm extracts the pure absorbance from highly distorted spectra. Thereafter, the authors trained a deep learning network via the deep convolutional descattering autoencoder (DSAE) on a set of corrected infrared spectra. These corrected spectra were obtained using the ME-EMSC algorithm. Additionally, different reference spectra were used in this study to reflect the large variability in chemical features. In conclusion, the DSAE approach reduced the highly demanding computational time needed in the ME-EMSC algorithm for scatter correction. The DSAE outperformed the ME-EMSC correction in speed, robustness, and noise levels, while preserving the same chemical information in the corrected spectra.

After enhancing the data, the next step is to extract relevant information, which is achieved by applying chemometrics and machine learning techniques to the data, either spectra or images. An overview of chemometrics, machine learning, and deep learning methods is shown in Figure 2. Recent trends in applying chemometrics, machine learning, and deep learning techniques on spectral and image data are shown separately below.

3 | CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING METHODS FOR THE ANALYSIS OF SPECTRAL DATA

Spectroscopic measurements produce high-dimensional profiles containing a high amount of information, which can be optimally exploited using chemometrics, machine learning, and deep learning methods. These methods aim to discover the underlying chemical properties of the sample more precisely and accurately. Recent advances in applying chemometrics, machine learning, and deep learning methods used on spectral data generated by nuclear magnetic resonance (NMR), mass spectroscopy (MS), vibrational spectroscopy, and X-ray spectroscopy are discussed in the following subsections.

3.1 | Nuclear magnetic resonance

Nuclear magnetic resonance (NMR) describes a measurement principle where the nuclei of specific atoms are irradiated by a static magnetic field and then exposed to a second oscillating magnetic field.[35] The analysis of NMR spectra is challenging, and it is not straightforward to draw a conclusion directly from the spectra or even interpret the spectra without the use of chemometric methods. Therefore, combining NMR spectra with chemometrics, machine learning, and deep learning methods is beneficial, and some of the recent applications are discussed below.

The analysis of the 1D ^{1}H NMR spectra of metabolomics samples is challenging since resonances overlap in specific chemical shift regions. However, Pérez et al[36] suggested using a chemometric approach through multivariate curve resolution-alternating least squares (MCR-ALS) to facilitate the steps of metabolites profiling and resonance integration. The authors proved the ability of this method to extract the concentrations and resonances from untargeted metabolites. Their approach was validated using 1D ^{1}H NMR spectra from metabolomic profiling of zebrafish upon acrylamide exposure. Consequently, the authors recommended using their approach to identify spectral features and as biomarker discovery. Another issue for metabolomics NMR-based was presented by Miros et al.[37] in which the growth of *Hypericum perforatum*, or St. John's wort, plants are monitored under different light conditions. The authors developed a toolkit combining 1D ^{1}H NMR spectra with multivariate analysis to extract differences in chemical profiles. As a result, specific metabolites were identified as markers for the difference between the plant growths under different light conditions and glutamine, sucrose, and fructose were found to be chemical markers of light conditions. Another area of interest was related to herbal medicines in which Zhao et al[38]

doi.org/10.1002/ansa.202000162

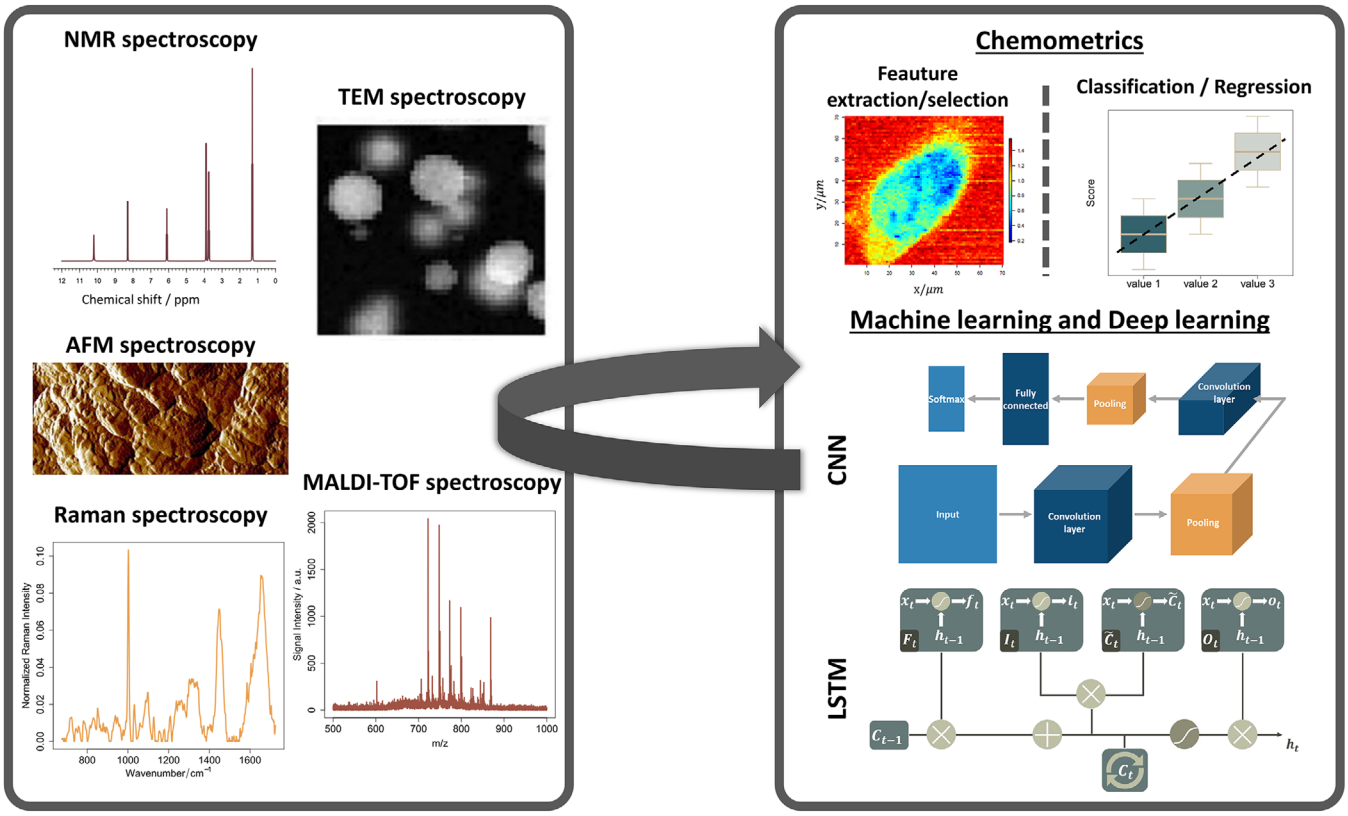

FIGURE 2 An overview of chemometrics, machine learning, and deep learning methods applied to spectroscopic measurements. The AI-based methods like chemometrics, machine learning, and deep learning visualized on the right are utilized to the chemical/spectroscopic data. The data on the left is not representative for all data sources, but we focused in this review on these data types. [LSTM network, AFM, and TEM images are retrieved from references,[30,33,34] respectively]

analyzed the effects of the multistep processing on the chemical changes. The authors applied a chemometric method on ^{1}H NMR spectra to provide comprehensive information on the chemical changes during the processing steps of the Danshen extract. For instance, the hierarchical classification analysis (HCA) clustered the samples according to the processing steps, which indicates that ^{1}H NMR enables the identification of the critical control points based on information of the organic compounds present in the sample. Additionally, a principal component analysis (PCA) and an orthogonal partial least squares discriminant analysis (OPLS-DA) were applied to distinguish the major metabolite differences between the intermediates before and after the critical control point. The combination of ^{1}H NMR and chemometrics proved to be an effective process quality control tool. Therefore, the authors proposed to apply their approach to other herbal medicines to identify critical control points and potential chemical markers. Marion et al[39] developed a new method, adaptive clustering around latent variables (AdaCLV), for simultaneous dimensionality reduction and variable clustering. This method was applied to NMR spectra and can be used for the identification of potential biomarkers. Briefly, AdaCLV filters out variables that do not vary significantly between samples and approximates cluster membership degrees on the remaining variables. The potential overlapping clusters are identified, and the ranking of variable importance within a cluster is achieved. Compared with other clustering methods, AdaCLV estimated latent variables and cluster membership with higher or equivalent precision, and it showed less sensitivity regarding the used hyperparameters. Further analysis was performed by Coimbra et al[40] through merging a time-domain nuclear magnetic resonance (TD-NMR) spectra with chemometric methods to determine the presence of formaldehyde in raw milk samples. PCA, partial least squares (PLS), and soft independent modeling of class analogy (SIMCA) were used to discriminate the samples regarding the level of milk adulteration. SIMCA overcame PCA and PLS with a good discrimination and a high predictive index. Consequently, TD-NMR combined with chemometrics proved its effectiveness for the dairy industry to check the formaldehyde levels in raw milk. Zhang et al[41] combined ^{1}H NMR with chemometric methods to classify the monofloral Chinese honey based on botanical and geographical origins. ^{1}H NMR spectra on samples of 8 classes were collected across China. A PCA could be used to successfully classify their botanic origins while the classification at different geographical levels was effectively distinguished using orthogonal partial least square discriminant analysis (OPLS-DA). This study reported several benefits, including a small sample amount, a simple preparation, a short testing time, and a non-targeted multi-species detection. Recent trends for chemical analysis also include research on the application of deep learning techniques. For instance, Kong et al.[42] proposed a combination of deep learning via a convolutional neural network (CNN) and the sparse matrix completion method to speed up 2D nanoscale NMR spectroscopy, which is vital for molecular structure determination. The use of a CNN successfully suppressed the

observation noise and improved the sensitivity. An additional challenge was presented by Hou et al,[43] where the authors aim to assess the use of deep learning as a tool for rapid and accurate identification of edible oils. Therefore, two-dimensional CNN (2D-CNN) and one-dimensional CNN (1D-CNN) were used to classify different types of edible oils using different low-field nuclear magnetic resonance (LFNMR) spectra. The results showed that transverse relaxation decay signals analyzed by the 1D-CNN presented the best classification ability.

3.2 | Mass spectroscopy

Mass spectroscopy (MS) measures the mass of charge ratio of molecules. It is used to quantify known materials, identify unknown compounds within a sample, and elucidate the structure and chemical properties of different molecules. The fundamental principle involves the fragmentation of a compound or molecule into charged species, which are accelerated, deflected, and finally focused on a detector according to their mass and charge ratio. Ion deflection is based on charge, mass, and velocity, ions separation is based on mass to charge (m/z) ratio, and detection is proportional to the abundance of these ions.[44]

With rapidly growing chemometrics, machine learning, and deep learning methods, MS took its share of the pie. For instance, Duan et al[45] developed a new software, QPMASS, to analyze large-scale gas chromatography-mass spectrometry (GC-MS) data. GC-MS analysis generates many fragment ions for each analytic compound, making the tasks of sample deconvolution and peak alignment very challenging. To deal with this issue, the authors implemented parallel computing with an advanced dynamic programming approach. This approach aligns peaks from multiple samples based on the similarity of each pair calculated using retention time and mass spectra. The diagram of using dynamic programming and the parallel peak alignment in QPMASS is illustrated in Figure 3. As a result, QPMASS enabled fast processing of large-scale datasets and reduced false positive and false negative errors to be less than 5%.

A further challenge for GC-MS was tackled by Alkhalifah et al,[46] in which the search for automated algorithmic clustering methods was discussed. The authors developed VOCCluster, a python-based algorithm that quickly and efficiently analyzes features of deconvolved GC-MS breath data. Compared to a manual volatile organic compound (VOC) panel, the results showed a superior and faster performance with an accurate clustering of 96% of VOCs. Additional trends were introduced by Papagiannopoulou et al.[47] to identify pathogenic bacteria cells in urine samples. First, the authors implemented matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) on individual bacterial cells and identified species rapidly and with acceptable accuracy. In this regard, a deep learning technique via CNN was applied. The results showed similar performance compared to traditional supervised machine learning algorithms, including logistic regression, random forests, and k-nearest neighbor classification. Further issues were tackled by Li et al,[48] who worked to decrease the false positive rate and to improve the low sensitivity arising from a database search engine. This engine is used to identify significant histocompatibility complex (MHC)-binding peptides in mass spectroscopy. The authors developed DeepRescore, a post-processing tool, to improve the sensitivity and reliability in peptide identification. Their approach combines peptide features derived from deep learning predictions with previously used features to rescore peptide-spectrum matches. The results showed that rescoring by DeepRescore on two public immunopeptidomics datasets increases both the sensitivity and reliability of the prediction of MHC-binding peptides. As well, it showed that the deep learning-derived features improved the performance.

3.3 | Vibrational spectroscopy

Vibrational spectroscopy is a non-destructive identification method that measures the vibrational energies of molecular vibrations in the sample. Each chemical bond has a unique vibrational energy, which will be different from one compound to another. This unique energy provides each compound with a unique fingerprint, which is vital in determining compound structures, identifying, and characterizing compounds, and identifying impurities. There are two types of vibrational spectroscopy discussed herein: infrared absorption and Raman spectroscopy. The main difference between these measurement techniques is that in infrared spectroscopy, the absolute frequencies at which a sample absorbs radiation are measured, while Raman spectroscopy measures inelastic scattering in a relative manner. These both are complementary as the vibrations feature different selection rules and both methods are essential to extract the full picture of the vibrational modes in a molecule.[49] Similar to other spectroscopic techniques, vibrational spectroscopy requires advanced data processing to extract meaningful information from spectra. Recent trends about the use of chemometrics, machine learning, and deep learning methods in Raman and infrared spectra are presented below.

Akpolat et al[50] discussed the use of a handheld Raman spectroscopic device with pattern recognition techniques for classification and quantification of different types of tomato carotenoids, which is of great interest for health issues. In this regard, samples with varying carotenoids profiles were non-destructively measured via a handheld Raman spectrometer. The derived spectra were analyzed for classification and quantification purposes using soft independent modeling of class analogy (SIMCA), artificial neural network (ANN), and partial least squares regression (PLSR). A good classification of tomatoes based on their carotenoid profile of 93% and 100% is shown using SIMCA and ANN, respectively. Besides this result, PLSR and ANN were able to achieve a good quantification of all-*trans*-lycopene. Consequently, the authors suggested using their approach as a tool for breeders to provide real-time information on carotenoid profiles. Another study aimed at the quantification of complex mixtures is discussed in Han et al[51] The authors developed a two-stage algorithm based on Bayesian modeling and implemented it on Raman spectra. First, a hierarchical Bayesian model was constructed to learn the peak representation for a target analyte spectrum. A reversible-jump

doi.org/10.1002/ansa.202000162

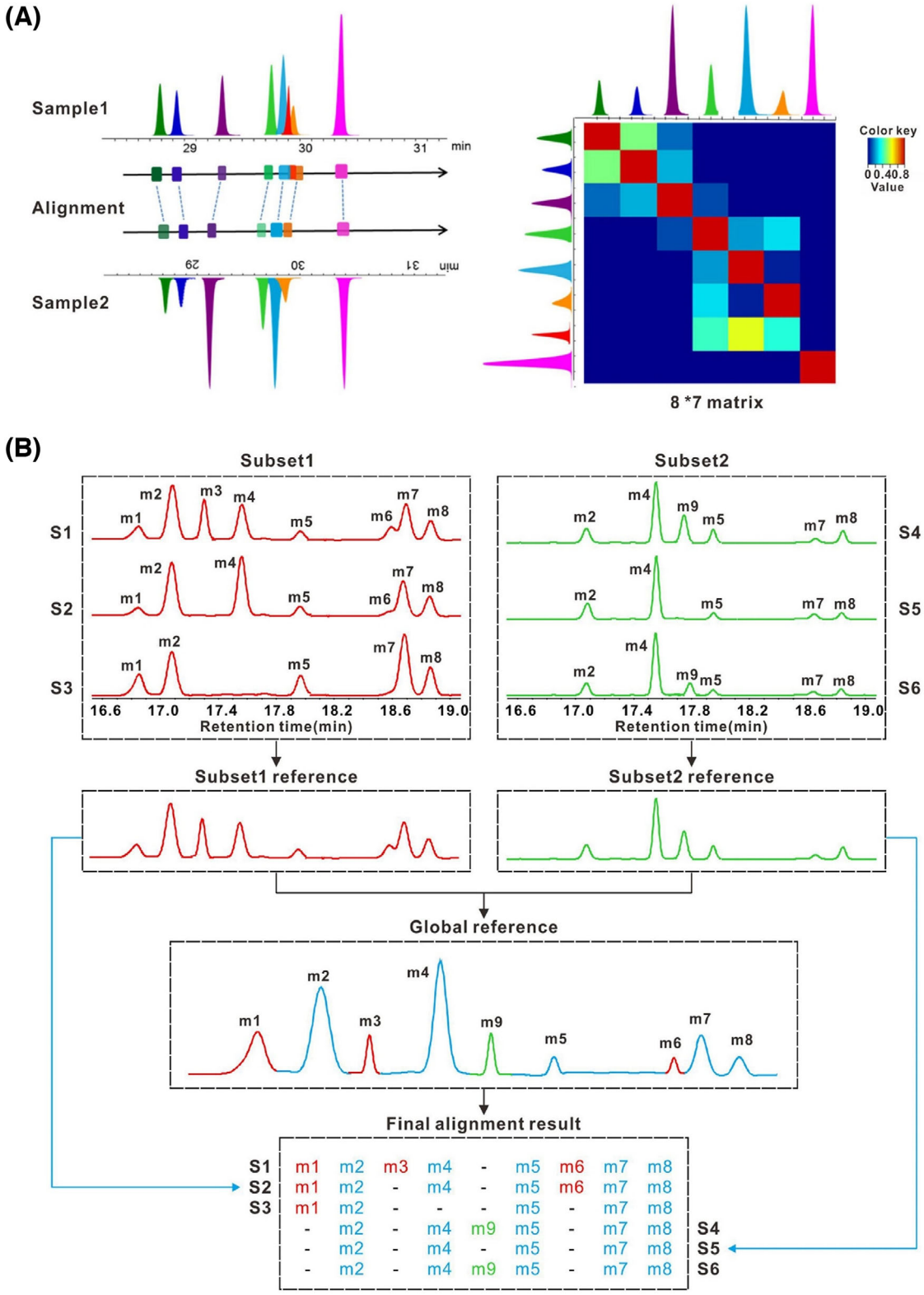

FIGURE 3 Dynamic programming and parallel peak alignment in QPMASS. The diagram in panel (A) describes the dynamic programming for peak alignment. The left panel shows the matched peaks after alignment, and the right panel shows the score matrix for the peak alignment of two samples. In panel (B), the diagram illustrates the parallel peak alignment. The subset references are the consensus sample derived from the average mass spectra and retention time. The final alignment result describes the origin of the detected peaks. [Retrieved and adapted from reference[45]]

Markov chain Monte Carlo (RJMCMC) is then used to estimate the target analyte concentration in a mixture using the peak variables learned in the first step. Their approach was implemented on both simulated and experimental spontaneous Raman spectral datasets and showed good quantification of glucose concentration. As a result, the authors suggested using this algorithm as a complementary tool for Raman spectroscopy-based mixture quantification studies. Since *Arcobacter* is an emerging foodborne pathogen that has become more important in recent years, Wang et al[52] were interested in fast identification of Arcobacter. The authors combined Raman spectroscopy with deep learning via a CNN to identify various species of *Arcobacter*. Their method achieved a high identification accuracy (97.2%) at the species level. Furthermore, a fully connected artificial neural network (ANN) was constructed on Raman spectra to determine the actual ratio of a specific *Arcobacter* species in a bacterial mixture. In their approach, the accuracy of Raman spectroscopy for bacterial species determination was improved and enabled rapid identification of *Arcobacter*. A further challenge in assessing endoscopic disease severity in Ulcerative colitis (UC) patients using Raman spectroscopy was explored by Kirchberger-Tolstik et al.[53] In this study, the endoscopic disease severity evaluation was performed according to the four Mayo subscores. The authors coupled Raman spectra with a one-dimensional CNN (1D-CNN) to identify the level of colonic inflammation and then applied first-order Taylor expansion to extract the important Raman bands for this classification. Their approach indicated a good classification performance and can be used further as a complementary method for UC characterization and diagnosis. Zafar et al[54] introduced a novel method that monitors the oxyhemoglobin changes produced by neuronal activations using functional near-infrared spectroscopy (fNIRS). The authors suggested using a kernel-based recursive least squares (KRLS) algorithm to reduce the detection time in fNIRS signals from the neuronal activation. In this manner, the KRLS algorithm with a Gaussian kernel was used. It showed the best performance for estimating both changes in oxyhemoglobin and deoxyhemoglobin. Therefore, a neuronal activation can be determined in about 0.1 s with fNIRS using KRLS prediction, enabling almost real-time detection if combined with electroencephalography. A different matter that involves the detection of petroleum presence in soil mixtures was covered by J. Galán-Freyle et al[55] A remote-sensed tool that combines artificial intelligence and a portable mid-infrared quantum cascade laser spectroscopy (QCL) system was developed. First, remote sensing combined with support vector machine (SVM) was used to detect the presence or absence of traces of petroleum in soil. Then, PCA, PLS-DA, and SVM were implemented to discriminate between the different soil types. Additionally, a statistical analysis method was developed to calculate limits of detection (LOD) and limits of decision (LD) from fits of the detection probability. As a result, a SVM provided better identification probabilities of soils that contains traces of petroleum. Le[56] proposed the use of deep learning with NIR for rapid analysis of cereal characteristics. First, the author applied the deep learning-stacked sparse autoencoder (SSAE) method on the corn and the rice datasets. This deep learning tool reduces the NIR data dimension, eliminates the interference information, and obtains advanced data features. A combination of the affine transformation (AT) and the extreme learning machine (ELM) was then established to predict the different types of cereals. Their approach provides a fast, efficient, and cost-effective method for cereal characteristics analysis. Xu et al[57] used functional near-infrared spectroscopy (fNIRS) to investigate hemodynamic fluctuations in the bilateral temporal cortices for typically developing (TD) children and children with autism spectrum disorder (ASD). The authors proposed the use of fNIRS time series to estimate the global time-varying behavior of brain activity and then combined two deep learning networks, LSTM and CNN, to explore the potential patterns of temporal variation for ASD identification. This global time-varying behavior is measured through the Augmented Dickey-Fuller (ADF) test. The ADF test showed that ASD children performed weaker stationarity in hemodynamic fluctuation variation than controls, as illustrated in Figure 4. Also, the proposed deep learning approach was able to differentiate between ASD and TD children accurately.

Accordingly, the characterization of the time-varying behavior of brain activity holds promising potential for a better understanding of the underlying causes of ASD. Furthermore, the deep learning framework has the potential for diagnosing children with the risk of ASD.

3.4 | X-ray

X-ray spectroscopy and X-ray intensity measurements enable the characterization of materials. This is done by X-ray excitation, for example, high energetic electromagnetic radiation, which results in the emission of characteristic wavelengths for the elements of the specimen/sample. These specific wavelengths can be used to generate insights in the elemental composition of the sample. X-ray spectroscopy can be used to address a range of scientific questions, from interactions of simples molecules to the structure of the human brain.[58] Chemometrics, machine learning, and deep learning methods proved to be great tools to solve practical problems, especially in chemistry and spectroscopy. However, their application in the X-ray field is not as broad as in other spectroscopic areas. Therefore, recent trends concerning the application of chemometrics, machine learning, and deep learning methods on X-ray spectra are mentioned below.

Otsuka et al.[59] investigated the effect of humidity-controlled storage of amorphous rebamipide (RB), RB form I, and RB solid dispersion with different surfactant and polymers. Their method is based on applying PCA on the dataset generated from power X-ray diffractograms (PXRD) and NIR spectra. The authors showed that the fusion of data from different sources resulted in correlations between NIR spectra and diffraction patterns in both neat RB and solid dispersion samples. As a result, the presented methods can be a useful model for evaluating amorphous active pharmaceutical ingredients without a standard sample. An additional study was motivated by the recent improvements of portable X-ray devices to detect meteorites from hot and cold deserts. In this study, chemometrics was used for the analysis of the X-ray data, which Allegreta et al.[60] measured using a portable energy dispersive X-ray fluorescence spectroscopy (pED-XPF) instrument. Moreover, the meteorite classification was achieved

doi.org/10.1002/ansa.202000162

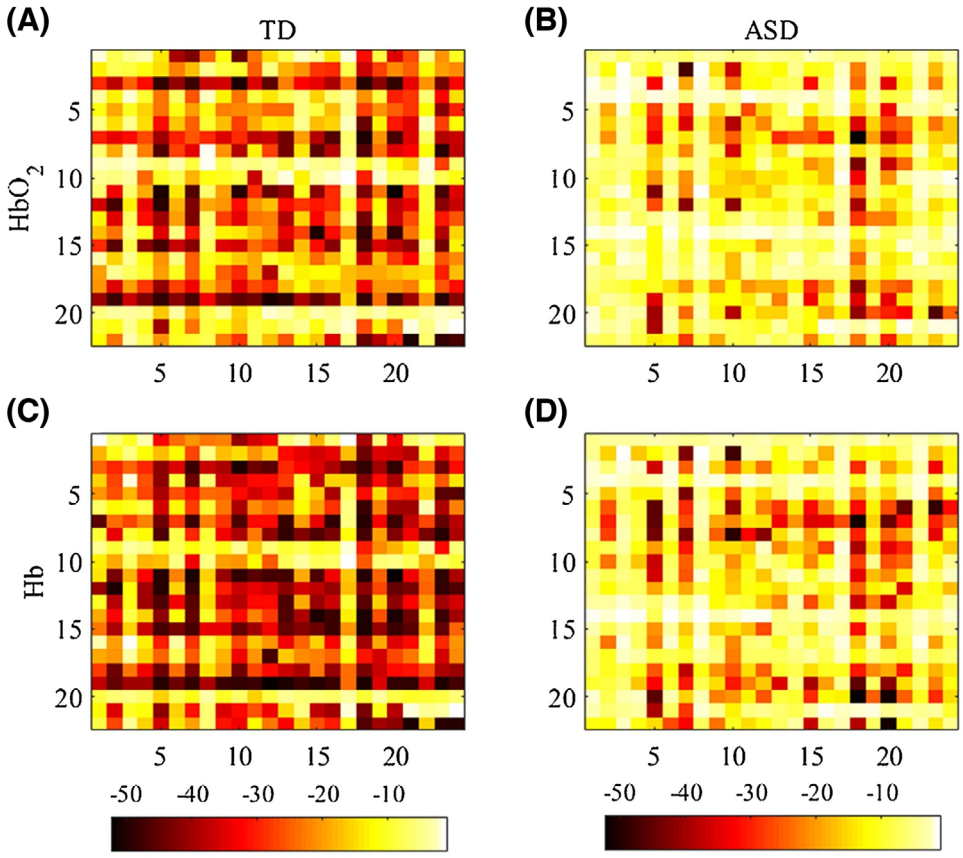

FIGURE 4 ADF color map values for ASD and TD children. The *x*-axis indicates the number of optical channels, while the *y*-axis shows the numbering of subjects. A clear distinction in the response between ASD and TD children is concluded. [Retrieved and adapted from reference[57]].

by applying various chemometrics methods on standardized X-ray fluorescence (XRF) spectra. In this regard, PCA, cubic support vector machine (CSVM), fine kernel nearest neighbor (FKNN), subspace discriminant-ensemble classifiers (SD-EC), and subspace discriminant KNN-EC (SKNN-EC) methods were tested. Their approach allowed the rapid and trustable classification and discrimination of meteorites in macro-groups. 100% accuracy in sample classification was obtained using each of these machine learning methods. Consequently, their approach proved effective and promising for the differentiation and classification of real or supposed meteorites. Carbone et al.[61] proposed a graph-based deep learning architecture to predict the X-ray absorption near-edge structure (XANES) spectra of molecules. Briefly, XANES encodes vital information about the local chemical environment, but significant challenges arise from the material's complexity associated with chemical composition and structure. The author proved with their approach that the predicted spectra reproduce all prominent peaks, with 90% of the predicted peak locations within 1 eV of the ground truth. This method can also be used to provide a general-purpose, high-throughput capability for predicting spectral information of a broad range of materials, including molecules, crystals, and interfaces. Other issues were researched by Mullaliu et al,[62] who investigated the electrochemical activity in manganese hexacyanoferrate (MnHCF) by varying the interstitial ion content. The authors combined X-ray absorption spectroscopy and a MCR-ALS. Their approach intent to assess the structural and electronic modifications during Na release and Li insertion. As a result, MCR-ALS showed that water absorption affects the reaction dynamics only at the Fe site. Besides, the Mn local environment encountered a substantial yet reversible Jahn-Teller effect upon interstitial ion removal due to the formation of trivalent Mn. Furthermore, this is associated with decreased of equatorial Mn—N bond lengths by 10%.

4 | CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING METHODS FOR THE ANALYSIS OF IMAGE DATA

AI-based techniques like chemometrics, machine learning, and deep learning are used to analyze chemical image data for decades. These tools were adapted to different structures and types of images, from the simplest grayscale image to hyperspectral images, and provided new insights on their spatial and spectroscopic information. Recent trends in applying chemometrics, machine learning, and deep learning methods on image data from atomic force microscopy (AFM), electron microscopy (EM), and 2D chromatography are presented below.

4.1 | Atomic force microscopy (AFM) and electron microscopy (EM)

Initially, atomic force microscopy (AFM) is a high-resolution imaging technique where a small probe with a sharp tip is scanned back and forth in a controlled manner across a sample to measure its surface. This procedure results in a topography of the sample at atomic resolution. AFM microscopy techniques can be performed in various scanning modes that enable nanoscale characterization of different material properties such as electrical, magnetic, and mechanical properties.[63]

On the other hand, Electron microscopy (EM) uses an electron beam to create an image of a sample. Because of the higher energy of the electrons compared with visible light, an electron microscope has a much higher resolution than a light microscope. EM can be used to investigate the microstructure of a wide range of biological and inorganic specimens. Moreover, it provides morphologic and crystallographic information.[64]

The incorporation of chemometrics, machine learning, and deep learning methods for AFM imaging techniques is limited, and recent applications are mentioned below. For instance, Yablon et al[65] investigated three different applications of machine learning in AFM imaging. In their first application, the authors applied two AI-based methods like neural networks and CNNs. These networks are trained to differentiate two different multicomponent polymer blends based on their AFM phase images. The results showed that CNN performs perfectly with 100% accuracy on the test data. A feature extraction approach was investigated in their second application to detect particles in an image with a complex background and many aggregates. In this manner, an initial logistic regression model was trained. The corresponding output was fed to a Hessian blob detection algorithm to isolate particles with a circular shape. Their proposed method significantly improves the particle identification compared with a commercial particle analysis package. Finally, the authors discussed the current status of autonomous instrumentation in AFM and its limitation, which needs a large amount of optimization. The additional issue concerned with the time required by the oscillating tip to reach the steady-state motion in AFM imaging was discussed in Javazm et al.[66] Due to AFM restricted scanning speed, the authors proposed an innovative imaging technique based on an artificial intelligence-based algorithm. Thereby, multiple artificial intelligence-based methods were investigated, including multilayer perceptron, radial basis function neural networks, and adaptive neural fuzzy inference system networks. Their approach aims to show the capability of artificial intelligence methods to estimate the surface topography directly. Therefore, the results showed that the multilayer perceptron overcame the other techniques in terms of surface characteristics estimation. In conclusion, the authors suggested using their approach to reach an accurate and fast estimation of the surface topography in AFM imaging. Further advantages of their method are that no closed-loop controller is needed, and the capability of estimating simultaneous the topography, the Hamaker parameter, and the tip-sample interaction force. Regarding the AFM limitation mentioned above, a further investigation was performed by Payam et al[67] The authors intended to explore the probe-sample interactions in dynamic AFM. Therefore, a novel approach for dynamic AFM data acquisition and imaging based on wavelet transform was applied in the photodetector data stream. By use of simulations, their approach was able to produce data including information about the transient response of amplitude and phase with the variation of material and sample topography properties. The method reliability was tested by comparing it with a standard lock-in amplifier (LIA) analysis. It showed the ability to reconstruct amplitude and phase images of standard samples, starting from time-domain data of actual measurements. The authors indicated that using their method would improve the measurement speed, reduce the loss of information and give access to a wealth of information about the transient response, which leads to the possibility of analyzing material properties in dynamic AFM.

Some of the challenges in electron microscopy (EM) imaging that are resolved using chemometrics, machine learning, and deep learning methods are presented below. For instance, Yu et al[68] addressed the limitations of traditional image recognition methods, such as the inability to obtain the complete pore space characteristics in scanning electron microscopy (SEM) images. Additionally, traditional image recognition methods for SEM images lead to poor segmentation results and a low accuracy. The authors implemented a semantic image segmentation technique based on artificial intelligence to analyze the pore characteristics and explore the relationship between the microscopic pore characteristics and the macroscopic permeability parameters of the sandstone in the SEM images. The results showed that the application of deep learning via CNNs accurately recognize images and allowed the automatic processing of microscopic images. Furthermore, this significantly improved the accuracy of the pore identification in rock samples. Li et al.[34] investigated a reaction-convection-diffusion model to track spatial-temporal patterns in scanning transmission electron microscopy (STEM) videos of Pt nanoparticle formation and graphene contamination. The authors developed a data-driven approach utilizing pixel-level information to infer the underlying partial differential equation (PDE) that governs the spatial-temporal patterns in STEM videos. The PDE model resulted in a redundant basis matrix, leading to non-unique numerical solutions. Therefore, the least angle regression algorithm (LARS) was utilized to reduce the ambiguity and to improve its interpretation. Additionally, the optimal parameter λ, used to balance model parsimony and descriptive capability, is determined by Mallow's Cp criteria (Cp). The Pearson correlation coefficient (PCC) is used to track discrepancies between experimental and estimated frames. The analysis applied to STEM multiple Pt particles video is illustrated in Figure 5.

Both the simulated and experimental datasets proved that the use of the PDE models has the potential to capture the characteristic behavior of spatial-temporal patterns at a mesoscopic scale in STEM videos and can be of great help for the investigations of complex time-evolving processes.

4.2 | Two-dimensional chromatography

Two-dimensional chromatography is a chromatographic technique that yields information on the chemical composition of a sample, by

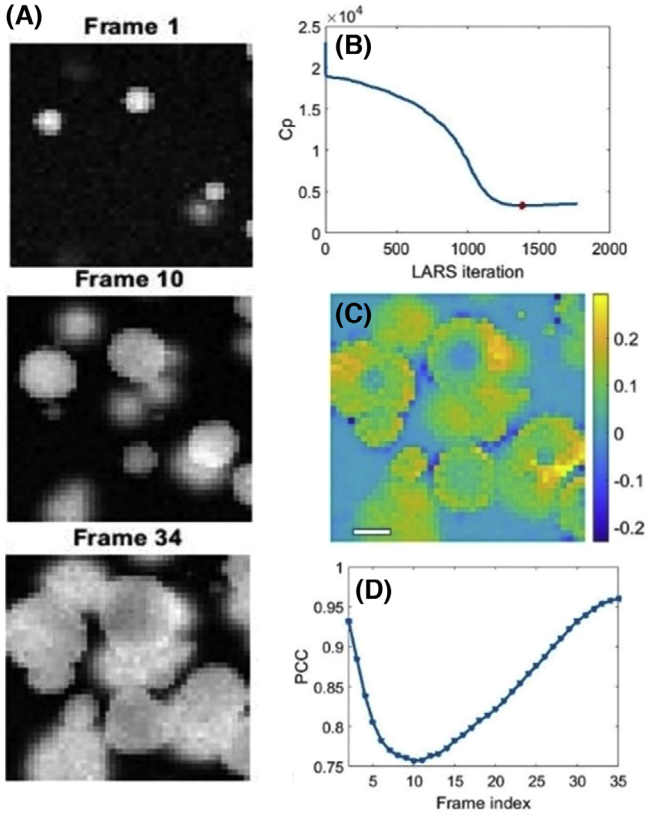

FIGURE 5 Estimation based on multiple Pt particles STEM video. Snapshots of experimental data are illustrated in (A). The optimal parameter calculated using Cp criteria is shown in red in (B). The estimated frame is shown in (C). PCC in (D) proved the convergence to the experimental data between frame 11 and 35. [Retrieved from reference [34]]

combining two separation systems. Typically, two different chromatographic columns are connected in sequence, and an aliquot from the first column is injected in the second column. As the separation systems are often working differently in the columns, peaks that could not be separated using the first column can be separated using the second column in 2D chromatography.[69] A large amount of information is contained in high-resolution chromatography, and it is complicated to extract all relevant information and deduce correct and straightforward solutions. However, recent research for efficient chemometric data-processing strategies is presented below.

Huygens et al.[70] reported three evolutionary algorithms to enhance searches in the method development spaces of 1D- and 2D-chromatography, including genetic algorithms (GA), evolution strategies (ES), and covariance matrix adaptation-evolution strategy (CMA-ES). The authors compared these algorithms to a plain grid search. The results showed significant outperformance, especially in terms of the number of search runs needed to achieve a given separation quality. Additionally, the ES and GA performance followed a hyperbolic law in the large search run number limit. Subsequently, the convergence rate in the hyperbolic function can quantify the difference in the required number of search runs in these algorithms. A further problem on two-dimensional liquid chromatography was tackled by Pérez-Cova et al.[71] The authors employed a two-dimensional liquid chromatography method hyphenated simultaneously to two different detectors on a mixture of 31 pharmaceutical compounds. MCR-ALS was then used to evaluate the obtained two-dimensional chromatograms. The authors perform two evaluations. First they assessed the multilinear behavior of the high-dimensional data for each of the two detection modes. Second, they check the model performance for the multiset data obtained by fusion of the data coming from both detectors. Additionally, their approach proved that data fusion from the two detectors increased the ability for compound identification. Nagai et al.[72] identified a novel biomarker of hepatocellular carcinoma (HCC) in the human liver using multivariate analysis methods such as PCA and OPLS-DA. The data was extracted using an ultrahigh-performance liquid chromatography/quadrupole time-of-flight mass spectroscopy (UHPLC/QTOFMS) instrument equipped with a mixed-mode column. The results showed that novel biomarkers for HCC were identified with the global metabolomics/metabolic profiling (G-Met) method. Besides this, the difference in fatty acid species of triglyceride in tumor regions was demonstrated by high definition mass spectroscopy (HDMS) combined with UHPLC/QTOFMS. It showed localization in cryosections using desorption electrospray ionization-mass spectroscopy imaging (DESI-MSI). In conclusion, G-Met combined with UHPLC/QTOFMS and HDMS and distribution analysis by DESI-MSI is useful for characterizing tumor cell progression and discovering prospective biomarkers.

5 | SUMMARY AND OUTLOOK

Developments of artificial intelligence-based techniques like chemometrics, machine learning and deep learning have occupied the interest of researchers for decades. These data analysis techniques combined with spectroscopic measurements in chemistry and chemical data have gained popularity and yielded promising application possibilities in various fields from the food industry to biomedical applications. This review paper discussed the recent investigations on AI-based techniques for specific spectroscopic measurements and imaging approaches including NMR, MS, vibrational spectroscopy, X-ray, AFM, EM, and 2D chromatography. Each of these measurement techniques and the application tasks requires specific properties of the analysis methods. In that sense, the analysis methods are task and data-dependent and are discussed separately. However, the enhancement of the data quality is a common procedure in most of the reviewed studies. This enhancement is achieved by either applying inverse problems or preprocessing techniques. In this regard, deep learning techniques via CNN and LSTM became popular as solutions for the inverse problem for spectroscopic data. On the other hand, the modification of existing preprocessing techniques or their applications in new areas is a very common trend. Following the enhancement of data quality, data modeling for a variety of tasks was reviewed. For instance, the discrimination between different groups and the

quantification of a specific variable require either classification or regression methods. In the reviewed studies, PCA, PLS, SVM, SIMCA, ANN, PLSR, and deep learning were implemented for the spectroscopic measurements and the chemical data discussed herein (NMR, MS, vibrational spectroscopy, X-ray, AFM, EM, and 2D chromatography).

However, attempts to improve the predictive quality and robustness of artificial intelligence-based methods such as chemometrics, machine learning, and deep learning are performed in many application fields. Also, investigations to develop new AI-based techniques are increasing as well. Furthermore, strategies to overcome the lack of available data, especially in biomedical applications, and the advancement of data fusion methods are still subjects in further research.

ACKNOWLEDGMENTS

The financial support from the Deutsche Forschungsgemeinschaft (DFG) via the CRC 1076 AquaDiva is highly appreciated.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Berry RJ, Ozaki Y. Comparison of wavelets and smoothing for denoising spectra for two-dimensional correlation spectroscopy. *Appl Spectrosc*. 2002;56(11):1462-1469. https://doi.org/10.1366/00037020260377779
2. Chen H, Xu W, Broderick N, Han J. An adaptive denoising method for Raman spectroscopy based on lifting wavelet transform. *J Raman Spectrosc*. 2018;49(9):1529-1539. https://doi.org/10.1002/jrs.5399
3. Guo S, Bocklitz T, Popp J. Optimization of Raman-spectrum baseline correction in biological application. *Analyst*. 2016;141(8):2396-2404. https://doi.org/10.1039/C6AN00041J
4. Shao L, Griffiths PR. Automatic baseline correction by wavelet transform for quantitative open-path fourier transform infrared spectroscopy. *Environ Sci Technol*. 2007;41(20):7054-7059. https://doi.org/10.1021/es062188d
5. Xi Y, Rocke DM. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*. 2008;9(1):324. https://doi.org/10.1186/1471-2105-9-324
6. Efitorov A, Burikov S, Dolenko T, Laptinskiy K, Dolenko S. Significant feature selection in neural network solution of an inverse problem in spectroscopy*. *Procedia Computer Science*. 2015;66:93-102. https://doi.org/10.1016/j.procs.2015.11.012
7. Dolenko SA, Burikov SA, Dolenko TA, Persiantsev IG. Adaptive methods for solving inverse problems in laser raman spectroscopy of multicomponent solutions. *Pattern Recognit Image Anal*. 2012;22(4):550-557. https://doi.org/10.1134/S1054661812040049
8. Sankaran S, Ehsani R. Visible-near infrared spectroscopy based citrus greening detection: Evaluation of spectral feature extraction techniques. *Crop Prot*. 2011;30(11):1508-1513. https://doi.org/10.1016/j.cropro.2011.07.005
9. Zhu C, Palmer GM, Breslin TM, Harter J, Ramanujam N. Diagnosis of breast cancer using diffuse reflectance spectroscopy: Comparison of a Monte Carlo versus partial least squares analysis based feature extraction technique. *Lasers Surg Med*. 2006;38(7):714-724. https://doi.org/10.1002/lsm.20356
10. Balabin RM, Smirnov SV. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Anal Chim Acta*. 2011;692(1):63-72. https://doi.org/10.1016/j.aca.2011.03.006
11. Li S, Chen G, Zhang Y, et al. Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques. *Opt Express, OE*. 2014;22(21):25895-25908. https://doi.org/10.1364/OE.22.025895
12. Faix O. Classification of Lignins from Different Botanical Origins by FT-IR Spectroscopy. *Holzforschung*. 1991;45(s1):21-28. https://doi.org/10.1515/hfsg.1991.45.s1.21
13. Geballe TR, Knapp GR, Leggett SK, et al. Toward Spectral Classification of L and T Dwarfs: Infrared and Optical Spectroscopy and Analysis. *Astrophys J*. 2002;564(1):466-481. https://doi.org/10.1086/324078
14. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl Spectrosc*. 2000;54(3):413-419. https://doi.org/10.1366/0003702001949500
15. Minasny B, McBratney AB. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemom Intell Lab Syst*. 2008;94(1):72-79. https://doi.org/10.1016/j.chemolab.2008.06.003
16. Chatzidakis M, Botton GA. Towards calibration-invariant spectroscopy using deep learning. *Sci Rep*. 2019;9(1):2126. https://doi.org/10.1038/s41598-019-38482-1
17. Kyathanahally SP, Döring A, Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. *Magn Reson Med*. 2018;80(3):851-863. https://doi.org/10.1002/mrm.27096
18. Mishra P, Biancolillo A, Roger JM, Marini F, Rutledge DN. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC, Trends Anal Chem*. 2020;132(o):116045. https://doi.org/10.1016/j.trac.2020.116045
19. Torniainen J, Afara IO, Prakash M, Sarin JK, Stenroth L, Töyräs J. Open-source python module for automated preprocessing of near infrared spectroscopic data. *Anal Chim Acta*. 2020;1108:1-9. https://doi.org/10.1016/j.aca.2020.02.030
20. Martyna A, Menżyk A, Damin A, et al. Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components. *Chemom Intell Lab Syst*. Published online 2020:104029. https://doi.org/10.1016/j.chemolab.2020.104029
21. Roger J-M, Biancolillo A, Marini F. Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemom Intell Lab Syst*. 2020;199:103975. https://doi.org/10.1016/j.chemolab.2020.103975
22. Zhang C, Zhou L, Zhao Y, Zhu S, Liu F, He Y. Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. *Chemom Intell Lab Syst*. Published online 2020:104063.
23. Raulf AP, Butke J, Menzen L, et al. Deep Neural Networks for the Correction of Mie Scattering in Fourier-Transformed Infrared Spectra of Biological Samples. arXiv preprint arXiv:200207681. Published online 2020.
24. Wahl J, Sjödahl M, Ramser K. Single-Step Preprocessing of Raman Spectra Using Convolutional Neural Networks. *Appl Spectrosc*. 2020;74(4):427-438.
25. Argoul P. Overview of Inverse Problems. 17.
26. Nakamura G. *Inverse Modeling An Introduction to the Theory and Methods of Inverse Problems and Data Assimilation*. IOP Publishing; 2015. https://doi.org/10.1088/978-0-7503-1218-9
27. Yuan Y, Demers H, Brodusch N, Wang X, Gauvin R. Inverse modeling for quantitative X-ray microanalysis applied to 2D heterogeneous materials. *Ultramicroscopy*. 2020;219:113117.
28. Hong S, Qin S, Dong P, et al. Quantification of rust penetration profile in reinforced concrete deduced by inverse modeling. *Cem Concr Compos*. Published online 2020:103622.
29. Takeda S, Hama T, Hsu H-H, et al. AI-driven Inverse Design System for Organic Molecules. *arXiv preprint arXiv*:200109038. Published online 2020.

30. Houhou R, Barman P, Schmitt M, Meyer T, Popp J, Bocklitz T. Deep learning as phase retrieval tool for CARS spectra. *Opt Express*. 2020;28(14):21002-21024.
31. Guo S, Mayerhöfer T, Pahlow S, Hübner U, Popp J, Bocklitz T. Deep learning for 'artefact'removal in infrared spectroscopy. *Analyst*. 2020;145(15):5213-5220.
32. Magnussen EA, Solheim JH, Blazhko U, et al. Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells. *J Biophotonics*. n/a(n/a):e202000204. https://doi.org/10.1002/jbio.202000204
33. Hassenkam T, Fantner GE, Cutroni JA, Weaver JC, Morse DE, Hansma PK. High-resolution AFM imaging of intact and fractured trabecular bone. *Bone*. 2004;35(1):4-10. https://doi.org/10.1016/j.bone.2004.02.024
34. Li X, Dyck O, Unocic RR, Ievlev AV, Jesse S, Kalinin SV. Statistical learning of governing equations of dynamics from in-situ electron microscopy imaging data. *Materials & Design*. 2020;195:108973. https://doi.org/10.1016/j.matdes.2020.108973
35. The Basics of NMR. Accessed November 17, 2020. https://www.cis.rit.edu/htbooks/nmr/inside.htm
36. Pérez Y, Casado M, Raldúa D, et al. MCR-ALS analysis of 1 H NMR spectra by segments to study the zebrafish exposure to acrylamide. *Anal Bioanal Chem*. 2020;412(23):5695-5706.
37. Miros FN, Murch SJ, Shipley PR. Exploring feature selection of St John's wort grown under different light spectra using 1H-NMR spectroscopy. *Phytochem Anal*. Published online 2020.
38. Zhao F, Li W, Pan J, Chen Z, Qu H. A novel critical control point and chemical marker identification method for the multi-step process control of herbal medicines via NMR spectroscopy and chemometrics. *RSC Advances*. 2020;10(40):23801-23812.
39. Marion R, Govaerts B, von Sachs R. AdaCLV for Interpretable Variable Clustering and Dimensionality Reduction of Spectroscopic Data.; 2020.
40. Coimbra PT, Bathazar CF, Guimarães JT, et al. Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control*. 2020;110:107006.
41. Zhang J, Chen H, Fan C, Gao S, Zhang Z, Bo L. Classification of the botanical and geographical origins of Chinese honey based on 1H NMR profile with chemometrics. *Food Res Int*. 2020;137:109714.
42. Kong X, Zhou L, Li Z, et al. Artificial intelligence enhanced two-dimensional nanoscale nuclear magnetic resonance spectroscopy. *npj Quantum Information*. 2020;6(1):1-10.
43. Hou X, Wang G, Wang X, Ge X, Fan Y, Nie S. Convolutional neural network based approach for classification of edible oils using low-field nuclear magnetic resonance. *J Food Compos Anal*. Published online 2020:103566.
44. Rajawat J, Jhingan G. Chapter 1 - Mass spectroscopy. In: Misra G, ed. *Data Processing Handbook for Complex Biological Data Sources*. Academic Press; 2019:1-20. https://doi.org/10.1016/B978-0-12-816548-5.00001-0
45. Duan L, Ma A, Meng X, Shen G, Qi X. QPMASS: A parallel peak alignment and quantification software for the analysis of large-scale gas chromatography-mass spectrometry (GC-MS)-based metabolomics datasets. *J Chromatogr A*. Published online 2020:460999.
46. Alkhalifah Y, Phillips I, Soltoggio A, et al. VOCCluster: Untargeted Metabolomics Feature Clustering Approach for Clinical Breath Gas Chromatography/Mass Spectrometry Data. *Anal Chem*. 2019;92(4):2937-2945.
47. Papagiannopoulou C, Parchen R, Rubbens P, Waegeman W. Fast Pathogen Identification Using Single-Cell Matrix-Assisted Laser Desorption/Ionization-Aerosol Time-of-Flight Mass Spectrometry Data and Deep Learning Methods. *Anal Chem*. 2020;92(11):7523-7531.
48. Li K, Jain A, Malovannaya A, Wen B, Zhang B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics*. Published online 2020:1900334.
49. Vibrational Spectroscopy: Definition & Types - Video & Lesson Transcript. Study.com. Accessed November 17, 2020. https://study.com/academy/lesson/vibrational-spectroscopy-definition-types.html
50. Akpolat H, Barineau M, Jackson KA, et al. High-Throughput Phenotyping Approach for Screening Major Carotenoids of Tomato by Handheld Raman Spectroscopy Using Chemometric Methods. *Sensors*. 2020;20(13):3723.
51. Han N, Ram RJ. Bayesian modeling and computation for analyte quantification in complex mixtures using Raman spectroscopy. *Computational Statistics & Data Analysis*. 2020;143:106846.
52. Wang K, Chen L, Ma X, et al. Arcobacter Identification and Species Determination Using Raman Spectroscopy Combined with Neural Networks. *Appl Environ Microbiol*. 2020;86(20).
53. Kirchberger-Tolstik T, Pradhan P, Vieth M, et al. Towards an interpretable classifier for characterization of endoscopic Mayo scores in ulcerative colitis using Raman Spectroscopy. *Anal Chem*. Published online 2020.
54. Zafar A, Hong K-S. Reduction of onset delay in functional near-infrared spectroscopy: prediction of HbO/HbR signals. *Frontiers in Neurorobotics*. 2020;14:10.
55. Galán-Freyle NJ, Ospina-Castro ML, Medina-González AR, Villarreal-González R, Hernández-Rivera SP, Pacheco-Londoño LC. Artificial intelligence assisted Mid-infrared laser spectroscopy in situ detection of petroleum in soils. *Applied Sciences*. 2020;10(4):1319.
56. Le BT. Application of deep learning and near infrared spectroscopy in cereal analysis. *Vib Spectrosc*. 2020;106:103009. https://doi.org/10.1016/j.vibspec.2019.103009
57. Xu L, Liu Y, Yu J, et al. Characterizing autism spectrum disorder by deep learning spontaneous brain activity from functional near-infrared spectroscopy. *J Neurosci Methods*. 2020;331:108538.
58. X-ray - Fundamental characteristics. Encyclopedia Britannica. Accessed November 17, 2020. https://www.britannica.com/science/X-ray
59. Otsuka Y, Utsunomiya Y, Umeda D, Yonemochi E, Kawano Y, Hanawa T. Effect of polymers and storage relative humidity on amorphous rebamipide and its solid dispersion transformation: Multiple spectra chemometrics of powder X-Ray diffraction and near-infrared spectroscopy. *Pharmaceuticals*. 2020;13(7):147.
60. Allegretta I, Marangoni B, Manzari P, et al. Macro-classification of meteorites by portable energy dispersive X-ray fluorescence spectroscopy (pED-XRF), principal component analysis (PCA) and machine learning algorithms. *Talanta*. 2020;212:120785.
61. Carbone MR, Topsakal M, Lu D, Yoo S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Phys Rev Lett*. 2020;124(15):156401. https://doi.org/10.1103/PhysRevLett.124.156401
62. Mullaliu A, Aquilanti G, Conti P, Giorgetti M, Passerini S. Effect of Water and Alkali-Ion Content on the Structure of Manganese (II) Hexacyanoferrate (II) by a Joint Operando X-ray Absorption Spectroscopy and Chemometric Approach. *ChemSusChem*. 2020;13(3):608-615.
63. AFM Microscopes - Introduction to Atomic Force Microscopy | Bruker. Bruker.com. Accessed November 17, 2020. https://www.bruker.com/products/surface-and-dimensional-analysis/atomic-force-microscopes/campaigns/afm-microscopes.html
64. *Electron Microscopy introduction*. WUR. Published March 21, 2014. Accessed November 17, 2020. https://www.wur.nl/en/Value-Creation-Cooperation/Facilities/Wageningen-Electron-Microscopy-Centre/Electron-Microscopy-intro.htm
65. *Machine learning to enhance atomic force microscopy analysis and operation*. Wiley Analytical Science. Accessed November 10, 2020. https://analyticalscience.wiley.com/do/10.1002/was.00010012

66. Javazm MR, Pishkenari HN. Observer Design for Topography Estimation in Atomic Force Microscopy Using Neural and Fuzzy Networks. *Ultramicroscopy*. Published online 2020:113008.
67. Payam AF, Biglarbeigi P, Morelli A, Lemoine P, McLaughlin J, Finlay D. Data acquisition and imaging using wavelet transform: a new path for high speed transient force microscopy. *Nanoscale Advances*. Published online 2020.
68. Yu Q, Xiong Z, Du C, et al. Identification of rock pore structures and permeabilities using electron microscopy experiments and deep learning interpretations. *Fuel*. 2020;268:117416.
69. Kilz P. Two-Dimensional Chromatography as an Essential Means for Understanding Macromolecular Structure. *Chromatographia*. 2004;59(1):3-14. https://doi.org/10.1365/s10337-003-0106-7
70. Huygens B, Efthymiadis K, Nowé A, Desmet G. Application of evolutionary algorithms to optimise one-and two-dimensional gradient chromatographic separations. *J Chromatogr A*. 2020;1628:461435.
71. Pérez-Cova M, Tauler R, Jaumot J. Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior. *Chemom Intell Lab Syst*. Published online 2020:104009.
72. Nagai K, Uranbileg B, Chen Z, et al. Identification of novel biomarkers of hepatocellular carcinoma by high-definition mass spectrometry: Ultrahigh-performance liquid chromatography quadrupole time-of-flight mass spectrometry and desorption electrospray ionization mass spectrometry imaging. *Rapid Commun Mass Spectrom*. 2020;34(S1):e8551. https://doi.org/10.1002/rcm.8551

How to cite this article: Houhou R, Bocklitz T. Trends in artificial intelligence, machine learning and chemometrics applied to chemical data. *Anal Sci Adv*. 2021;2:128–141. https://doi.org/10.1002/ansa.202000162