Analytica Chimica Acta 1161 (2021) 338403

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

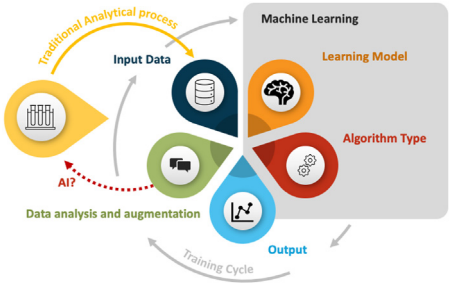
Review

Taking the leap between analytical chemistry and artificial intelligence: A tutorial review

Lucas B. Ayres [a], Federico J.V. Gomez [b], Jeb R. Linton [c], Maria F. Silva [b], Carlos D. Garcia [a, *]

[a] Department of Chemistry, Clemson University, Clemson, SC, 29634, USA
[b] Instituto de Biología Agrícola de Mendoza (IBAM-CONICET), Facultad de Ciencias Agrarias, Universidad Nacional de Cuyo, Mendoza, Argentina
[c] IBM Watson and Cloud Platform, Armonk, NY, 10504, USA

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 4 December 2020
Received in revised form
2 March 2021
Accepted 3 March 2021
Available online 15 March 2021

Keywords:
Artificial intelligence
Deep learning
Analytical
Sensors
Spectroscopy

ABSTRACT

The last 10 years have witnessed the growth of artificial intelligence into different research areas, emerging as a vibrant discipline with the capacity to process large amounts of information and even intuitively interact with humans. In the chemical world, these innovations in both hardware and algorithms have allowed the development of revolutionary approaches in organic synthesis, drug discovery, and materials' design. Despite these advances, the use of AI to support analytical purposes has been mostly limited to data-intensive methodologies linked to image recognition, vibrational spectroscopy, and mass spectrometry but not to other technologies that, albeit simpler, offer promise of greatly enhanced analytics now that AI is becoming mature enough to take advantage of them. To address the imminent opportunity of analytical chemists to use AI, this tutorial review aims to serve as a first step for junior researchers considering integrating AI into their programs. Thus, basic concepts related to AI are first discussed followed by a critical assessment of representative reports integrating AI with various sensors, spectroscopies, and separation techniques. For those with the courage (and the time) needed to get started, the review also provides a general sequence of steps to begin integrating AI into their programs.

Contents

* Corresponding author. 211 S. Palmetto Blvd, Clemson, SC, 29634, USA.
E-mail address: cdgarci@clemson.edu (C.D. Garcia).

https://doi.org/10.1016/j.aca.2021.338403
0003-2670/

1. Introduction

Artificial intelligence (AI) is a relatively new field of study that aims at solving problems using computers, running code that mimics the cognitive processes of the human brain. Most modern algorithms involved in AI are not only able to connect inputs with outputs but also to adapt to environmental clues and take actions, increasing the chances of providing an accurate answer. Probably the main advantage of AI is that it provides a tremendous capacity to get meaningful and unbiased information from datasets that are so large and/or so complex, no human could possibly analyze them [1]. Moreover, the constant increase of computers' processing ability, along with the development of powerful algorithms and their distribution under open-source platforms (e.g. frameworks, APIs and training data) have allowed the application of AI to many fields of science [2–7]. Among those, GPS navigation apps, self-driving cars, voice-assistants (human language recognition), and IBM's Watson [8] are probably some of the most tangible examples of advancements in AI and its penetration in our daily activities. In terms of chemistry, AI has enabled ground-breaking advances linked to drug discovery [9], material's design [10–12] as well as organic synthesis [13]. Innovations in the latter category are particularly notable due to the capabilities of new computational approaches (molecular design algorithms) that allow exploring broad chemical spaces and bolster research in areas such as molecule property prediction [14], molecule design [15], retro-synthesis [16], reaction conditions predictions [17] and reaction outcome prediction [18]. Despite the advances in learning, the release of user-friendly frameworks, and the availability of pre-trained neural networks, the use of AI for analytical purposes has not been intensively explored and remains (comparatively speaking) poorly understood. These issues can be attributed to the gap between the existing academic training and the complexity of modern algorithms applied in data science.

To address this problem, that has hindered the implementation of AI in the area of analytical chemistry for several decades already [19], this tutorial review aims to provide basic information to junior researchers considering the integration of AI into their research programs. We believe these professionals are uniquely positioned to take advantage of these technologies to speed up developments. The review provides an overview of the most relevant aspects of AI (e.g. algorithms, training models and outputs) with emphasis on recent advances in deep learning and its uses in relevant areas of chemistry. Besides providing examples linked to chemical applications, the review also presents a critical assessment of the application of AI into different analytical sub-disciplines (e.g. separations, spectroscopy, detection systems, etc.) as well as a quick discussion of potential directions where AI could have the biggest impacts in the near future. In order to provide realistic expectations, we also present a quick introduction to the steps required for AI implementation, with links to pertinent training material.

2. Brief history of AI

Although the *concept* of artificial intelligence can be found in Greek mythology (where some *things* were able to move autonomously) [20], the modern idea of AI really started when questions such as *"can computers mimic the human mind?"* began to draw the attention of the scientific community [21]. That said, it is generally accepted that AI (as a field of study) was born the summer of 1956, after a famous conference by John McCarthy [22], at Dartmouth College. Since then, AI is defined as a field of computer science

aimed at increasing the ability of computer systems to analyze a certain environment and take actions to solve specific problems [23]. The period following this conference provided remarkable demonstrations of AI's applications such as the ELIZA program [24] that was a chatterbot capable of processing natural language and, the system STUDENT [25], capable of solving straightforward algebra problems. This period is known as the Golden Years (1956–1974) in which several (over) optimistic projections of AI's uses were done in parallel with massive funding from governmental agencies such as the U.S. DARPA. This was the period that allowed several major developments in neural networks. However, the successive failures to solve real-world problems and to reach ambitious projections, along with technical and theoretical limitations contributed to reductions in funding for this field as well as a decrease in the overall enthusiasm about the idea. It is important to clarify that, at that point, most of the tasks were solved by preprogrammed Boolean expressions (IF/THEN), without any real *learning*. In other words, these algorithms were only able to respond to a series of preprogrammed decisions. Despite these issues and taking advantage of the increased capacity to store and read data, the following years (1980–2010) were dedicated to developing systems with the capacity to learn from a dataset and then provide reasonable outputs such as regression, clustering, and classification. These accomplishments, commonly referred to as *machine learning* (ML), were also made possible due to advances in statistical tools as well as an increase in the processing power of computers (integrated circuits with more transistors) to solve and analyze complex mathematical tasks. Such breakthroughs came with a new enthusiastic boom with applications in many scientific fields such as robotics [26], computational vision [27], and cognitive science [28]. The most recent growth and applicability of AI can be also attributed to the continuous improvement of machine learning algorithms based on complex artificial neural networks (e.g. deep learning) [29,30]. Moreover, the "big data era" [31], the easy access to AI frameworks based on online platforms, and the availability of supercomputers have all facilitated the training and sharing of complex models to predict unusual relationships as well as to perform more complicated tasks. Along with the shift from traditional machine learning techniques to the use of deep neural networks, progress in the field has been facilitated by the development of powerful (and accessible) graphics processing units (GPUs/Graphics Cards) which can perform massive matrix math calculations far faster than a typical CPU. This progress has also been the primary driving factor in the democratization of AI, since traditional machine learning required far more extensive engineering of input data (usually referred to as Feature Engineering). This *new* ability of computers to extract patterns from high-volumes of data (even when poorly structured) is the engine behind bots that already demonstrated high-level functioning in games such as DOTA2 [32], chess [33], and AlphaGo [34].

It is also important to mention that although most of the papers discussed in this review do not make an explicit difference between AI and machine learning, there is a subtle distinction between the terms. The former is a much wider concept that includes not only the learning process from a dataset (technically defined as machine learning) but also the simulation of human thinking and behavior in response to a situation. Based on these advances, computers can now assist in a wide number of fields related to chemistry including healthcare [35–38], energy [39], drinking water [40], food processing [41], fluid dynamics [30,42], and even review scientific articles [43]. Again, and as summarized in Section 7 of this review, the application of these technologies is slowly (but surely) evolving from classic machine learning into the AI domain, providing real opportunities to develop *smart* protocols that can respond to the environment in ways that traditionally required the involvement of a human. Examples supporting specific approaches are provided in the review.

3. Classification of machines and algorithms

Due to the previously described advantages, it is clear that AI represents a powerful approach to explore the chemical space and can help chemists with many specific tasks. Aiming to clarify discussions about this topic, the following paragraphs describe some key technical aspects about AI, including a general overview, common algorithms, and training models. It is not our goal to provide mathematical details about each algorithm, as such deductions can be found elsewhere [29,44]. Readers are also encouraged to consult previously-published reviews linked to applications of AI to different aspects of chemistry [3,45–51], noting that a significant fraction of them are oriented towards organic synthesis and drug discovery.

3.1. General overview

The main purpose of any AI system is to get meaningful information from a dataset. This dataset (or input) could be a collection of variables from the chemical space (e.g. x, y, z) connected to an output via a regression, classification, or clustering problem. In other words, and regardless of the complexity, computer-based strategies are simply used to link the input data to an output result. To accomplish this task, many algorithms have been developed, each showing its own advantages and limitations, relative to the nature of the input, the environment, and the desired output. Thus, and while not intended to be comprehensive, this section provides a brief survey of the most common algorithms currently used in analytical chemistry as well as different strategies (supervised, unsupervised or reinforcement) that can be applied to train them. Fig. 1 summarizes a general overview of these topics. Additional information related to common data scientific terms and their definitions can be found in Ref. [52].

3.2. Learning in machine learning

3.2.1. Supervised learning

Supervised learning is by far the most common form of machine learning and refers to the practice of training a model using labeled data. For example, a dataset of photographs containing pictures of cats and dogs, with each picture labeled as either *cat* or *dog*, or with labels specifying the breed or some other useful characteristic (this specific example is later discussed in Section 6). This dataset could be used to train an AI system (a 'classifier') to predict the appropriate label for new, previously unseen pictures. In a similar way, a dataset containing the vapor pressures of chemical compounds could be used to train an AI system to perform a regression and predict the vapor pressures of other compounds, not found in its training data set. In effect, if vapor pressure is the desired output, the detailed list of known vapor pressures becomes the 'label' attached to each compound. These training datasets are typically created and maintained (or 'curated' in AI parlance) by humans, often at great effort. Because these data are used by the AI as the ultimate source of information, they are often referred to as *Ground Truth*. A variation of supervised learning known as self-supervised learning has gained recent notoriety. This is the practice of using inherent features of the data themselves as labels: for example, using text downloaded from Wikipedia and other sources, one can train a system to predict the next word in a sequence of words, or the missing word if a word is randomly dropped from a sequence. This technique, developed for language modeling, has also been applied to chemistry by training a neural network to predict

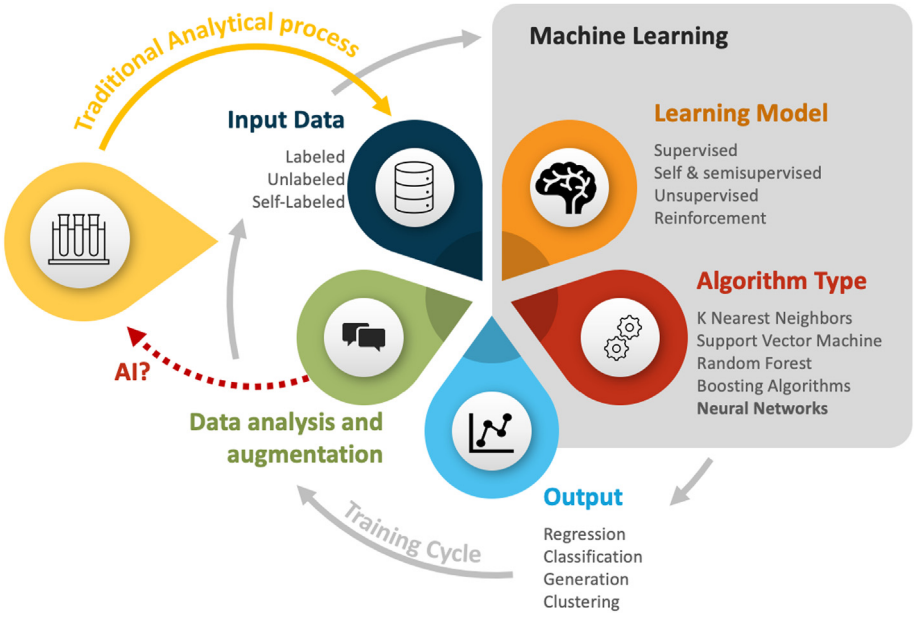


Fig. 1. Key aspects and representative examples of the components of artificial intelligence, noting how traditional methodologies (biosensors, chromatography, spectroscopy, etc) intersect with ML approaches and how ML could soon support the development of new analytical tools. Designed by PresentationGO.com.

missing atoms in the SMILES sequence of chemical reactions [53,54].

3.2.2. Unsupervised learning

Unsupervised Learning is another relatively common practice by which analysts examine unlabeled data to find trends and patterns, most typically using principal component analysis or clustering algorithms. This capacity to find patterns in the data over the noise (plain information with no meaning) was applied to investigate the degradation of a catalyst using 3D images [55] and is behind Mol2vec, an approach that learns vector representations of molecular substructures that point in similar directions for chemically-related substructures [56]. The application of this type of machine learning towards classification and coarse-graining of molecular simulations were recently reviewed [57]. As a slight modification of this approach, where the most important components or patterns in unlabeled data can be applied to create labels for subsequent supervised learning, known as semi-supervised learning, was recently applied to predict bioactivity of compounds [58] and to develop a fully automated methodology for validation of the identification of rarely occurring post-translational modifications for peptides by MS/MS [59].

3.2.3. Reinforcement learning

Reinforcement learning is a technique rarely used in practical business systems, but which has in recent years shown tremendous power in use cases where complex planning and strategy are required – especially in games. Reinforcement learning systems are trained by defining a scoring system and allowing the AI to run through a scenario many times – often tens or hundreds of thousands of times - with the objective of gradually learning to maximize the score. This approach has been used, for example, to design molecules with specified properties [60].

3.3. Examples of machine learning algorithms

Unfortunately, there is no set of golden rules to select the best algorithm to address a particular problem. This decision is often influenced by the amount and quality of the data available, the computing resources, a target training time, and desired performance. Recent publications led by Jimenez-Carvelo [61], Modaresi [62], Reichenbach [63], and Qin [64] provide a critical assessment of some of these strategies, when applied to the analysis of food, water, wine, and fish samples, respectively. Although the selection of the best algorithm to address a problem is often performed by preliminary comparative experiments (as exemplified in Refs. [65–67]), this section presents an overview of the most common algorithms currently used, presented in increasing order of sophistication. As a reference, Table 1 provides a brief classification of common algorithms, followed by a brief discussion of the most relevant for analytical applications (marked with *). A more comprehensive list is provided as Supplementary Material where, for example, reinforcement learning is included. It is important to note that these are not exclusive approaches and many groups have reported alternative ways to implement these algorithms.

3.3.1. K nearest neighbors

This is a well-known and simple algorithm, mostly used for classification of datasets, although its application for regression outcomes [61,65,68] is also feasible. The training model consists in a multidimensional query of pre-labeled data, defined by the number of nearest neighbors (k parameter) and the discriminating distance to be used. The functionality for classification of an unknown data is based on non-parametric method, in which a mathematical tool (Euclidian distance) is used to classify an object on its neighborhood according to the k most similar instance (Fig. 2).

Albeit simple, this strategy has allowed to classify representative elements into either metallic or non-metallic [69], to predict the melting point of 4119 structurally diverse organic molecules and 277 drug-like molecules [70], and to discriminate between active and inactive cyclooxygenase-2 (COX-2) inhibitors [71], among others [62,64]. In line with these examples, it is generally accepted that despite the simplicity and potential of K Nearest Neighbors to discover complex decision boundaries with very large datasets, random variations in smaller datasets used during the training

Table 1
Selected examples of machine learning algorithms and the corresponding learning approach.

	Supervised and self-supervised learning			Unsupervised learning	
	Regression	Classification	Generation	Clustering	Dimensionality reduction
Traditional Machine Learning					
Linear Regression	X				
Logistic Regression	X				
K Nearest Neighbors (k-NN)*	X	X			
Support Vector Machines (SVM)*	X	X			
Decision Tree (DT)	X	X			
Random Forest (RF)*	X	X			
Boosting Algorithms *	X	X			
K Means				X	
Principal Component Analysis (PCA)					X
Linear Discriminant Analysis (LDA)		X			X
Neural Networks (NN)*					
Multi-Layer Perceptron	X	X			
Convolutional NN	X	X	X		
Recurrent NN		X	X		
Transformer		X	X		
Autoencoder			X	X	X

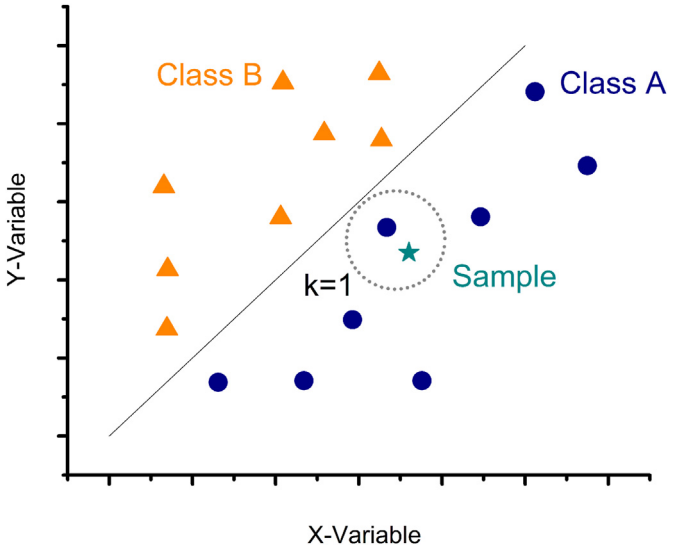


Fig. 2. A graphical representation of k Nearest Neighbors algorithm applied to classify an unknown sample. Note that the example specifically includes two (equivalent) classes of data, the ideal scenario for this approach.

steps as well as the presence of outliers (weighed equally than good points) typically lead to inaccurate predictions [72].

3.3.2. Support vector machines

Support vector machine (SVM) algorithms are mostly used to perform classification tasks, but with a higher complexity than those addressed by K-Nearest Neighbors. This technique is based on a best fit of a hyperplane which separates two (or more) distinct groups [73]. Here, the dataset is first split into a training and a testing set and the goal is to develop a model with the ability to find a decision boundary (hyperplane) with maximum distance to the closest points of the training set (support vectors) [74]. The main difference with respect to K-Nearest Neighbors is that SVM are able to assign more importance to the points in the vicinity of the boundary and therefore, they feature a much higher tolerance to heterogeneous datasets. With some considerations, SVM concepts can be also applied to regression problems [75,76]. The application of this algorithm can be used for linear separations (Fig. 3A) or nonlinear separations (Fig. 3B). In the first case, the groups can be divided by using a hyperplane (line) in only two dimensions. On other hand, nonlinear separation requires the use of kernel equations to convert a two-dimensional space into a three-dimensional space aiming the use of the optimal hyperplane (n-dimensional line).

This ability of SVM to model non-linear relationships has allowed their extensive use for sensing [77], and spectroscopy applications [78]. Specifically, this approach has allowed predicting mechanistic details about multiple-pathway chemical reactions [79] and solubility of drugs [80], classifying active or inactive compounds to prioritize screening [81], and provided objective means to perform classification of coffee [82], wine [83,84], and illegal drugs [85]. In addition, Ghasemi-Varnamkhasti et al. [86] combined principal component analysis (PCA), linear discriminant analysis (LDA) and a support vector machine to reach 100% accuracy in the binary classification of beers using an e-nose.

3.3.3. Random forests

A random forest is an algorithm based on the bootstrap aggregation ("bagging") method, generally used to solve classification and/or regression problems. The approach derives from the ensemble of multiple decision regression trees (featuring lower overfitting), approach that has been used for several chemical predictions [87–90]. This algorithm has building blocks denominated as decision trees (DT) and, each one of them, could be considered an independent learning model with low bias and high variance (Fig. 4). The overall functionality consists in feeding the decisions trees with bagging samples (rows and features) from the dataset and/or training model to infer predictions according to a set of questions. Each set of questions will lead to a distinct pathway and, consequently, multiple possible outputs. Regarding the outputs, the final class is chosen based on majority voting for classification tasks, while the average output from the decision trees is used to solve regression problems. Even though each decision tree has high variance, the sum of all of them yields low variance which makes this algorithm powerful and widely used for numerous applications in machine learning. To gauge the potential complexity of the algorithm, a set of 500 random decision trees was applied to process differential mobility spectra to quantitatively determine condensed phase physicochemical properties of molecules [91]. Despite its advantages, one should consider balancing the size of the dataset vs the structure of the decision tree and the number of possible outputs to minimize bias.

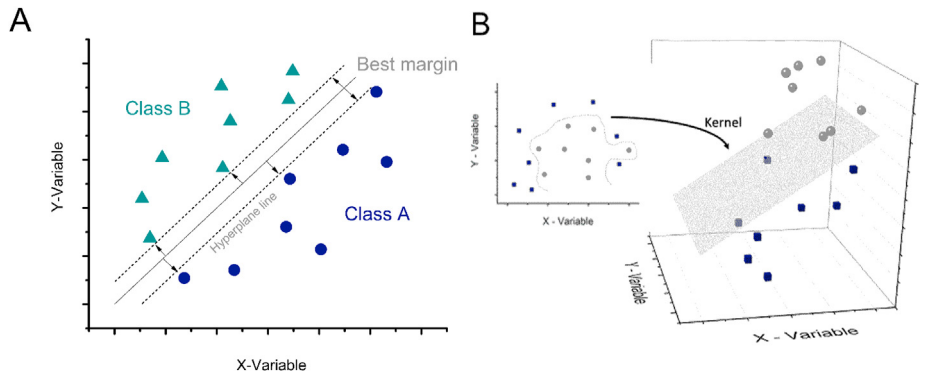


Fig. 3. Application of a support vector machine towards a linear (A) and non-linear (B) separation or two populations.

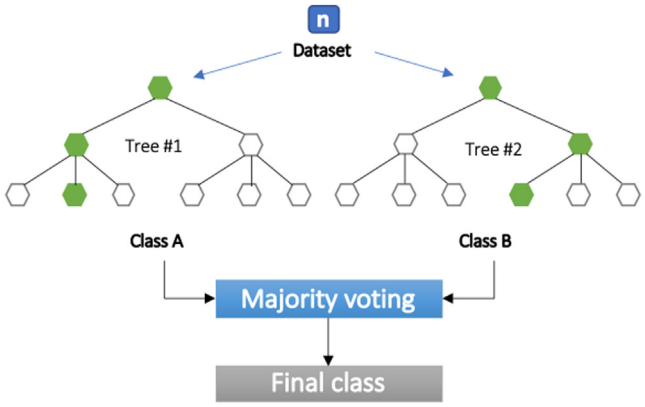


Fig. 4. An example of Decision tree algorithm and it is general functionality.

As expected, this approach allows tackling even more complex problems with many more variables and provides the means to predict the activity of drugs against cancer cells [92], the taste of chemical compounds [93], and the liver toxicity of chemicals [94]. Random forests have also been recently applied to interpret sophisticated Raman spectra of biological samples [95], to evaluate infrared spectra for food adulteration detection [96], to classify chemical threat agents using 2D-GC with time-of-flight MS detection [89,97], to remove errors in large-scale lipidomics [98] and to analyze metabolomics data [99].

3.3.4. Boosting algorithms

Boosting algorithms are a class of related methods including the popular and powerful AdaBoost and Extreme Gradient Boosting (XGB) mechanisms. What they have in common is that they build ensembles of smaller, less sophisticated mechanisms which together compose much more powerful structures [100]. This looks similar to the 'bagging' used in Decision Trees, but rather than randomly creating decision trees and averaging or voting on outputs, these algorithms use mathematical calculations to determine where best to cleave the decision space in order to improve accuracy. As a result, algorithms such as XGB usually improve on Decision Trees for complex tasks. Like decision trees, in operation they are completely deterministic and therefore easier to interpret than neural nets. In fact, while the world of AI has been moving strongly toward neural nets for the past few years, boosting algorithms are the last holdout of traditional machine learning algorithms for competition-level state-of-the-art performance. As of early 2021 these algorithms are still the most popular choice for top competitors working with data other than imagery and natural language; and can be used very efficiently – often more efficiently than a neural net - to achieve state of the art results on many tasks including the quantitative analysis of wheat maltose [101] or the evaluation of beef quality attributes [67] example that was used to demonstrate the importance of matching the algorithm with the problem.

3.3.5. Neural nets

A neural net (more formally referred to as an *Artificial Neural Network* or *ANN*) can be defined as a processing data method that resembles the way human brains solve problems. Neural nets can be applied to many types of problems including classification, regression, and pattern recognition [80]. The primary building block of this system is denominated a neuron, where each one has an input, an ability to mathematically process values and, give an output. Concerning the system's general architecture, the layered structure is the most common one, where the neurons are linked from the input layer, passing through one or more (usually many) hidden layers, to the output layer (Fig. 5A). It is the capacity of these networks to process non-linear activation functions that allows learning, via complex interactions between the neurons. In fact, it has been demonstrated that this combination of linear and non-linear functions can (theoretically) be used to approximate any function, to any level of accuracy – a law known as the Universal Approximation Theorem [102].

Neural nets, therefore, can be used to tackle almost any problem in data science. They are not always the best choice, especially in cases of very limited data or when it is imperative that the functioning of a model be fully deterministic and explainable. That said, and as a general rule, they are the most universally-useful tool available. For this reason, and because they are relatively easy to learn and apply, we will focus primarily on neural nets for the remainder of the review. In the simplest case, a neural net is composed of the input layer, one hidden layer (processing the information), and the output layer (providing the results) - a relatively rare structure called a *shallow* neural net. More commonly, systems that integrate two or more hidden layers are called *deep* neural networks (DNNs). By processing the information through multiple hidden layers, these structures can address more complex problems than *shallow* neural nets, though they often require larger datasets for training [50,52,103]. The general practice of using deep neural nets is often referred to as Deep Learning (DL) [29,44,48,104–106].

According to the arrangement of neurons in each layer and between layers, many types of architectures can be constructed. The simplest, in which every neuron of each layer is connected to every neuron in the layers above and below, is traditionally called a *Multi Layer Perceptron* (MLP), though this arrangement is most

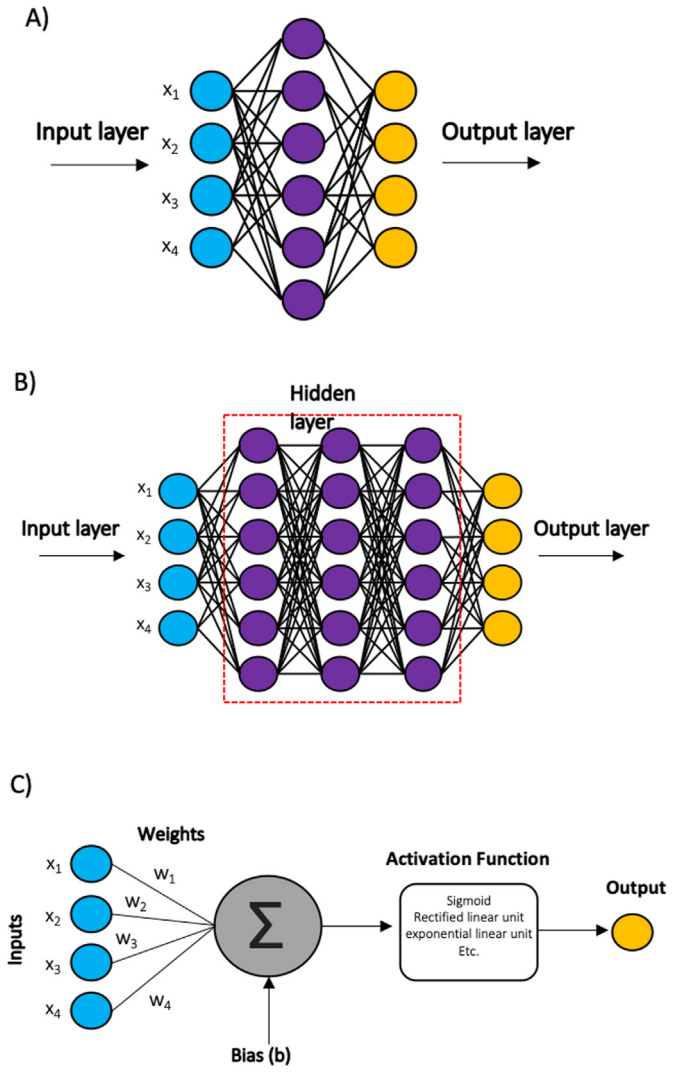


Fig. 5. A single (A) and multiple (B) layer's structure of an Artificial Neural Network, and the functionality of each neuron (C).

commonly mixed in with other types of layers and simply called *fully connected* or *dense* layers. MLPs are still a favored choice when working with tabular data, especially when using *embeddings* to represent categorical data. The most mature and certainly one of the most powerful architectures is the Convolutional Neural Net (CNN), which is used mostly for 2-D images [107] but has been applied successfully to 1-D and 3-D data and to many analytical tasks [108–112] where translation invariance is desirable. Time-series and other sequential data such as written text have mostly been handled by Recurrent Neural Nets (RNNs) [113], but recently these have largely fallen out of use in favor of *transformers* [114], a class of architecture which makes heavy use of *attention layers* of various types. MLPs, CNNs, and transformers can handle practically any supervised or self-supervised use case likely to occur in analytical chemistry. For unsupervised learning, *Autoencoders* are a class of architecture which can perform dimensionality reduction and clustering (for example, in LC-MS data [115] or to increase the LOD in gas sensors [116]). It is worth bearing in mind that there is an unlimited number of neural net architectures, and many useful architectures are mixtures of the types described above.

Another important aspect of neural nets is that these systems can be readily adapted to new data without the need for extensive Feature Engineering, the often-cumbersome practice of input data manipulation and enhancement, endemic in most other techniques. Therefore, and for exemplification reasons, the following discussion is centered on deep learning methods in which several hidden layers are interrelated in a logical sequence (Fig. 5B). Regarding the functionality: for each layer, the outputs of the previous layer are multiplied by a matrix of Weights and then a Bias value is added to produce the initial output. After this linear math function, a nonlinear mathematical operation called an Activation Function must be applied before passing the output to the next layer. (Fig. 5C). This may be any almost nonlinear function, such as a Sigmoid function, a Logistic or Exponential Linear Unit; but in current practice it is usually a Rectified Linear Unit ('ReLU'), which simply replaces any negative values with zero. A forward propagation will occur through all the layers until the output layer that is going to provide a certain value. In other words, the weights and biases will directly impact on the activation value of each neuron and, consequently, on the final output.

When training a neural net using Supervised Learning, each training example is fed into the neural net, and then the output of the neural net is compared to the label attached to the input example. This comparison is done by what is called the Loss Function and when aggregated across many training examples, it is a calculation to estimate how good the neural network model is: the lower the lossfunction's value, the better the neural net. In this way, the direct purpose of training the neural net is to make the loss function as low as possible. Recent advances in computer science have led to programming languages which can automatically calculate the gradients of complex multidimensional functions. By automatically finding the negative gradient of the loss function, the weight and bias parameters can be incrementally adjusted to reduce the loss. This mechanism of aggregating the loss function over a number of training samples, then finding the gradient of that loss function and incrementally updating the weights and biases to descend the gradient is called Stochastic Gradient Descent (SGD), and is the basis of most Neural Net training.

The many weight and bias parameters of a deep neural network are analogous to a huge multi-dimensional space. For visualization purposes to understand how neural net training works, this high-dimensional space can be represented as a three-dimensional landscape where the loss function represents the height of the surface in the z dimension and the weights and biases are represented by the x and y dimensions (Fig. 6A, also see Ref. [117] and video visualizations at https://losslandscape.com/). In these visualizations, following the downward slope of the landscape by incrementally changing x and y gradually – like a ball rolling down a complex and bumpy hillside - reduces the loss function's value towards a minimum that represents the lowest difference between the training data set and unknown data (Fig. 6B).

One of the challenges in neural network training is that there may be many local minimum values in the landscape where the "ball" can roll no lower in the immediate vicinity, and yet much lower minima exist elsewhere. Complex algorithms have been developed for avoiding getting stuck in these "local minima" – should the ball have momentum? Should it start with small steps, gradually move faster, and then slow down again? Should it leap forward for several aggressive steps and then jump back to an average of the previous steps? Modern AI libraries such as PyTorch (https://pytorch.org/) and TensorFlow (https://www.tensorflow.org/) provide researchers the tools required to experiment with many of these techniques, but most practitioners can feel comfortable simply using the default settings which will generally do a good job for most tasks when not trying to break any records for training time and accuracy. Will your "ball" eventually find the true "global minimum"? In a neural net with many layers and

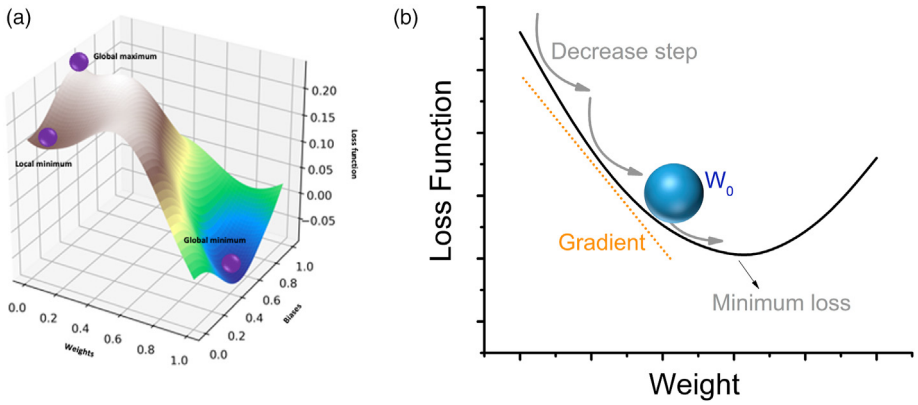


Fig. 6. Three-dimensional shape of gradient descendent (A) and its representation as a function of two variables (B).

thousands to millions of weight and bias parameters, there is generally no way to know … the more practical question is: Is the neural net accurate enough to do what we want it to do?

To answer these questions, it is important to understand that the adjustment to weights and biases is done thousands of times, with each adjustment being a single small step. But how can individual adjustments to thousands or millions of parameters be calculated from each training sample? This adjustment is done by a Back Propagation algorithm: the loss function is calculated for the training samples in batch, and an adjustment is made to the very last layer of the neural net to adjust its parameters slightly in the direction of lower loss. Then layer by layer, this update is continually applied backwards through the neural net using a simple built-in matrix calculus function (based on the Chain Rule of differential calculus) to determine how much of an update to apply to each weight and bias parameter at each layer.

As the size of the dataset increases, so generally does the accuracy of trained model and its ability to make accurate predictions (i.e. classifications or regressions) on previously unknown inputs. The recent breakthroughs in neural nets are due to Deep Learning's strengths, such as it is ability to provide high quality results from a large dataset without the need for traditional feature engineering and for its practical applicability to multiple real-world domains. Thus, neural nets have been extensively used in conjunction with analytical applications and applied towards the classification of olive oils [118], prediction of flash points of pure compounds [119], properties of materials [106] and to support various spectroscopies [109,120,121]. The added complexity of neural nets has also allowed discriminating myeloma patients using MS data [122] or treating chromatograms [123]. Additional applications linked to food industry [41,61,118,124,125] have also been recently reviewed.

4. Teaching chemistry to computers

4.1. Sources of data and libraries

There have been multiple attempts to incorporate AI next to the lab benches [126]. However, many of the past efforts can be considered outstanding examples of computer algorithms, programmed to execute a pre-defined set of logic steps to search for a pre-programmed solution [50]. In contrast, and considering that AI systems are trained (not programmed) to provide a probabilistic answer, most of the breakthroughs of AI in chemistry can be traced back not only to recent developments in neural nets (and deep learning in particular) but also to the increase in the volume and quality of chemical datasets [37,61,127–130]. It is important to state that having access to unbiased, structured data (or a way to extract data from unstructured sources) has been critically important not only for analytical chemists [103,131] but also for those interested in the design of models with the capacity to optimize methodologies and to predict reaction schemes [132,133]. In this regard, and while a handful of reports have used relatively small databases (1026 odorants [134], 981 primary and 852 secondary metabolites [135]) most machine learning approaches require datasets with (hundreds of) thousands of unique entries [136–138].

Highlighting the importance of locating and mining important information from large databases, Buryak's group [110] recently reported a deep convolutional neural network (CNN) that is able to rank candidates for the EI-MS library search (low resolution spectra) in the NIST 17 database (retention index data and GC conditions for 99,400 compounds on various columns, including MS spectra for 72361 of them; https://chemdata.nist.gov/). While many of these datasets are publicly available, many of them are associated with AI platforms such as RXN for chemistry (https://rxn.res.ibm.com, a free AI machine powered by IBM with the capacity to predict organic chemical reactions); Pistachio (https://www.nextmovesoftware.com/pistachio.html, a platform able to extract synthetic conditions from unstructured text in databases) [130] ChemOS [139] (a software package supporting material science and drug discovery); or DeepMind (https://deepmind.com, a collaborative platform with broad applications, including quantum chemistry [140] and protein folding [141]). In addition to commercial tools such as MatLab [142], MoleculeNet [143] (http://moleculenet.ai/) allows testing machine learning methods of molecular properties, integrated as parts of the open source DeepChem package (MIT license). One remarkable aspect of this project is that they also present a collection of databases, with information related to quantum mechanics, physical chemistry, biophysics, and physiology. These repositories allow users to develop applications beyond those traditionally used in analytical chemistry, several of them summarized in Table 2.

Numerous open-source programming libraries are also available to make these data sources accessible to chemists, although with some programming ability is often required. One particularly good example is the PubChemPy library (https://pubchempy.readthedocs.io/) which makes it easier to query PubChem directly through Python. It is also important to note that a common practice among data scientists is to work with datasets in interactive environments such as Jupyter Notebooks, which enable interactive experimentation (the process known among data scientists as Experimental Data Analysis) of data one or a few lines of programming at a time. As a brief example, the Supplementary Information section describes the steps required to apply the PubChemPy library to use a database to find synonyms for a compound name.

Table 2
Selected examples of searchable databases to obtain (input) chemical information.

Name	Remarks	Reference/Link
GDB databases	Small molecules	[144]
Chemical Abstract Service	Maintained by ACS, includes SciFinder, patents, methods, etc.	https://www.cas.org/
PubChem	Maintained by NIH, includes chemical and physical properties, biological activities, safety, and toxicity, etc.	https://pubchem.ncbi.nlm.nih.gov/
Chemistry webbook	Maintained by NOST, includes spectra, properties, etc.	http://webbook.nist.gov/chemistry/
Protein Data Bank	Protein Structures in 3D formats	http://www.rcsb.org/
ProtDataTherm	Thermostability analysis of proteins	[145]
ZINC	Commercially-available compounds for docking as well as search for analogues	http://zinc.docking.org/
ChemSpider	Text and structure search, includes references, properties, spectra, and suppliers	http://www.chemspider.com/
Merck Index	Maintained by RSC, properties, reactions, structures, constants and conversions	https://www.rsc.org/merck-index
NOMAD CoE	Applied to materials science	https://www.nomad-coe.eu/
Spectral Database for Organic Compounds SDB	NMR, IR, Raman, etc. spectra	http://sdbs.riodb.aist.go.jp/sdbs/cgi-bin/cre_index.cgi
MassBank	Public repository for sharing mass spectral data for life sciences	http://www.massbank.jp/
Crystallography Open Database	Organic, inorganic and complex crystal structures	http://www.crystallography.net/
United States Patent and Trademark Office	Patents	http://patft.uspto.gov/
eMolecules	Chemical structure	http://www.emolecules.com/
Chemical Structure Lookup Service	Bioactivity, metabolite, patents, drug development, imaging agents, crystal structures, natural products, reaction databases, toxicology and environmental data	http://cactus.nci.nih.gov/cgi-bin/lookup/search
CRC Handbook of Chemistry and Physics Online	Physical and chemical properties, structure, safety, etc.	http://hbcponline.com/
NextMove Patent Reaction Dataset	Reactions extracted from patents from 1976 to 2016	https://depth-first.com/articles/2019/01/28/the-nextmove-patent-reaction-dataset/
Malaria Drug Accelerator	Database linking phenotypic hits to function for malaria.	http://winzeler.ucsd.edu/malda/

4.2. Descriptors

There are multiple ways to represent chemical structures in a way that computers can learn from them. Probably the most common one is SMILES (Simplified Molecular-Input Line-Entry System) [53,54]. This is a rather simple notation in which molecules are represented by a string of characters representing atom and bonds, without including hydrogen atoms. Because this notation uses letters (to represent atoms), parentheses (to represent branching points) and numbers (to designate ring connection points), it can be also interpreted by (some) humans. For example, the characters -, = , #, *, and . represent single bond, double bond, triple bond, aromatic bond, and disconnected structures, respectively. Thus, a molecule of phenol is represented by C1=CC=C(C=C1)O. One limitation of this notation is, however, that it doesn't encode chirality. In cases where chirality is important, the International Chemical Identifier (InChI, https://www.inchi-trust.org/) can be used. Note that the PubChemPy library described above is also capable of searching for compounds using their SMILES and InChI representations. Although their specific implementation is not discussed in detail here, these strings of characters can be processed into chemically-related information for use in various neural net architectures, including transformers [114]. Building from language-based solutions, these neural networks may treat reactions (reactants → products) as a translation task [146,147], or a language classification task [148], and are able to process multiple input vectors at the same time [49]. In fact, Winter et al. [149] described a system that was able to translate between two semantically equivalent but syntactically different representations of molecular structures, compressing the meaningful information both representations have in common in a low-dimensional representation vector. It is also important to mention that there are multiple ways to perform these transformations, based on numerical parameters and/or graphical interpretations [128,150–155]. Again, and despite the decades of development [156,157], many of these systems have found their way mostly into the description or prediction of organic synthesis [15,49,133,158,159].

4.3. Output interpretation

A side effect of the fast expansion of the predictive capabilities of AI has been the ability to adopt some of these technologies without paying too much attention to how the learning process is implemented and what is the chemical link (reactivity) between the input and the output. In such cases, the AI method behaves as a dangerous black box that is able to chew data and spit out answers without providing meaningful information about the system [65,66,160]. This issue is not exclusive to our field [161] and addressing it (even as a grey box [77]) is particularly important for projects focusing on reaction schemes [45,60,79,154]. On the other hand, treating the algorithm as a black box could be reasonable for analytical problems where the output itself (classification, calibration, signal improvement, etc) is more important than understanding how the model derived it. This could be the case of an algorithm predicting results based on a well-established chemical mechanism and under limited experimental conditions. From our perspective, and beyond the philosophical issue of not knowing how and why the prediction was made, this approach could lead to a) limited ability to rationally troubleshoot the process, b) potential mistakes during generalization, and c) inability to identify and mitigate bias. The latter is particularly important because (for instance, and unless specifically corrected), predictions extracted with one machine learning approach (Mol2vec, IVS2vec) could carry-over an unrelated bias embedded in the original algorithms (Word2vec, [162]).

5. AI in analytical chemistry

AI algorithms are beneficial as substitute methods to conventional approaches or as components of embedded systems. Despite their advantages and limitations towards solving analytical problems [130], AI has been applied for optimization and data interpretation of several analytical approaches including biosensors,

microfluidics, chromatography, mass spectrometry, Raman, hyperspectral imaging and NIR. Here, we provide a few examples of AI applied to different analytical applications considering that the list is likely to be incomplete and will probably become outdated soon.

5.1. Colorimetric sensors

Integration of AI approaches including pattern analysis and classification algorithms with sensors can bridge the gap between the data acquisition and analysis, resulting in a new and revolutionary cross-disciplinary research area [163]. In this regard, computers can now easily identify patterns of shapes and measure color intensity greatly increasing the throughput and accuracy of various sensing approaches. For example, Jiang et al. [164], applied an extreme learning machine (optimized after including 20 hidden layers) and an evolutionary algorithm (available at https://www.ntu.edu.sg/home/egbhuang/elm_kernel.html), to the quantitative determination of tea polyphenols using a homemade color sensitive sensor composed of nine color spots. Similarly, Kim et al. [165] was able to enhance the selectivity of sensors developed using M13 bacteriophages modified with a surface protein (pVIII). Upon demonstrating that these pages generate unique color patterns when interacting with certain analytes, they implemented an automated clustering approach through RGB distance values. Luo et al. [166], applied a machine learning model to the analysis of total organic carbon (TOC) using inkjet-printed colorimetric sensors. The machine learning model trough pattern recognition describe the relationship between the sensor and TOC value. Using a combination of machine learning and colorimetry, Duan et al. [167], presented a strategy based on color-spectral images to discriminate mixtures of amino acids. Authors evaluated the performance of six common convolutional neural networks (LeNet, Vanilla CNN, Residual Network, SqueezeNet, VGGNet and GoogLeNet Inception v1) and concluded that the VGG model provided the most convenient compromise between the architecture (number of layers and channels), accuracy, and speed. The same group also applied a similar approach for the analysis of pollutants (mercuric chloride, lead nitrate, tetracycline, nickel sulfate, cupric nitrate and chromic chloride) and found that despite the sensitivity issues, the Inception v1 model displayed a better astringency and a higher accuracy [111]. A significant improvement was recently discussed by Yu et al. [168] that implemented a convolutional neural network model (two hidden convolutional layers, two subsampling layers, a fully connected layer, and an output layer) to treat videos compressed into static images. In this way, they preserved the morphology and motion of single cells and lowered the computational demands. Similar approaches describing the application of AI towards the development of holographic biosensors [169] and fuse terahertz spectroscopy and imaging [101] have been also described.

5.2. Electronic noses

These sensing approaches take advantage of AI algorithms for mimicking biological olfactory systems, connecting a particular sample with a signal pattern via a pattern recognition algorithm [134,170]. From the chemical standpoint, the approach is based on the use of sensors with low selectivity and thus the dynamic signal can be affected by the analyte (structure, reactivity, concentration, etc.) and the conditions (type of sensor, temperature, concentration, flow, pressure, interferences, etc.) thus generating rather complex patterns [171]. As an example of the capacity of AI to aid in the assessment of beer quality, Gonzalez-Viejo et al. [172] integrated 17 commercial sensors with a two-layer feedforward network model, which is composed of a tan-sigmoid function in the hidden layer and a linear transfer function in the output layer. Another interesting e-nose approach was also presented by Hayasaka and co-workers [173], where instead of using a sensor array they used a single graphene field-effect transistor (GFET) and the power of machine learning to achieve gas selectivity. To accomplish this, the conductivity profiles were recorded, decoupled into four distinctive physical properties generated when gas molecules interact with the graphene and finally projected onto a feature space as 4D output vectors. This approach was able to quantitatively classify water, methanol, and ethanol with high accuracy when tested individually. Systems based on other sensors (quartz crystal microbalance [174]) and with different applications (beer [86], waste-water [175]) have been also recently published.

5.3. Biosensors

Smartphone-based sensing systems have received special attention with the worldwide popularity of smartphones. Due to the integration of numerous sensors and functions such as processing and communication, smartphone-based platforms are now coupling AI approaches and gaining importance in analytical chemistry [163]. Rodríguez-Rodríguez et al. [176] proposed the possibility to forecast glucose levels in diabetic people by executing predictive algorithms locally using portable devices such as smartphones and a Raspberry Pi. While the idea of running highly-demanding algorithms coupled to a wearable amperometric glucose sensor (Freestyle Libre, [177]) is challenging due to the restricted computational resources, limited battery power, and risks of being disconnected from the Internet; they found that it is possible to do predict blood glucose levels with a root mean squared error of 11.65 mg/dL in just 16.15 s, employing a 10-min sampling of the past 6 h of data (interstitial-glucose levels) and a Random Forest. With the Raspberry Pi, the computational effort increases to 56.49 s, but authors stated that this can be improved to 34.89 s if SVM are applied achieving an error of 19.90 mg/dL. Besides these examples, it is worth mentioning the applicability of intelligent approaches to enhance the power of other electrochemical biosensors designed for the analysis of prostate cancer antigen [178] or *e. coli* [179]. A few additional examples can be found in a recent review [180], specifically focused on connecting machine learning with biosensors.

Linked to a common reaction pathway and taking advantage of the connectivity and remote processing ability of smartphones, Solmaz et al. [181] presented a custom smartphone application ("ChemTrainer") based on machine learning algorithms for automatic quantification of peroxide content on colorimetric test strips. The most interesting aspect in this development is that the app captures, crops, processes the active region of the strip, and then communicates with a remote server that contains the learning model through a Cloud-hosted service. LS-SVM and Random Forest classifiers were fed with the mean RGB, Hue-Saturation-Value (HSV) and labeled values of the test strip image under various illumination conditions. The model detects the color change in peroxide strips with over 90% success rate. Advancing this concept, Draz et al. applied a convolutional neural network (using Inception v3 architecture, which was transfer-learned using Google's TensorFlow framework) to detect various virus (hepatitis B, hepatitis C, and Zika) following just the visual patters of gas bubbles using just a cell phone [182].

The 2019 SARS CoV-2 (COVID-19) pandemic has illustrated the need for rapid and accurate diagnostic protocols. In this sense, biosensors play an outstanding role for detection of COVID as can be inferred trough the large number of publications in this area in the last months. Nevertheless, the combination of AI with biosensors could bring outstanding advantages such as automation, patients'

classification, high processing capacity and fast statistical analysis. Cady et al. [183] developed a multiplexed grating-coupled fluorescent plasmonic (GC-FP) biosensor for measure antibodies against COVID-19 in human blood serum and dried blood spot samples. Also, a support vector machine based machine learning approach was developed to score patient samples for prior COVID-19 infection, using antibody binding data for all three COVID-19 antigens used in the test. Recent studies have shown that one important feature of COVID-19 is the abnormal respiratory status caused by viral infections. Taking advantage of this, a portable non-contact method to screen the health conditions of people wearing masks through analysis of the respiratory characteristics from RGB-infrared sensors was proposed [184]. The authors developed a portable screening device to get the thermal and RGB videos from target people. In order to extract respiration data from faces in thermal videos, face recognition method to capture people's masked areas was applied. Then a bidirectional GRU neural network with an attention mechanism (BiGRU-AT) model to work on the classification task was used. The results of validation experiments show that the model can identify the health status of respiratory with 84% accuracy, 90% sensitivity and 76% specificity on the real-world dataset.

5.4. Mass spectrometry

Because MS is one of the most utilized detection techniques in analytical chemistry, several AI algorithms have been applied for classification and interpretation of spectra [67,122]. Along these lines, Hwang et al. [185] reported the classification of N-glycopeptides as core- and outer-fucosylated types using LC-MS/MS and machine learning algorithms such as deep neural networks (DNN) and support vector machines. Following the workflow described in Fig. 7, they reported an accuracy of more than 99% against manual characterization for 82 fucosylated N-glycopeptides (54 core, 24 outer and 4 dual) from 22 glycoproteins.

Other examples of machine learning approaches to evaluate metabolomic data have been also recently reported [67,154,155,186]. Not only large proteins but also small molecules have been targeted by MS-AI. Jang et al. [187] developed a software named AI-SIDA (artificial intelligence screener of illicit drugs and analogues) for screening unknown erectile dysfunction drugs and analogues using LC–MS/MS. AI-SIDA consists of three different layers: LC-MS/MS viewer, AI classifier, and Identifier. The second AI classifier layer, an artificial neural network classification model, was constructed by training 149 LC–MS/MS spectra (including 27 sildenafil-type, 6 vardenafil-type, 11 tadalafil-type ED drugs/analogues and other 105 compounds). This model was found to show 100% classification accuracy for the 187 LC–MS/MS modeling and test data sets. Donati's group [188] demonstrated the usefulness of supervised and unsupervised machine learning methods to accurately evaluate matrix effects, one of the parameters that most affect analytical signals. Although their interesting study is useful and complete for the prediction of the effects caused by carbon and easily ionizable elements on inductively coupled plasma optical emission spectrometry (ICP-OES), it is of upmost importance because it opens the door to start considering AI strategies to address different effects, including severe matrix effects of real world samples particularly affecting extractions, sample injection in different instrumental analysis techniques and detection in mass spectroscopy.

5.5. Vibrational spectroscopy techniques

There are several reports on the use of AI involving spectroscopy techniques such as visible (VIS), near-infrared (NIR), mid-infrared (MIR), fluorescence, Raman, and nuclear magnetic resonance (NMR). As recently described by Yang et al. a number of studies have already demonstrated the advantages of deep neural networks to extract critical patterns from raw spectra, significantly reducing the demand for feature engineering [120]. Among those, a number of applications have focused on obtaining classification (geographical, botanical, etc.) data and managing the large amounts of information obtained by hyperspectral imaging. Hyperspectral imaging (HSI) techniques have become a very popular and powerful tool to inspect food and agricultural products, making their combination with AI algorithms quite natural. In this sense, Quin et al. [64] developed multimode hyperspectral imaging techniques to detect substitution and mislabeling of fish fillets. Using reflectance in visible and near infrared region, fluorescence, reflectance in short-wave infrared region, and Raman in combination with 24 machine learning classifiers in six categories (i.e., decision trees, discriminant analysis, Naive Bayes classifiers, support vector

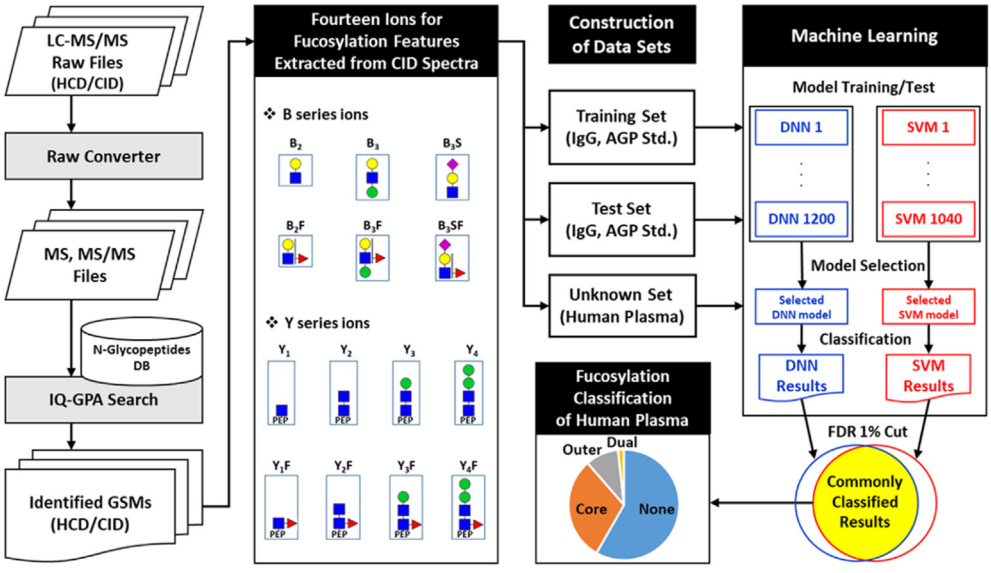


Fig. 7. The computational workflow for classifying the fucosylation of N-glycopeptides using machine learning. Extracted from Ref. [185].

machines, k-nearest neighbors classifiers, and ensemble classifiers) were able to examine fish species and freshness. The different combinations of machine learning classifier, spectral type, and dataset provided an intuitive way to compare their performances and identify the best combination. Recently, Near Infrared Hyperspectral Imaging (HSI-NIR) combined with machine learning for supervised detection and classification of *Cannabis sativa L.* was demonstrated [189]. Samples of leaves of *Cannabis sativa L.* and containing other similar plants were used to simulate an illegal plantation. Their approach was able to recognize cannabis plants under natural conditions with only four spectral bands. Spectral images with chemical mapping analysis can provide chemical and spatial information; low-cost drone imaging systems could significantly contribute to forensic, food and environmental sciences. In additional to other developments linked to the use of hyperspectral spectroscopy for the detection of bacteria [190], Ghaffari et al. recently presented a protocol to select only the samples and variables in a data matrix that carry essential features, reaching reduction rates of over 99% within seconds [191].

In addition to the previously-described examples of Infrared Spectroscopy coupled to AI for biosensing purposes [151,152], the combination of IR-AI has been also applied to identify the geographical origin of several products, thus avoiding frauds and adulteration tactics [96]. Bona et al. [82], applied near infrared (NIRS) and mid infrared spectroscopy in combination with support vector machines to obtain geographical classification of different genotypes of arabica coffee, leading to the correct classification of all samples with a sensitivity and specificity of 100%. In a similar way, a method for discriminating the origin of the medicinal plant *Tetrastigma hemsleyanum* was proposed using NIRS and deep learning models. In this case, the convolutional neural network (CNN) was faster and more accurate than the tested support vector machine (SVM) and the self-adaptive evolutionary extreme learning machine (SAE-ELM) [192]. A rapid and specific method for the detection of fish allergen parvalbumin was developed based on IR spectroscopy [193]. Inception ResNet (IRN), support vector machine, and random forest models were successfully trained, established and compared using parvalbumin IR spectra from 16 fish species. Although the tree studied models were able to identify the fish allergen parvalbumin, IRN model was selected since it provided the highest accuracy (up to 97.3%). Accurate estimation of soil organic matter (SOM) is essential in understanding and to identify areas that need fertilization. The SOM have been estimated using Visible and near-IR spectroscopy coupled to extreme learning machine (ELM) and support vector machine. ELM models yielded superior predictability relative to SVM models [194]. Due to the growing complexity of these datasets, machine learning (based on *ab initio* molecular dynamics simulations) has also been applied to predict highly accurate molecular infrared spectra with unprecedented computational efficiency [195]. Other applications of AI linked to infrared spectroscopy include the analysis of microplastics [196] and seeds [197] and herb extracts [198].

Gomes et al. [199] studied the combination of surface plasmon resonance (SPR) based sensors with different machine learning techniques aiming to improve and attest to the quality of the real-time SPR responses so-called sensorgrams. They were able to create intelligent SPR sensors to give a safe, reliable, and auditable analysis of sensorgram responses.

Considering the links between DNA damage and the development of diseases, Chen et al. [104] applied a deep-learning-based open-source pipeline (FociNet) to automatically segment full-field fluorescent images and dissect DNA damage of each cell. Beyond the development of the system (deployed using standard scripts in https://www.kaggle.com/keegil/keras-u-net-starter-lb-0-277/comments?scriptVersionId=2164855/notebook, a sigmoid activation function for the last layer, a binary_crossentropy loss function, and Adam optimizer), the work provides a tool to evaluate DNA damage on the basis of microscopy images as well as a potential strategy for high-content screening of active compounds. Not surprisingly, other groups have also pointed to the advantages of using the Adam optimizer [200], especially for spectroscopy data [120]. Also using a configuration similar to the one reported by Chen [104], Cui and co-workers [201] reported the combination of deep learning and three-dimensional (3D) fluorescence difference spectroscopy for rapid identification of mixtures of illicit drugs in biofluids. In this case, they evaluated the performance of several supervised models and concluded that the selected generative adversarial network (containing a 3-layer generator and a 3-layer discriminator) provided the best generalization capabilities. This approach allowed the identification of various drugs considering the signal of the mixtures including codeine, 4,5-methylene-dioxy amphetamine, 3,4-methylene dioxy-methamphetamine, meperidine, and methcathinone; regardless of their fluorescence. Also focusing on fluorescence analysis, Olivieri's group described the possibility to apply artificial neural networks to estimate the sensitivity parameter without extensive computational costs [202].

Because it is one of the most versatile analytical techniques, several groups have explored the possibility to couple AI to Raman Spectroscopy [95]. In this regard, Li et al. proposed a combination of SERS with three variable selected regression methods for the determination of thiabendazole in apple juices, reaching for the best model a recovery range of 83–93% [203]. In this case, the extreme learning machine model was built after the analysis of the nonlinear correlation between SERS spectra of thiabendazole and their corresponding concentrations. Complementary, Zhu et al. [204] proposed coupling SIMPLS with interval selection, model population analysis (MPA), weighted bootstrap sampling (WBS) and soft shrinkage for simultaneously predicting the volume ratios of various pesticides by SERS spectra. Albeit the SIMPLS algorithm should be strictly classified as a chemometric method [205], its combination with MPA/WBS showed good prediction performance and illustrates the potential overlap between ML algorithms and traditional chemometric approaches [61,86,197,206]. The same group [112] also recently combined SERS with a one-dimensional convolutional neural network (1D CNN) for the onsite determination of pesticides and applied this approach towards tea samples. Their results were contrasted with conventional approaches with satisfactory results. Interestingly, the identification of pesticide was performed on the cloud server and then the trained 1D CNN models can be applied for subsequent analysis.

Since CNNs are the most mature type of neural net and are great for any sort of image classification, a number of researchers have applied them for bioanalytical purposes, especially coupled to Raman imaging [108]. Along these lines, Lu et al. recently presented a novel method that uses a convolutional neural network (ConvNet, Fig. 8) to analyze biological Raman spectra and identify the microbes at a single-cell level and with an accuracy of 95%. To accomplish this, authors utilized a clever 10-fold cross-validation where the shuffled random data were split into 10 sets of approximately equal size to perform either testing or training the ConvNet model (repeated 10 times and each of the 10 sets acted as test data once) [207].

Taking advantage of the unique merits of SERS, Shi et al. [208] explored the set up a SERS-based database of deoxyribonucleic acid (DNA), suitable for artificial intelligence-based sensing applications. This database was built using silver nanoparticles (AgNPs)-decorated silicon wafer (Ag NPs@Si) SERS chips, followed by training with a deep neural network. Three representative tumor suppressor genes were discriminated in a label-free manner with a recognition accuracy of 90%.

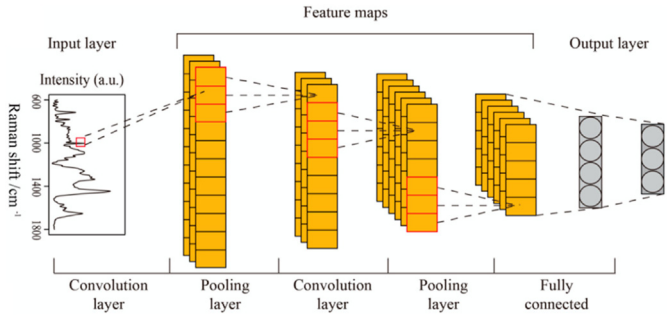


Fig. 8. The diagrammatic sketch of the structure of Convolutional Neural Networks. The ConvNet contains two convolutional layers, two max-pooling layers, and one fully connected layer. Adapted from Ref. [207].

5.6. Separations

Yali et al. [209] developed a quantitative structure–retention relationship methodology that was employed to predict retention times of polychlorinated biphenyl (PCBs) using previously published data set. Multiple linear regression (MLR) and support vector machine methods were applied for generation of linear and nonlinear models. SVM model performed better than MLR which represents the nonlinear relationship between the structural feature and retention time of PCBs on this chromatographic condition. Similar approaches based on machine learning has been also applied to GC-MS data to classify marijuana varieties [210], essential oils [211], and steroids [212]. Following the work from Strozier [97], Tao's group [63] presented a comparison of different machine learning approaches to when applied to classification problems of different degrees of difficulty and using GCxGC data and demonstrated the utility of relatively simple approaches such as SVM. As one of the most recent examples of GC-AI, Vrzal et al. [213] presented the DeepRel (available on https://github.com/TomasVrzal/DeepRel), a user-friendly model with a graphical user interface based on deep learning using SMILES as input to predict retention indexes in GC based on molecular structure. Their approach is extremely useful for non-targeted analyses due to its ability to characterize unknown compounds, even when not present in the libraries.

Multiple attempts have been described to predict and interpret retention times in liquid chromatography. Probably one of the most significant ones was described by Stanstrup et al. [214] who not only proposed the idea of crossing information between systems to predict retention time, but also developed an open database (http://predret.org), where users can upload their own data. This concept has been expanded using an artificial neural network (47 input neurons, 1 hidden layer with 23 hidden neurons, and an output layer with 1 output neuron) for use with MolFind [215], or deep learning approach (4 fully connected hidden layers with 1000, 500, 200 and 100 units, respectively, activated via a non-linearity function (ReLu), connected to an output layer consisting of one unit with no activation) leading to the prediction of retention time of small molecules (±46 s) in nano-HPLC [216]. Additional examples of applications of various neural networks to predict retention times [135], screen for unknown erectile dysfunction drugs [187], improve separations [84,217] or extract features from chromatograms [123], have been also reported.

One area that is surprisingly underdeveloped is the application of AI to capillary electrophoresis (CE). In the last 20 years or so and since the first attempts to apply ANN to CE were published by J. Havel [218], only a few developments have been presented. Among those, it is worth mentioning Jiao et al. [219], who proposed a quantitative structure property relationship (QSPR) to predict the electrophoretic mobility of aromatic acids and Taylor et al. [220], who applied a rather simple ANN (input layer with 1206 neurons, 1 hidden layer with 100 neurons, and one output layer with 5 neurons) to read electropherograms. Although targeting a slightly different problem, Woodruff's group used CE (coupled to MS) to obtain a dataset to demonstrate the potential of machine-learning algorithms to provide concentration estimates for chemicals without the corresponding analytical standards [186].

Last year, a symbolic regression (approach not discussed in the review) was applied to remove artifacts in DNA separations [221]. Despite the great potential of these articles, they have only been cited a handful of times, possibly suggesting that the field is just moving in a different direction or that the perceived simplicity of the technique does not justify an AI intervention.

5.7. Other applications

Song et al. presented an open-source algorithm for the discovery of binding ligands from high-throughput sequencing data of SELEX libraries [222]. (The code of the algorithm (called SMART-Aptamer) is available to the scientific community online at https://github.com/songjiajia2018/SMART-Aptamer-v1.git and could be applied for the discovery of binding ligands for a variety of biomedical applications. Although seldom applied towards analytical problems [77], various machine learning tools have been applied for the analysis of various electrochemical systems [223–225],

6. Implementation of simple AI machines

As stated before, this tutorial review aims to provide basic information to junior analytical chemists considering integrating AI into their programs. Therefore, we provide a general guide to get started with AI, focusing first on neural nets. The approach is recommended because neural nets are among the most powerful AI systems and can be applied to many practical problems with very little education beforehand – in particular, neural nets often greatly reduce the amount of preparatory data manipulation and custom architecture coding compared to other AI algorithms. It has become common for new practitioners of AI to apply powerful deep neural networks to novel (though typically simple) problems within the first few hours of working with them, and several popular AI libraries can enable powerful novel work with only a handful of lines of Python code. There are now very few use-cases that neural nets can't address; so they provide a great balance of broad applicability, tremendous analytical power, and short time to value for the beginner. Readers are encouraged to take one of the several popular free online machine learning courses. Perhaps the most popular is Andrew Ng's excellent course which begins with the fundamentals and is very comprehensive (http://www.ml-class.org/). A very popular recent alternative is Jeremy Howard's Fast.ai course (https://course.fast.ai/) and accompanying python library, which teaches a 'top-down' approach – i.e. starting immediately with practical application of the most powerful neural net techniques and then gradually introducing the fundamentals. This course is less comprehensive, focusing only on the technologies still at the cutting edge today, but presents a much faster path to practical applications of the most powerful techniques, primarily deep neural nets. These courses require only minimal programming experience and basic math knowledge to get started; students with no experience in Python programming are advised first to become basically familiar with that language in order to better absorb the course material mentioned above. In addition, there are additional on-line sources (for instance, https://d2l.ai/) that provide background information and examples. Readers may also find a

practical guide to apply deep learning in spectral analysis in this outstanding review [120]. That said, and while it is possible to train a neural net with just a few minutes of practice (see examples in Supplementary Information), it is important to approach this process with realistic expectations and consider that developments of new and valuable work would generally take weeks or months of work. For those willing to give it a try, a step-by-step tutorial to begin using machine learning is provided as Supplementary Information.

7. Conclusions and opportunities

As shown throughout this review, a common trend observed in the literature is the development of algorithms (of various complexities) to analyze large volumes of data and extract both patterns and meaningful information out of differences (even minute) from individual measurements. Probably this is the main reason supporting the initial developments of AI linked to image recognition, vibrational spectroscopy, and mass spectrometry. Unfortunately, simpler approaches (for example single-analyte biosensors and several separation techniques) may not yet fully justify the effort required to develop the databases needed to train these systems. As these solutions still rely on identifying and utilizing rules specific to each data item at the cost of a substantial human effort, AI systems have already matured enough to address many of these tasks and provide unique opportunities for analytical chemists to seek new developments beyond calibration, noise reduction or traditional recognition of patterns.

While these tasks have been approached by chemometrics, now considered a routine tool for analytical chemists, ML methods are seen (by many) as complicated tools. It is also important to note that despite the clear division made by researchers when describing machine learning and chemometrics methods, both approaches have roots in common and could be mutually supportive. Even though chemometrics and design of experiments clearly excel at method development, analysis of results, and classification tasks;problems involving complex nonlinear analytical issues often require the use of predictive AI. Machine learning approaches can tackle the problem from a very different perspective: merging analytical technologies with larger datasets, performing more complicated analysis, and feeding into automated AI approaches. While the main triumph of the application of AI for complex analytical systems is to enable the development of predictive models, the lack of interpretability of some algorithms, especially neural nets, also carries the potential risk to lose the connection with the fundamental physiochemical phenomenon involved. To alleviate this issue, many cases allow processing some data before their use as inputs to the model. This processing step could be done with a number of processes, including PCA, and can give important information about the most important variables and responses. In this way, a more adequate interpretation of the analytical responses (*e.g.* identification of m/z values related to a chemical species, regions of spectra attributable to functional groups, etc.) can be reached. That said, in the next few years algorithms are likely to overcome these issues and potentially support the analysis to a very high degree through data mining, pattern recognition, deep thinking, and advanced control of experiments.

The ability of AI to elucidate and profit from a blend of multidimensional data will also improve our capabilities to implement smart systems and automated workflows, especially if feedback loops are incorporated in the analytical sequence. Current reports suggest that mining information to build broadly applicable and accessible databases will soon enable interdisciplinary queries, promote collaborations, and transform the way we develop analytical methodologies. This concept is illustrated in Fig. 1, where we anticipate (possibly speculate) that a future role for AI could be to close the loop, enabling autonomous experiments (dotted red arrow) that create empirical 'ground truth' data to fuel other AI work, forming a virtuous cycle of accelerated discovery of new chemical reactivity. While optimizing procedures to get "chemical meaning" from words in data repositories and journals looks a daunting task, the recent release of the Generative Pre-trained Transformer 3 (GPT3) is a huge step forward. This approach is able to interpret human-oriented text (natural language process), read chemical recipes from papers, and provide instructions to robotic instrumentation for chemical synthesis [144]. In such way, a trained model based on a large dataset of analytical procedures might soon be able to improve or even predict the experimental pathway of a certain analysis and reach a higher level of artificial intelligence. An additional trend identified in the literature is the development of interconnected devices. Just like these systems have been able to revolutionize our homes and cars, the Internet of Chemical Things (IoCT) has the potential to change the way we do chemistry. Basically, this concept involves the optimization of internet resources, interconnecting users and manufacturers, analytical devices (lab equipment and portable devices), software suites, mobile devices, and online resources [226]. This has the potential to drastically change the way we think about health care [37,38,227–229], especially if spectrometers [230] and wearable devices [198] become simpler and more affordable, potentially leading to a full network of wearable things [199].

All things considered, the addition of AI technologies into analytical labs is moving forward and despite being in its very early stages, has already been incredibly productive. While it is impossible to predict the impact of AI on analytical chemistry, the community is probably ready to move beyond the interpretation of data and dive into the development of novel algorithms that can not only identify unique chemical features in datasets but also predict more efficient routes of chemical reactivity and foster the development of new analytical methodologies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors would like to acknowledge partial financial support to this project from Clemson University, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and Facultad de Ciencias Agrarias, Universidad Nacional de Cuyo (Mendoza, Argentina). Authors also thank Dr. Hector Goicoechea (UNL) for helpful discussions related to chemometrics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.338403.

References

[1] S. Russel, P. Norvig, Artificial Intelligence: a Modern Approach, third ed., Pearson, 2010, 2020.
[2] A. Pannu, Artificial intelligence and its application in different areas, IJEIT 4 (2015) 79–84.
[3] A. Zhavoronkov, Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry, Mol. Pharm. 15 (2018) 4311–4313.
[4] C.A.M. La Porta, S. Zapperi, Explaining the dynamics of tumor aggressiveness: at the crossroads between biology, artificial intelligence and complex

systems, Semin, Canc. Biol. 53 (2018) 42–47.
[5] M.E. Dilsizian, E.L. Siegel, Machine meets biology: a primer on artificial intelligence in cardiology and cardiac imaging, Curr. Cardiol. Rep. 20 (2018) 139.
[6] E.P.V. Le, Y. Wang, Y. Huang, S. Hickman, F.J. Gilbert, Artificial intelligence in breast imaging, Clin. Radiol. 74 (2019) 357–366.
[7] K. Yao, R. Unni, Y. Zheng, Intelligent nanophotonics: merging photonics and artificial intelligence at the nanoscale, Nanophotonics 8 (2019) 339–366.
[8] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, E.T. Mueller, Watson: beyond Jeopardy!, vols. 199–200, Artificial Intelligence, 2013, pp. 93–105.
[9] H.C.S. Chan, H. Shan, T. Dahoun, H. Vogel, S. Yuan, Advancing drug discovery via artificial intelligence, Trends Pharmacol. Sci. 40 (2019) 592–604.
[10] C.-H. Yu, Z. Qin, M.J. Buehler, Artificial intelligence design algorithm for nanocomposites optimized for shear crack resistance, Nano Future (2019) 3.
[11] J.P. Janet, S. Ramesh, C. Duan, H.J. Kulik, Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization, ACS Cent. Sci. 6 (2020) 513–524.
[12] D. Mrdjenovich, M.K. Horton, J.H. Montoya, C.M. Legaspi, S. Dwaraknath, V. Tshitoyan, A. Jain, K.A. Persson, Propnet: a knowledge graph for materials science, Matter 2 (2020) 464–480.
[13] D.E. Blanco, B. Lee, M.A. Modestino, Optimizing organic electrosynthesis through controlled voltage dosing and artificial intelligence, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 17683–17689.
[14] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.R. Muller, A. Tkatchenko, Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. 6 (2015) 2326–2331.
[15] M. Sumita, X. Yang, S. Ishihara, R. Tamura, K. Tsuda, Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies, ACS Cent. Sci. 4 (2018) 1126–1133.
[16] J.L. Baylon, N.A. Cilfone, J.R. Gulcher, T.W. Chittenden, Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification, J. Chem. Inf. Model. 59 (2019) 673–688.
[17] H. Gao, T.J. Struble, C.W. Coley, Y. Wang, W.H. Green, K.F. Jensen, Using machine learning to predict suitable conditions for organic reactions, ACS Cent. Sci. 4 (2018) 1465–1476.
[18] C.W. Coley, R. Barzilay, T.S. Jaakkola, W.H. Green, K.F. Jensen, Prediction of organic reaction outcomes using machine learning, ACS Cent. Sci. 3 (2017) 434–443.
[19] Z. Hippe, Problems in the application of artificial intelligence in analytical chemistry, Anal. Chim. Acta 150 (1983) 11–21.
[20] G.J. Henderson, Gods and Robots: Myths, Machines, and Ancient Dreams of Technology by Adrienne Mayor, vol. 60, Technology and Culture, 2019, pp. 1100–1101.
[21] A.M. Turing, I.—computing machinery and intelligence, Mind LIX (1950) 433–460.
[22] J. Moor, The Dartmouth college artificial intelligence conference: the next fifty years, AI Mag. 27 (2006) 87–91.
[23] N.J. Nilsson, Principles of Artificial Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1981, p. 112.
[24] J. Weizenbaum, ELIZA—a computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1966) 36–45.
[25] D.G. Bobrow, Natural Language Input for a Computer Problem Solving System, 1964, p. 36.
[26] M. Brady, Artificial intelligence and robotics, Artif. Intell. 26 (1985) 79–121.
[27] S. Ullman, Artificial intelligence and the brain: computational studies of the visual system, Annu. Rev. Neurosci. 9 (1986) 1–26.
[28] H. Simon, Cognitive science: the newest science of the artificial, Cognit. Sci. 4 (1981) 33–46.
[29] C.C. Aggarwal, Neural Networks and Deep Learning: A Textbook, first ed., Springer, 2018.
[30] M. Bahiraei, S. Heshmatian, H. Moayedi, Artificial intelligence in the field of nanofluids: a review on applications and potential future directions, Powder Technol. 353 (2019) 276–301.
[31] M. Chen, S. Mao, Y. Liu, Big data: a survey, Mobile Network. Appl. 19 (2014) 171–209.
[32] J.M. Font, T. Mahlmann, Dota 2 bot competition, IEEE Trans. Games 11 (2019) 285–289.
[33] O.E. David, N.S. Netanyahu, L. Wolf, DeepChess: end-to-end deep neural network for automatic learning in chess, Int. Conf. Artificial Neural Netw. (ICANN) 9887 (2016) 88–96.
[34] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359.
[35] A. Richardson, B.M. Signor, B.A. Lidbury, T. Badrick, Clinical chemistry in higher dimensions: machine-learning and enhanced prediction from routine clinical chemistry data, Clin. Biochem. 49 (2016) 1213–1220.
[36] M. Poostchi, K. Silamut, R.J. Maude, S. Jaeger, G. Thoma, Image analysis and machine learning for detecting malaria, Transl. Res. 194 (2018) 36–55.
[37] Y. Mohamadou, A. Halidou, P.T. Kapen, A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19, Appl. Intell. 50 (2020) 3913–3925.
[38] A. Shaban-Nejad, M. Michalowski, D.L. Buckeridge, Health intelligence, How artificial intelligence transforms population and personalized health, NPJ Digit Med. 1 (2018) 53.
[39] Y. Wang, Y. Yu, S. Cao, X. Zhang, S. Gao, A review of applications of artificial intelligent algorithms in wind farms, Artif. Intell. Rev. 53 (2019) 3447–3500.
[40] L. Li, S. Rong, R. Wang, S. Yu, Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review, Chem. Eng. J. 405 (2021).
[41] J. Nayak, K. Vakula, P. Dinesh, B. Naik, D. Pelusi, Intelligent food processing: journey from artificial neural network to deep learning, Comput. Sci. Rev. 38 (2020), 100297.
[42] S.L. Brunton, B.R. Noack, P. Koumoutsakos, Machine learning for fluid mechanics, Annu. Rev. Fluid Mech. 52 (2020) 477–508.
[43] M.J. Mrowinski, P. Fronczak, A. Fronczak, M. Ausloos, O. Nedic, Artificial intelligence in peer review: how can evolutionary computation support journal editors? PloS One 12 (2017), e0184711.
[44] A. Shrestha, A. Mahmood, Review of deep learning algorithms and architectures, IEEE Access 7 (2019) 53040–53065.
[45] O. Engkvist, P.O. Norrby, N. Selmi, Y.H. Lam, Z. Peng, E.C. Sherer, W. Amberg, T. Erhard, L.A. Smyth, Computational prediction of chemical reactions: current status and outlook, Drug Discov. Today 23 (2018) 1203–1218.
[46] C.W. Coley, W.H. Green, K.F. Jensen, Machine learning in computer-aided synthesis planning, Acc. Chem. Res. 51 (2018) 1281–1289.
[47] J. Panteleev, H. Gao, L. Jia, Recent applications of machine learning in medicinal chemistry, Bioorg. Med. Chem. Lett 28 (2018) 2807–2815.
[48] A.C. Mater, M.L. Coote, Deep learning in chemistry, J. Chem. Inf. Model. 59 (2019) 2545–2559.
[49] A.F. de Almeida, R. Moreira, T. Rodrigues, Synthetic organic chemistry driven by artificial intelligence, Nat. Rev. Chem. 3 (2019) 589–604.
[50] J. Gasteiger, Chemistry in times of artificial intelligence, ChemPhysChem 21 (2020) 2233–2242.
[51] F. Strieth-Kalthoff, F. Sandfort, M.H.S. Segler, F. Glorius, Machine learning the ropes: principles, applications and directions in synthetic chemistry, Chem. Soc. Rev. 49 (2020) 6154–6168.
[52] E. Szymańska, Modern data science for analytical chemical data – a comprehensive review, Anal. Chim. Acta 1028 (2018) 1–10.
[53] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Model. 28 (1988) 31–36.
[54] N.M. O'Boyle, Towards a Universal SMILES representation - a standard method to generate canonical SMILES based on the InChI, J. Chem. Inf. Model. 4 (2012) 22.
[55] Y. Tan, H. Matsui, N. Ishiguro, T. Uruga, D.-N. Nguyen, O. Sekizawa, T. Sakata, N. Maejima, K. Higashi, H.C. Dam, M. Tada, Pt–Co/C cathode catalyst degradation in a polymer electrolyte fuel cell investigated by an infographic approach combining three-dimensional spectroimaging and unsupervised learning, J. Phys. Chem. C 123 (2019) 18844–18853.
[56] S. Jaeger, S. Fulle, S. Turk, Mol2vec: unsupervised machine learning approach with chemical intuition, J. Chem. Inf. Model. 58 (2018) 27–35.
[57] M. Ceriotti, Unsupervised machine learning in atomistic simulations, between predictions and understanding, J. Chem. Phys. 150 (2019) 150901.
[58] Y. Zhang, A.A. Lee, Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning, Chem. Sci. 10 (2019) 8154–8163.
[59] N. Dong, D.M. Spencer, Q. Quan, J.C.Y. Le Blanc, J. Feng, M. Li, K.W.M. Siu, I.K. Chu, rPTMDetermine: a fully automated methodology for endogenous tyrosine nitration validation, site-localization, and beyond, Anal. Chem. 92 (2020) 10768–10776.
[60] Z. Zhou, S. Kearnes, L. Li, R.N. Zare, P. Riley, Optimization of molecules via deep reinforcement learning, Sci. Rep. 9 (2019) 10752.
[61] A.M. Jimenez-Carvelo, A. Gonzalez-Casado, M.G. Bagur-Gonzalez, L. Cuadros-Rodriguez, Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity - a review, Food Res. Int. 122 (2019) 25–39.
[62] F. Modaresi, S. Araghinejad, A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification, Water Resour. Manag. 28 (2014) 4095–4111.
[63] S.E. Reichenbach, C.A. Zini, K.P. Nicolli, J.E. Welke, C. Cordero, Q. Tao, Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition, J. Chromatogr. A 1595 (2019) 158–167.
[64] J. Qin, F. Vasefi, R.S. Hellberg, A. Akhbardeh, R.B. Isaacs, A.G. Yilmaz, C. Hwang, I. Baek, W.F. Schmidt, M.S. Kim, Detection of fish fillet substitution and mislabeling using multimode hyperspectral imaging techniques, Food Contr. 114 (2020).
[65] M. Su, G. Feng, Z. Liu, Y. Li, R. Wang, Tapping on the black box: how is the scoring power of a machine-learning scoring function dependent on the training set? J. Chem. Inf. Model. 60 (2020) 1122–1136.
[66] S. Zhong, K. Zhang, D. Wang, H. Zhang, Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds, Chem. Eng. J. 405 (2021), 126627.
[67] D.A. Gredell, A.R. Schroeder, K.E. Belk, C.D. Broeckling, A.L. Heuberger, S.Y. Kim, D.A. King, S.D. Shackelford, J.L. Sharp, T.L. Wheeler, D.R. Woerner, J.E. Prenni, Comparison of machine learning algorithms for predictive modeling of beef attributes using rapid evaporative ionization mass spectrometry (REIMS) data, Sci. Rep. 9 (2019) 5721.

[68] Y. Zhang, A. Li, B. Deng, K.K. Hughes, Data-driven predictive models for chemical durability of oxide glass under different chemical conditions, NPJ Mater. Degrad. 4 (2020) 14.
[69] J.E. Vidueira Ferreira, C.H.S. da Costa, R.M. de Miranda, A.F. de Figueiredo, The use of the k nearest neighbor method to classify the representative elements, Educ. Quím. 26 (2015) 195–201.
[70] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, J.B. Mitchell, Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization, J. Chem. Inf. Model. 46 (2006) 2412–2422.
[71] G.W. Kauffman, P.C. Jurs, QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors, J. Chem. Inf. Comput. Sci. 41 (2001) 1553–1560.
[72] S. Sathe, C.C. Aggarwal, Nearest Neighbor Classifiers versus Random Forests and Support Vector Machines, IEEE International Conference on Data Mining (ICDM), 2019, pp. 1300–1305, 2019.
[73] E.H. Houssein, M.E. Hosney, D. Oliva, W.M. Mohamed, M. Hassaballah, A novel hybrid Harris hawks optimization and support vector machines for drug design and discovery, Comput. Chem. Eng. 133 (2020).
[74] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, A comparative study of K-nearest neighbour, support vector machine and multi-layer Perceptron for thalassemia screening, Chemometr. Intell. Lab. Syst. 69 (2003) 13–20.
[75] H. Drucker, C.J. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Syst. 9 (1996) 155–161.
[76] M. Awad, R. Khanna, Support Vector Regression, Efficient Learning Machines, Apress, Berkeley, CA, 2015, pp. 67–80.
[77] M. Aliramezani, A. Norouzi, C.R. Koch, A grey-box machine learning based model of an electrochemical gas sensor, Sensor. Actuator. B Chem. 321 (2020), 128414.
[78] R.M. Balabin, E.I. Lomakina, Support vector machine regression (SVR/LS-SVM)–an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst 136 (2011) 1703–1712.
[79] G. Grazioli, S. Roy, C.T. Butts, Predicting reaction products and automating reactive trajectory characterization in molecular simulations with support vector machines, J. Chem. Inf. Model. 59 (2019) 2753–2764.
[80] P. Rajendra, A. Subbarao, G. Ramu, V. Brahmajirao, Prediction of drug solubility on parallel computing architecture by support vector machines, Netw. Model. Analys. Health Informat. Bioinformat. 7 (2018).
[81] V.G. Maltarollo, T. Kronenberger, G.Z. Espinoza, P.R. Oliveira, K.M. Honorio, Advances with Support Vector Machines for Novel Drug Discovery, Expert Opin Drug Discov., 2019, pp. 23–33.
[82] E. Bona, I. Marquetti, J.V. Link, G.Y.F. Makimori, V. da Costa Arca, A.L. Guimarães Lemes, J.M.G. Ferreira, M.B. dos Santos Scholz, P. Valderrama, R.J. Poppi, Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee, Food Sci. Technol. 76 (2017) 330–336.
[83] N.L. da Costa, I.A. Castro, R. Barbosa, Classification of cabernet sauvignon from two different countries in south America by chemical compounds and support vector machines, Appl. Artif. Intell. 30 (2016) 679–689.
[84] N.L. Costa, L.A.G. Llobodanin, I.A. Castro, R. Barbosa, Using support vector machines and neural networks to classify merlot wines from south America, Inf. Process. Agric. 6 (2019) 265–278.
[85] C. Maione, V.C.d.O. Souza, L.R. Togni, J.L. da Costa, A.D. Campiglia, F. Barbosa, R.M. Barbosa, Establishing chemical profiling for ecstasy tablets based on trace element levels and support vector machine, Neural Comput. Appl. 30 (2016) 947–955.
[86] M. Ghasemi-Varnamkhasti, S.S. Mohtasebi, M. Siadat, H. Ahmadi, S.H. Razavi, From simple classification methods to machine learning for the binary discrimination of beers using electronic nose data, Eng. Agric., Environ. Food 8 (2015) 44–51.
[87] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.
[88] D.S. Palmer, N.M. O'Boyle, R.C. Glen, J.B. Mitchell, Random forest models to predict aqueous solubility, J. Chem. Inf. Model. 47 (2007) 150–158.
[89] G. Purcaro, P.H. Stefanuto, F.A. Franchina, M. Beccaria, W.F. Wieland-Alter, P.F. Wright, J.E. Hill, SPME-GCxGC-TOF MS fingerprint of virally-infected cell culture: sample preparation optimization and data processing evaluation, Anal. Chim. Acta 1027 (2018) 158–167.
[90] C. De Stefano, G. Lando, C. Malegori, P. Oliveri, S. Sammartano, Prediction of water solubility and Setschenow coefficients by tree-based regression strategies, J. Mol. Liq. 282 (2019) 401–406.
[91] S.W.C. Walker, A. Anwar, J.M. Psutka, J. Crouse, C. Liu, J.C.Y. Le Blanc, J. Montgomery, G.H. Goetz, J.S. Janiszewski, J.L. Campbell, W.S. Hopkins, Determining molecular properties with differential mobility spectrometry and machine learning, Nat. Commun. 9 (2018) 5096.
[92] A.P. Lind, P.C. Anderson, Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties, PloS One 14 (2019), e0219774.
[93] P. Banerjee, R. Preissner, BitterSweetForest: a random forest based binary classifier to predict bitterness and sweetness of chemical compounds, Front. Chem. 6 (2018) 93.
[94] S. Chavan, N. Scherbak, M. Engwall, D. Repsilber, Predicting chemical-induced liver toxicity using high-content imaging phenotypes and chemical descriptors: a random forest approach, Chem. Res. Toxicol. 33 (2020) 2261–2275.
[95] S. Seifert, Application of random forest based approaches to surface-enhanced Raman scattering data, Sci. Rep. 10 (2020) 5436.
[96] F.B. de Santana, W. Borges Neto, R.J. Poppi, Random forest as one-class classifier and infrared spectroscopy for food adulteration detection, Food Chem. 293 (2019) 323–332.
[97] E.D. Strozier, D.D. Mooney, D.A. Friedenberg, T.P. Klupinski, C.A. Triplett, Use of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection and random forest pattern recognition techniques for classifying chemical threat agents and detecting chemical attribution signatures, Anal. Chem. 88 (2016) 7068–7075.
[98] S. Fan, T. Kind, T. Cajka, S.L. Hazen, W.H.W. Tang, R. Kaddurah-Daouk, M.R. Irvin, D.K. Arnett, D.K. Barupal, O. Fiehn, Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data, Anal. Chem. 91 (2019) 3590–3596.
[99] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding, Anal. Chim. Acta 879 (2015) 10–23.
[100] G. Chen, X. Zhang, Z. Wu, J. Su, G. Cai, An efficient tea quality classification algorithm based on near infrared spectroscopy and random Forest, J. Food Process. Eng. 44 (2020), e13604.
[101] Y. Jiang, H. Ge, Y. Zhang, Quantitative analysis of wheat maltose by combined terahertz spectroscopy and imaging based on Boosting ensemble learning, Food Chem. 307 (2020) 125533.
[102] D.A. Winkler, T.C. Le, Performance of deep and shallow neural networks, the universal approximation Theorem, activity cliffs, and QSAR, Mol. Informat. 36 (2017), 1600118.
[103] G. Gauglitz, Artificial vs. Human Intelligence in Analytics: Do Computers Outperform Analytical Chemists?, Analytical and Bioanalytical Chemistry, Springer Verlag, 2019, pp. 5631–5632.
[104] X. Chen, D. Xun, R. Zheng, L. Zhao, Y. Lu, J. Huang, R. Wang, Y. Wang, Deep-learning-assisted assessment of DNA damage based on foci images and its application in high-content screening of lead compounds, Anal. Chem. 92 (2020) 14267–14277.
[105] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. (2016) 770–778.
[106] D. Jha, L. Ward, A. Paul, W.K. Liao, A. Choudhary, C. Wolverton, A. Agrawal, ElemNet: deep learning the chemistry of materials from only elemental composition, Sci. Rep. 8 (2018) 17593.
[107] A. Dhillon, G.K. Verma, Convolutional neural network: a review of models, methodologies and applications to object detection, Progr. Artificial Intell. 9 (2020) 85–112.
[108] H. He, M. Xu, C. Zong, P. Zheng, L. Luo, L. Wang, B. Ren, Speeding up the line-scan Raman imaging of living cells by deep convolutional neural network, Anal. Chem. 91 (2019) 7070–7077.
[109] X. Zhang, T. Lin, J. Xu, X. Luo, Y. Ying, DeepSpectra: an end-to-end deep learning approach for quantitative spectral analysis, Anal. Chim. Acta 1058 (2019) 48–57.
[110] D.D. Matyushin, A.Y. Sholokhova, A.K. Buryak, Deep learning driven GC-MS library search and its application for metabolomics, Anal. Chem. 92 (2020) 11818–11825.
[111] Q. Duan, Z. Xu, S. Zheng, J. Chen, Y. Feng, L. Run, J. Lee, Machine learning based on holographic scattering spectrum for mixed pollutants analysis, Anal. Chim. Acta 1143 (2021) 298–305.
[112] J. Zhu, A.S. Sharma, J. Xu, Y. Xu, T. Jiao, Q. Ouyang, H. Li, Q. Chen, Rapid on-site identification of pesticide residues in tea by one-dimensional convolutional neural network coupled with surface-enhanced Raman scattering, Spectrochim. Acta Mol. Biomol. Spectrosc. 246 (2021), 118994.
[113] M. Tashiro, Y. Imamura, M. Katouda, De novo generation of optically active small organic molecules using Monte Carlo tree search combined with recurrent neural network, J. Comput. Chem. 42 (2020) 136–143.
[114] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv, 2017.
[115] Z. Rong, Q. Tan, L. Cao, L. Zhang, K. Deng, Y. Huang, Z.-J. Zhu, Z. Li, K. Li, NormAE: deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data, Anal. Chem. 92 (2020) 5082–5090.
[116] S.-Y. Cho, Y. Lee, S. Lee, H. Kang, J. Kim, J. Choi, J. Ryu, H. Joo, H.-T. Jung, J. Kim, Finding hidden signals in chemical sensors using deep learning, Anal. Chem. 92 (2020) 6529–6537.
[117] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, Adv. Neural Inf. Process. Syst. (2018) 6389–6399.
[118] I. Gonzalez-Fernandez, M.A. Iglesias-Otero, M. Esteki, O.A. Moldes, J.C. Mejuto, J. Simal-Gandara, A critical review on the use of artificial neural networks in olive oil production, characterization and authentication, Crit. Rev. Food Sci. Nutr. 59 (2019) 1913–1926.
[119] A. Alibakshi, Strategies to develop robust neural network models: prediction of flash point as a case study, Anal. Chim. Acta 1026 (2018) 69–76.
[120] J. Yang, J. Xu, X. Zhang, C. Wu, T. Lin, Y. Ying, Deep learning for vibrational spectral analysis: recent progress and a practical guide, Anal. Chim. Acta 1081 (2019) 6–17.
[121] Y. Liu, S. Zhou, W. Han, W. Liu, Z. Qiu, C. Li, Convolutional neural network for hyperspectral data analysis and effective wavelengths selection, Anal. Chim.

Acta 1086 (2019) 46–54.
[122] M. Deulofeu, L. Kolarova, V. Salvado, E. Maria Pena-Mendez, M. Almasi, M. Stork, L. Pour, P. Boadas-Vaello, S. Sevcikova, J. Havel, P. Vanhara, Rapid discrimination of multiple myeloma patients by artificial neural networks coupled with mass spectrometry of peripheral blood plasma, Sci. Rep. 9 (2019) 7975.
[123] D. Fichou, G.E. Morlock, Powerful artificial neural network for planar chromatographic image evaluation, shown for denoising and feature extraction, Anal. Chem. 90 (2018) 6984–6991.
[124] N.L. da Costa, M.S. da Costa, R. Barbosa, A review on the application of chemometrics and machine learning algorithms to evaluate beer authentication, Food Anal. Methods 14 (2020) 136–155.
[125] D.I. Patrício, R. Rieder, Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review, Comput. Electron. Agric. 153 (2018) 69–81.
[126] N.A.B. Gray, Artificial intelligence in chemistry, Anal. Chim. Acta 210 (1988) 9–32.
[127] H. Li, Z. Zhang, Z.-Z. Zhao, Data-mining for processes in chemistry, Mater. Eng., Process. 7 (2019) 151.
[128] D.D. Nguyen, Z. Cang, G.W. Wei, A review of mathematical representations of biomolecular data, Phys. Chem. Chem. Phys. 22 (2020) 4343–4367.
[129] T. Rodrigues, The good, the bad, and the ugly in chemical and biological data for machine learning, Drug Discov. Today Technol. 32–33 (2019) 3–8.
[130] N. Schneider, D.M. Lowe, R.A. Sayle, M.A. Tarselli, G.A. Landrum, Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter, J. Med. Chem. 59 (2016) 4385–4402.
[131] E. Szymanska, Modern data science for analytical chemical data - a comprehensive review, Anal. Chim. Acta 1028 (2018) 1–10.
[132] A.F. de Almeida, R. Moreira, T. Rodrigues, Synthetic organic chemistry driven by artificial intelligence, Nat. Rev. Chem. 3 (2019) 589–604.
[133] A.C. Vaucher, F. Zipoli, J. Geluykens, V.H. Nair, P. Schwaller, T. Laino, Automated extraction of chemical synthesis actions from experimental procedures, Nat. Commun. 11 (2020) 3601.
[134] L. Shang, C. Liu, Y. Tomiura, K. Hayashi, Machine-learning-based olfactometer: prediction of odor perception from physicochemical features of odorant molecules, Anal. Chem. 89 (2017) 11999–12005.
[135] P. Bonini, T. Kind, H. Tsugawa, D.K. Barupal, O. Fiehn, Retip: retention time prediction for compound annotation in untargeted metabolomics, Anal. Chem. 92 (2020) 7515–7522.
[136] J.E. Elias, F.D. Gibbons, O.D. King, F.P. Roth, S.P. Gygi, Intensity-based protein identification by machine learning from a library of tandem mass spectra, Nat. Biotechnol. 22 (2004) 214–219.
[137] P. Morgante, R. Peverati, ACCDB: a collection of chemistry databases for broad computational purposes, J. Comput. Chem. 40 (2019) 839–848.
[138] A.Y.-T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch, K.A. Persson, T.D. Sparks, Machine learning for materials scientists: an introductory guide toward best practices, Chem. Mater. 32 (2020) 4954–4965.
[139] L.M. Roch, F. Häse, A. Aspuru-Guzik, Chapter 16: ChemOS: an Orchestration Software to Democratize Autonomous Discovery, RSC Drug Discov. Ser., 2021-Janua, 2021, pp. 351–388.
[140] D. Pfau, J.S. Spencer, A.G.D.G. Matthews, W.M.C. Foulkes, Ab initio solution of the many-electron Schrödinger equation with deep neural networks, Phys. Rev. Res. 2 (2020), 033429.
[141] E. Callaway, It will change everything': DeepMind's AI makes gigantic leap in solving protein structures, Nature 588 (2020) 203–204.
[142] X.Y. Dong, X.Q. Niu, Z.Y. Zhang, J.S. Wei, H.M. Xiong, Red fluorescent carbon dot powder for accurate latent fingerprint identification using an artificial intelligence program, ACS Appl. Mater. Interfaces 12 (2020) 29549–29555.
[143] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, Chem. Sci. 9 (2018) 513–530.
[144] J.L. Reymond, The chemical space project, Acc. Chem. Res. 48 (2015) 722–730.
[145] H. Pezeshgi Modarres, M.R. Mofrad, A. Sanati-Nezhad, ProtDataTherm: a database for thermostability analysis and engineering of proteins, PloS One 13 (2018), e0191222.
[146] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, T. Laino, Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models, Chem. Sci. 9 (2018) 6091–6098.
[147] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C.A. Hunter, C. Bekas, A.A. Lee, Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction, ACS Cent. Sci. 5 (2019) 1572–1583.
[148] J. Jo, B. Kwak, H.S. Choi, S. Yoon, The message passing neural networks for chemical property prediction on SMILES, Methods 179 (2020) 65–72.
[149] R. Winter, F. Montanari, F. Noe, D.A. Clevert, Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations, Chem. Sci. 10 (2019) 1692–1701.
[150] A.A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J.L. Klug-McLeod, C.R. Butler, Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space, Chem. Commun. (Camb.) 55 (2019) 12152–12155.
[151] L. Wang, C. Zhang, R. Bai, J. Li, H. Duan, Heck reaction prediction using a transformer model based on a transfer learning strategy, Chem. Commun. (Camb.) 56 (2020) 9368–9371.
[152] K. Mao, P. Zhao, T. Xu, Y. Rong, X. Xiao, J. Huang, Molecular graph enhanced transformer for retrosynthesis prediction, bioRxiv (2020). https://doi.org/10.1101/2020.03.05.979773.
[153] J. Payne, M. Srouji, D.A. Yap, V. Kosaraju, BERT Learns (And Teaches) Chemistry, arXiv, 2020.
[154] P. Schwaller, R. Petraglia, V. Zullo, V.H. Nair, R.A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy, Chem. Sci. 11 (2020) 3316–3325.
[155] A.P. Shevchenko, R.A. Eremin, V.A. Blatov, The CSD and knowledge databases: from answers to questions, CrystEngComm 22 (2020) 7298–7307.
[156] N.A.B. Gray, Applications of artificial intelligence for organic chemistry: analysis of C-13 spectra, Artif. Intell. 22 (1984) 1–21.
[157] R.E. Valdés-Pérez, Machine discovery in chemistry: new results, Artif. Intell. 74 (1995) 191–201.
[158] F. Peiretti, J.M. Brunel, Artificial intelligence: the future for organic chemistry? ACS Omega 3 (2018) 13263–13266.
[159] C. Empel, R.M. Koenigs, Artificial-intelligence-Driven organic synthesis-en route towards autonomous synthesis? Angew Chem. Int. Ed. Engl. 58 (2019) 17114–17116.
[160] L.A. Griffin, I. Gaponenko, N. Bassiri-Gharb, Better, faster, and less biased machine learning: electromechanical switching in ferroelectric thin films, Adv. Mater. 32 (2020), e2002425.
[161] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.
[162] A. Hoyle, H. Wallach, I. Augenstein, R. Cotterell, Unsupervised Discovery of Gendered Language through Latent-Variable Modeling, 2019 arXiv preprint arXiv:1906.04760.
[163] X. Jin, C. Liu, T. Xu, L. Su, X. Zhang, Artificial intelligence biosensors: challenges and prospects, Biosens. Bioelectron. 165 (2020).
[164] H. Jiang, W. Xu, Q. Chen, Determination of tea polyphenols in green tea by homemade color sensitive sensor combined with multivariate analysis, Food Chem. 319 (2020) 126584.
[165] C. Kim, H. Lee, V. Devaraj, W.G. Kim, Y. Lee, Y. Kim, N.N. Jeong, E.J. Choi, S.H. Baek, D.W. Han, H. Sun, J.W. Oh, Hierarchical cluster Analysis of medical chemicals detected by a bacteriophage-based colorimetric sensor array, J. Nanomater. 10 (2020).
[166] R. Luo, G. Ma, S. Bi, Q. Duan, J. Chen, Y. Feng, F. Liu, J. Lee, Machine learning for total organic carbon analysis of environmental water samples using high-throughput colorimetric sensors, Analyst 145 (2020) 2197–2203.
[167] Q. Duan, J. Lee, S. Zheng, J. Chen, R. Luo, Y. Feng, Z. Xu, A color-spectral machine learning path for analysis of five mixed amino acids, Chem. Commun. (Camb.) 56 (2020) 1058–1061.
[168] H. Yu, W. Jing, R. Iriya, Y. Yang, K. Syal, M. Mo, T.E. Grys, S.E. Haydel, S. Wang, N. Tao, Phenotypic antimicrobial susceptibility testing with deep learning video microscopy, Anal. Chem. 90 (2018) 6314–6322.
[169] Y. Wu, A. Ray, Q. Wei, A. Feizi, X. Tong, E. Chen, Y. Luo, A. Ozcan, Deep learning enables high-throughput analysis of particle-aggregation-based biosensors imaged using holography, ACS Photonics 6 (2018) 294–301.
[170] Z. Chen, Z. Chen, Z. Song, W. Ye, Z. Fan, Smart gas sensor arrays powered by artificial intelligence, J. Semiconduct. 40 (2019) 111601.
[171] J. Tan, J. Xu, Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: a review, Artificial Intell. Agric. 4 (2020) 104–115.
[172] C. Gonzalez Viejo, S. Fuentes, A. Godbole, B. Widdicombe, R.R. Unnithan, Development of a low-cost e-nose to assess aroma profiles: an artificial intelligence application to assess beer quality, Sensor. Actuator. B Chem. 308 (2020), 127688.
[173] T. Hayasaka, A. Lin, V.C. Copa, L.P. Lopez, R.A. Loberternos, L.I.M. Ballesteros, Y. Kubota, Y. Liu, A.A. Salvador, L. Lin, An electronic nose using a single graphene FET and machine learning for water, methanol, and ethanol, Microsyst. Nanoeng. 6 (2020).
[174] T. Julian, S.N. Hidayat, A. Rianjanu, A.B. Dharmawan, H.S. Wasisto, K. Triyana, Intelligent mobile electronic nose system comprising a hybrid polymer-functionalized quartz crystal microbalance sensor array, ACS Omega 5 (2020) 29492–29503.
[175] G. Łagód, S.M. Duda, D. Majerek, A. Szutt, A. Dołhańczuk-Śródka, Application of electronic nose for evaluation of wastewater treatment process effects at full-scale WWTP, Processes 7 (2019) 251.
[176] I. Rodriguez-Rodriguez, J.V. Rodriguez, I. Chatzigiannakis, M.A. Zamora Izquierdo, On the possibility of predicting glycaemia 'on the fly' with constrained IoT devices in type 1 diabetes mellitus patients, Sensors 19 (2019) 4538.
[177] U. Hoss, E.S. Budiman, Factory-Calibrated continuous glucose sensors: the science behind the Technology, Diabetes Technol. Therapeut. 19 (2017) S44–S50.
[178] V.C. Rodrigues, J.C. Soares, A.C. Soares, D.C. Braz, M.E. Melendez, L.C. Ribas, L.F.S. Scabini, O.M. Bruno, A.L. Carvalho, R.M. Reis, R.C. Sanfelice, O.N. Oliveira Jr., Electrochemical and optical detection and machine learning applied to images of genosensors for diagnosis of prostate cancer with the biomarker PCA3, Talanta 222 (2021) 121444.
[179] Y. Xu, C. Li, Y. Jiang, M. Guo, Y. Yang, Y. Yang, H. Yu, Electrochemical impedance spectroscopic detection of E.coli with machine learning, J. Electrochem. Soc. 167 (2020), 047508.
[180] F. Cui, Y. Yue, Y. Zhang, Z. Zhang, H.S. Zhou, Advancing biosensors with machine learning, ACS Sens. 5 (2020) 3346–3364.
[181] M.E. Solmaz, A.Y. Mutlu, G. Alankus, V. Kılıç, A. Bayram, N. Horzum,

Quantifying colorimetric tests using a smartphone app based on machine learning classifiers, Sensor. Actuator. B Chem. 255 (2018) 1967–1973.
[182] M.S. Draz, A. Vasan, A. Muthupandian, M.K. Kanakasabapathy, P. Thirumalaraju, A. Sreeram, S. Krishnakumar, V. Yogesh, W. Lin, G.Y. Xu, Virus detection using nanoparticles and deep neural network–enabled smartphone system, Sci. Adv. 6 (2020), eabd5354.
[183] N.C. Cady, N. Tokranova, A. Minor, N. Nikvand, K. Strle, W.T. Lee, W. Page, E. Guignon, A. Pilar, G.N. Gibson, Multiplexed detection and quantification of human antibody response to COVID-19 infection using a plasmon enhanced biosensor platform, Biosens. Bioelectron. 171 (2021) 112679.
[184] Z. Jiang, M. Hu, Z. Gao, L. Fan, R. Dai, Y. Pan, W. Tang, G. Zhai, Y. Lu, Detection of respiratory infections using RGB-infrared sensors on portable device, IEEE Sensor. J. 20 (2020) 13674–13681.
[185] H. Hwang, H.K. Jeong, H.K. Lee, G.W. Park, J.Y. Lee, S.Y. Lee, Y.M. Kang, H.J. An, J.G. Kang, J.H. Ko, J.Y. Kim, J.S. Yoo, Machine learning classifies core and outer fucosylation of N-glycoproteins using mass spectrometry, Sci. Rep. 10 (2020) 318.
[186] D. Panagopoulos Abrahamsson, J.S. Park, R.R. Singh, M. Sirota, T.J. Woodruff, Applications of machine learning to in silico quantification of chemicals without analytical standards, J. Chem. Inf. Model. 60 (2020) 2718–2727.
[187] I. Jang, J.U. Lee, J.M. Lee, B.H. Kim, B. Moon, J. Hong, H.B. Oh, LC-MS/MS software for screening unknown erectile dysfunction drugs and analogues: artificial neural network classification, peak-count scoring, simple similarity search, and hybrid similarity search algorithms, Anal. Chem. 91 (2019) 9119–9128.
[188] J.A. Carter, L.M. O'Brien, T. Harville, B.T. Jones, G.L. Donati, Machine learning tools to estimate the severity of matrix effects and predict analyte recovery in inductively coupled plasma optical emission spectrometry, Talanta 223 (2021) 121665.
[189] J.F.Q. Pereira, M.F. Pimentel, J.M. Amigo, R.S. Honorato, Detection and identification of Cannabis sativa L. using near infrared hyperspectral imaging and machine learning methods. A feasibility study, Spectrochim. Acta Mol. Biomol. Spectrosc. 237 (2020) 118385.
[190] P. Lasch, M. Stammler, M. Zhang, M. Baranska, A. Bosch, K. Majzner, FT-IR hyperspectral imaging and artificial neural network analysis for identification of pathogenic bacteria, Anal. Chem. 90 (2018) 8896–8904.
[191] M. Ghaffari, N. Omidikia, C. Ruckebusch, Joint selection of essential pixels and essential variables across hyperspectral images, Anal. Chim. Acta 1141 (2021) 36–46.
[192] D. Zhou, Y. Yu, R. Hu, Z. Li, Discrimination of Tetrastigma hemsleyanum according to geographical origin by near-infrared spectroscopy combined with a deep learning approach, Spectrochim. Acta Mol. Biomol. Spectrosc. 238 (2020) 118380.
[193] X. Zhang, Y. Li, Y. Tao, Y. Wang, C. Xu, Y. Lu, A novel method based on infrared spectroscopic inception-resnet networks for the detection of the major fish allergen parvalbumin, Food Chem. 337 (2021) 127986.
[194] Y. Hong, S. Chen, Y. Zhang, Y. Chen, L. Yu, Y. Liu, Y. Liu, H. Cheng, Y. Liu, Rapid identification of soil organic matter level via visible and near-infrared spectroscopy: effects of two-dimensional correlation coefficient and extreme learning machine, Sci. Total Environ. 644 (2018) 1232–1243.
[195] M. Gastegger, J. Behler, P. Marquetand, Machine learning molecular dynamics for the simulation of infrared spectra, Chem. Sci. 8 (2017) 6924–6935.
[196] V.H. da Silva, F. Murphy, J.M. Amigo, C. Stedmon, J. Strand, Classification and quantification of microplastics (<100 mum) using a focal plane array-fourier transform infrared imaging system and machine learning, Anal. Chem. 92 (2020) 13724–13733.
[197] G. Larios, G. Nicolodelli, M. Ribeiro, T. Canassa, A.R. Reis, S.L. Oliveira, C.Z. Alves, B.S. Marangoni, C. Cena, Soybean seed vigor discrimination by using infrared spectroscopy and machine learning algorithms, Analyt. Methods 12 (2020) 4303–4309.
[198] C.S. Tan, S.Y. Leow, C. Ying, C.J. Tan, T.L. Yoon, C. Jingying, M.F. Yam, Comparison of FTIR spectrum with chemometric and machine learning classifying analysis for differentiating guan-mutong a nephrotoxic and carcinogenic traditional Chinese medicine with chuan-mutong, Microchem. J. 163 (2021), 105835.
[199] J.C.M. Gomes, L.C. Souza, L.C. Oliveira, SmartSPR sensor: machine learning approaches to create intelligent surface plasmon based sensors, Biosens. Bioelectron. 172 (2021) 112760.
[200] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.
[201] L. Ju, A. Lyu, H. Hao, W. Shen, H. Cui, Deep learning-assisted three-dimensional fluorescence difference spectroscopy for identification and semi-quantification of illicit drugs in biofluids, Anal. Chem. 91 (2019) 9343–9347.
[202] F.A. Chiappini, F. Allegrini, H.C. Goicoechea, A.C. Olivieri, Sensitivity for multivariate calibration based on multilayer Perceptron artificial neural networks, Anal. Chem. 92 (2020) 12265–12272.
[203] H. Li, M. Mehedi Hassan, J. Wang, W. Wei, M. Zou, Q. Ouyang, Q. Chen, Investigation of nonlinear relationship of surface enhanced Raman scattering signal for robust prediction of thiabendazole in apple, Food Chem. 339 (2021) 127843.
[204] J. Zhu, W. Ahmad, T. Jiao, J. Wang, H. Jiang, H. Li, Q. Chen, Interval combination iterative optimization approach coupled with SIMPLS (ICIOA-SIMPLS) for quantitative analysis of surface-enhanced Raman scattering (SERS) spectra, Anal. Chim. Acta 1105 (2020) 45–55.
[205] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (1993) 251–263.
[206] R.G. Brereton, Pattern recognition in chemometrics, Chemometr. Intell. Lab. Syst. 149 (2015) 90–96.
[207] W. Lu, X. Chen, L. Wang, H. Li, Y.V. Fu, Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification, Anal. Chem. 92 (2020) 6288–6296.
[208] H. Shi, H. Wang, X. Meng, R. Chen, Y. Zhang, Y. Su, Y. He, Setting up a surface-enhanced Raman scattering database for artificial-intelligence-based label-free discrimination of tumor suppressor genes, Anal. Chem. 90 (2018) 14216–14221.
[209] Z.P. Yali, A.P. Jadid, L.A. Samin, Modeling of retention time for polychlorinated biphenyl congeners in human adipose tissue using quantitative structure–retention relationship methodology, Int. J. Environ. Sci. Technol. 14 (2017) 2357–2366.
[210] A. McDaniel, L. Perry, Q. Liu, W.-C. Shih, J. Yu, Toward the identification of marijuana varieties by headspace chemical forensics, Forensic Chem. 11 (2018) 23–31.
[211] L. Lebanov, L. Tedone, A. Ghiasvand, B. Paull, Random Forests machine learning applied to gas chromatography - mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils, Talanta 208 (2020) 120471.
[212] G.M. Randazzo, A. Bileck, A. Danani, B. Vogt, M. Groessl, Steroid identification via deep learning retention time predictions and two-dimensional gas chromatography-high resolution mass spectrometry, J. Chromatogr. A 1612 (2020) 460661.
[213] T. Vrzal, M. Malečková, J. Olšovská, DeepReI: deep learning-based gas chromatographic retention index predictor, Anal. Chim. Acta 1147 (2021) 64–71.
[214] J. Stanstrup, S. Neumann, U. Vrhovsek, PredRet: prediction of retention time by direct mapping between multiple chromatographic systems, Anal. Chem. 87 (2015) 9421–9428.
[215] L.M. Hall, D.W. Hill, K. Bugden, S. Cawley, L.H. Hall, M.H. Chen, D.F. Grant, Development of a reverse phase HPLC retention index model for non-targeted metabolomics using synthetic compounds, J. Chem. Inf. Model. 58 (2018) 591–604.
[216] X. Domingo-Almenara, C. Guijas, E. Billings, J.R. Montenegro-Burke, W. Uritboonthai, A.E. Aisporna, E. Chen, H.P. Benton, G. Siuzdak, The METLIN small molecule dataset for machine learning-based retention time prediction, Nat. Commun. 10 (2019) 5811.
[217] W.N.L.d. Santos, M.C. da Silva Sauthier, A.M.P. dos Santos, D. de Andrade Santana, R.S. Almeida Azevedo, J. da Cruz Caldas, Simultaneous determination of 13 phenolic bioactive compounds in guava (Psidium guajava L.) by HPLC-PAD with evaluation using PCA and Neural Network Analysis (NNA), Microchem. J. 133 (2017) 583–592.
[218] G. Bocaz-Beneventi, F. Tagliaro, F. Bortolotti, G. Manetto, J. Havel, Capillary zone electrophoresis and artificial neural networks for estimation of the post-mortem interval (PMI) using electrolytes measurements in human vitreous humour, Int. J. Leg. Med. 116 (2002) 5–11.
[219] L. Jiao, X. Zhang, Y. Qin, X. Wang, H. Li, Hologram QSAR study on the electrophoretic mobility of aromatic acids, Chemometr. Intell. Lab. Syst. 157 (2016) 202–207.
[220] D. Taylor, D. Powers, Teaching artificial intelligence to read electropherograms, Forensic Sci. Int. Genet. 25 (2016) 10–18.
[221] J.D. Adelman, A. Zhao, D.S. Eberst, M.A. Marciano, Automated detection and removal of capillary electrophoresis artifacts due to spectral overlap, Electrophoresis 40 (2019) 1753–1761.
[222] J. Song, Y. Zheng, M. Huang, L. Wu, W. Wang, Z. Zhu, Y. Song, C. Yang, A sequential multidimensional analysis algorithm for aptamer identification based on structure analysis and machine learning, Anal. Chem. 92 (2020) 3307–3314.
[223] Y. Hou, C. Aldrich, K. Lepkova, L.L. Machuca, B. Kinsella, Analysis of electrochemical noise data by use of recurrence quantification analysis and machine learning methods, Electrochim. Acta 256 (2017) 337–347.
[224] J. Liu, F. Ciucci, The Gaussian process distribution of relaxation times: a machine learning tool for the analysis and prediction of electrochemical impedance spectroscopy data, Electrochim. Acta 331 (2020), 135316.
[225] H. Ma, Y.-L. Ying, Recent progress on nanopore electrochemistry and advanced data processing, Curr. Opin. Electrochem. (2020), 100675.
[226] S. Thompson, M.R. Kilbourn, P.J. Scott, Radiochemistry, PET imaging, and the internet of chemical things, ACS Cent. Sci. 2 (2016) 497–505.
[227] S. Nayak, N.R. Blumenfeld, T. Laksanasopin, S.K. Sia, Point-of-Care diagnostics: recent developments in a connected age, Anal. Chem. 89 (2017) 102–123.
[228] M.A. Booth, S.A.N. Gowers, C.L. Leong, M.L. Rogers, I.C. Samper, A.P. Wickham, M.G. Boutelle, Chemical monitoring in clinical settings: recent developments toward real-time chemical monitoring of patients, Anal. Chem. 90 (2018) 2–18.
[229] M. Mayer, A.J. Baeumner, A megatrend challenging analytical chemistry: biosensor and chemosensor concepts ready for the internet of things, Chem. Rev. 119 (2019) 7996–8027.
[230] N.M. Ralbovsky, I.K. Lednev, Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning, Chem. Soc. Rev. 49 (2020) 7428–7453.

Dr. Silva is a Professor at The National University of Cuyo and Principal Researcher of National Council for Research (CONICET) in Argentina. Her group (Green Analytical Chemistry) is dedicated on the development of methodologies aligned with the principles Green Analytical Chemistry for the extraction and determination of analytes of food and pharmaceutical interest and the study of the biological role of plant secondary metabolites. The latest projects are focused on the analytical applications of Natural Deep Eutectic Solvents (NADES). The outcomes of the research activities have been presented and recognized at national and international scientific meetings and are regularly published in peer-refereed journals. Her research has received support from the National Council for Research (CONICET), National Agency for Science and Technology (ANPCyT), and National University of Cuyo.

Jeb Linton holds a BSEE from Virginia Tech and currently is the CTO for Partner Ecosystem and Cognitive Security within IBM Cloud, Program Director for the National Capital Area Center for Advanced Studies, and founder of the IBM Cognitive Security initiative. Mr. Linton is an IBM Senior Technical Staff Member and Master Inventor and has worked for IBM since 2008 as technical strategist and architect on numerous Cloud Computing, Storage, Security and Cognitive Computing projects. He has also formerly acted as CTO of Security for Systems and Technology Group, co-lead of the IBM Watson Architecture Board, and as leader of the IBM Trusted Cloud initiative.

Federico J.V. Gomez. Gomez graduated with a BS in Chemistry (2009) and a PhD in Chemistry (2015) from the National University of San Luis, Argentina under the supervision of Dr. María Fernanda Silva. Currently, he is an independent researcher (CONICET) working on the development and characterization of portable electrochemical sensors for food-related compounds. Dr. Gomez is also an expert on the design and application of natural deep eutectic solvents (NADES), one of the hottest issues in the field of green analytical chemistry.

Dr. Carlos D. Garcia received his B.S. in Biochemistry and Ph.D. in Chemistry from the National University of Cordoba (Argentina) in 1996 and 2001, respectively. From 2002 to 2004, he was a postdoctoral fellow at Mississippi State University and Colorado State University. In September of 2004, he joined the faculty at UTSA where he reached the rank of Professor in 2014. In Aug 2015, he joined Clemson University. In 2018 he was elected Fellow of the Royal Society of Chemistry. His group is focused on the study of interactions of proteins with nanostructured surfaces and their use in analytical chemistry. Additionally, he is developing microfluidic devices to monitor biologically active compounds.

Lucas de Brito Ayres earned his Bachelors in Pharmacy and Biochemistry from the University of Sao Paulo (Brazil) and has already worked with Drs. Do Lago and Gutz on analytical Instrumentation, specifically by developing chemical instruments based on open-source technologies (e.g. Arduino, 3D printers and Android Software). He is currently pursuing his PhD in Chemistry and is interested in continuing working with low-cost sensors, instrumentation, and automation.