Showcasing research from Professor Modestino's laboratory, Tandon School of Engineering, New York University, New York, United States.

Machine learning enhanced spectroscopic analysis: towards autonomous chemical mixture characterization for rapid process optimization

Autonomous chemical process development and optimization methods use algorithms to explore the operating parameter space based on feedback from experimentally determined exit stream compositions. Measuring the compositions of multicomponent streams is challenging, requiring multiple analytical techniques. Herein, we describe a universal analytical methodology based on multitarget regression machine learning (ML) models to rapidly determine chemical mixtures' compositions from Fourier Transform Infrared (FTIR) absorption spectra. We demonstrate that ML models can accurately determine the compositions of multicomponent mixtures of similar species, enhancing spectroscopic chemical quantification for use in autonomous, fast process development and optimization.

As featured in:

See Miguel A. Modestino *et al.*, *Digital Discovery*, 2022, **1**, 35.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

Registered charity number: 207890

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2022, **1**, 35

Machine learning enhanced spectroscopic analysis: towards autonomous chemical mixture characterization for rapid process optimization†

Andrea Angulo, Lankun Yang, Eray S. Aydil and Miguel A. Modestino *

Autonomous chemical process development and optimization methods use algorithms to explore the operating parameter space based on feedback from experimentally determined exit stream compositions. Measuring the compositions of multicomponent streams is challenging, requiring multiple analytical techniques to differentiate between similar chemical components in the mixture and determine their concentration. Herein, we describe a universal analytical methodology based on multitarget regression machine learning (ML) models to rapidly determine chemical mixtures' compositions from Fourier transform infrared (FTIR) absorption spectra. Specifically, we used simulated FTIR spectra for up to 6 components in water and tested seven different ML algorithms to develop the methodology. All algorithms resulted in regression models with mean absolute errors (MAE) between 0–0.27 wt%. We validated the methodology with experimental data obtained on mixtures prepared using a network of programmable pumps in line with an FTIR transmission flow cell. ML models were trained using experimental data and evaluated for mixtures of up to 4-components with similar chemical structures, including alcohols (*i.e.*, glycerol, isopropanol, and 1-butanol) and nitriles (*i.e.*, acrylonitrile, adiponitrile, and propionitrile). Linear regression models predicted concentrations with coefficients of determination, R^2, between 0.955 and 0.986, while artificial neural network models showed a slightly lower accuracy, with R^2 between 0.854 and 0.977. These R^2 correspond to MAEs of 0.28–0.52 wt% for mixtures with component concentrations between 4–10 wt%. Thus, we demonstrate that ML models can accurately determine the compositions of multicomponent mixtures of similar species, enhancing spectroscopic chemical quantification for use in autonomous, fast process development and optimization.

Received 26th October 2021
Accepted 20th December 2021

DOI: 10.1039/d1dd00027f

rsc.li/digitaldiscovery

1. Introduction

Driven by an exponential increase in computational power and the ability to collect, store, and process massive amounts of data, machine learning (ML) has emerged as an invaluable tool for amplifying the performance of many technologies and businesses ranging from self-driving vehicles, targeted marketing, medical diagnostics to financial market forecasting. During the last three years, several studies implemented ML for automating and accelerating chemical process discovery, development, and optimization at the laboratory scale with impressive results,[1–8] but ML has not been fully exploited in this context. Advances on this front can have an enormous impact on chemical manufacturing.

The ML approaches used for chemical process development generally rely on a feedback loop between (1) an ML-guided high-throughput experimental system featuring a chemical reactor and (2) an analytic tool to determine the compositions of the process outlet streams (Fig. 1). Within this approach, an ML algorithm selects optimal experimental conditions to test (*e.g.*, inlet mixture composition, reactor operating conditions), which are then implemented in a reactor (*e.g.*, thermochemical or electrochemical) by an autonomous and automated system. The outlet streams from the reactors, containing mixtures of the desired chemicals byproducts, solvents, additives, and unreacted precursors, are characterized by an analytical tool to determine their composition and the initial ML algorithm uses this information to select the next set of experiments. Determining the composition of an unknown chemical mixture is a challenging task that requires a suite of analytical tools with varying costs and speed (*e.g.*, nuclear magnetic resonance, liquid and/or gas chromatography, mass spectrometry, and/or various optical spectroscopies, amongst others). Moreover, each technique or combination must be adapted to the chemical mixture of interest to provide complete compositional information.

An autonomous chemical process optimization system such as that depicted in Fig. 1 would ideally use a generally applicable, non-invasive, fast, and inexpensive spectrochemical

Department of Chemical and Biomolecular Engineering, Tandon School of Engineering, New York University, 6 Metrotech Ct., Brooklyn, NY 11201, USA. E-mail: modestino@nyu.edu

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1dd00027f

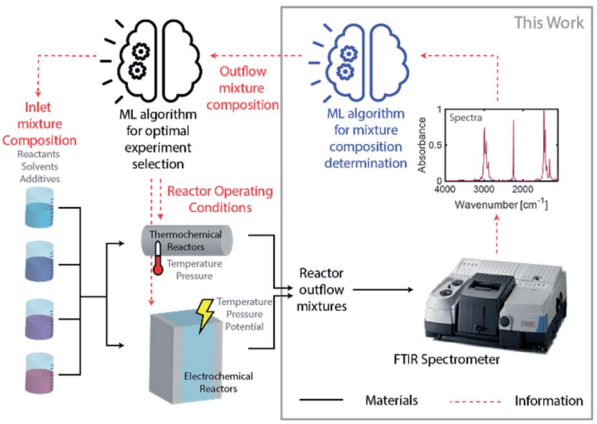

Fig. 1 Diagram of autonomous process discovery, development, and optimization system composed of an ML-guided high-throughput experimental subsystem and an analytic tool to determine the compositions of the process outlet streams. This work focuses on the development of an ML-enhanced FTIR analytical tool for reactor outflow mixture characterization.

characterization tool capable of quantifying the compositions of multicomponent mixtures based on unique identifying molecular spectral features. However, interpretation of spectra collected from mixtures can be complex, and their interpretation and quantification are often challenging because of spectral feature overlap and interactions between different species. We address this challenge in this study by developing and demonstrating a universal ML algorithm that enables rapid inline mixture characterization using an inexpensive Fourier transform infrared (FTIR) spectrometer. The approach we describe is particularly well suited for organic synthesis and aqueous molecular solutions comprising chemicals with vibrational fingerprints, a significant fraction of cases of interest.

FTIR spectroscopy is one of the most powerful and widespread analytical techniques to determine the presence of functional groups in molecules, the compositions of chemical solutions, and to study chemical processes *inline* or *in situ*.[9] FTIR-based methods often rely on the characterization of the position or absorbance of only a few spectroscopic features (absorption peaks) that are indicative of functional groups, while a large fraction of the spectra is ignored because overlapping features are difficult to discern, especially in the fingerprint region (*i.e.*, ~400–1500 cm^{-1}). Furthermore, when multiple analytes are present in the solution, absorption peaks from different molecules can overlap, and interactions between molecules can cause shifts in their positions, significantly increasing the complexity of the analysis.[10]

Machine learning (ML) algorithms can enhance humans' ability to extract information from complex spectral data by learning the correlations between mixture compositions and absorption features. Such algorithms and FTIR data have already been used in specific food and materials applications.[10–12] Previous studies have applied active learning to train classification algorithms and then use these algorithms to identify specific molecules in mixtures.[13–15] A few studies have used regression algorithms to determine species concentrations.[16,17] Recent examples of ML-enhanced FTIR analysis include the use of support vector machine (SVM) classifiers for rapid identification and quantification of components in artificial sweeteners with a prediction accuracy ranging between 60–94%,[17] and the use of linear regression to determine electrolyte composition in lithium-ion batteries within an absolute error of 3–5 wt%.[18] In the first case, the ML models were trained using only 131 absorbance points at selected wavenumbers, and the methodology included spectroscopy preprocessing methods (Savitzky–Golay, first derivative, and their combination). In the second, the ML methodology included multiple data preprocessing steps and manual selection of IR regions for specific functional groups pertaining to the species of interest. In both cases, the sample preparation was done by a lab operator.

Currently, there are multiple open-source and commercial software tools available that can facilitate the implementation of ML algorithms. These tools include MATLAB® PLS Toolbox software and Python's ScikitLearn, Keras, TensorFlow open-source library, among others.

Inspired by the successful implementation of ML in these specific applications, we developed a universal algorithm that uses supervised ML models to determine the concentrations of chemical species in solutions *via* multitarget regression with minimal human intervention. We first generate multicomponent mixture FTIR spectra by linearly combining pure species spectra using the respective molar fractions of each component as weights. These simulated multicomponent spectra are then used to train ML algorithms and develop an ML methodology to determine the compositions of real chemical mixtures. Finally, the ML algorithms are validated and evaluated by comparing their predictions of the compositions of experimental mixtures from their measured FTIR spectra. We used the reactants and possible products of two chemical reactions as model mixture components: electroreduction of acrylonitrile (AN) to adiponitrile (ADN), a nylon precursor, and the valorization of glycerol into other high-value C_3 products. We found that Artificial Neural Networks (ANN) and Linear Regression (LR) with Principal Component Analysis (PCA), also known as Principal Component Regressor (PCR), lead to the most accurate predictions, with R^2 values ranging between 0.854–0.986 and mean absolute errors (MAE) between 0.28–0.52 wt%, depending on the number and identity of components, and ML algorithm.

2. Results and discussion

2.1. Machine learning methodology development

To develop a robust ML approach, we evaluated the performance of various models using the absorbance (A.U.) at n different wavenumbers (wn), $\bar{A} = [A_1,\ldots,A_n]$, as predictor variables, and the concentrations of all (m of them) mixture components, $\bar{C} = [C_1,\ldots,C_m]$ as target variables. Both n and m can vary based on the spectrometer resolution and the number of mixture components, respectively. As a model system, we first considered mixtures of up to 6 components with similar absorption features and relevant to the electrochemical production of nylon precursors: acrylonitrile (AN), adiponitrile

(ADN), propionitrile (PN), ethylenediaminetetraacetic acid (EDTA), phosphate ions (PO_4^{3-}) and tetramethylammonium ions (TMA), in aqueous solutions.[19] The individual spectra of each one of these components are shown in Fig. 2. Mixture FTIR spectra were generated by linearly combining pure species spectra according to Beer's law,

$$A_j = \sum_{i=1}^{m} C_i A_j^i \quad (1)$$

where A_j is the absorbance of the multicomponent solution at the j^{th} wn, A_j^i is the absorbance of the pure species spectra at the j^{th} wn for the i^{th} component, and C_i is the molar concentration of the i^{th} species. Beer's law can be used to estimate the component absorption at low concentrations when there is no significant interaction between functional groups that cause characteristic peaks to shift in the spectra.[20] Signal-to-Noise ratio (S/N) can also be an important variable and was considered. S/N can vary depending on the acquisition speed, the light source's intensity, the sample and spectrometer environment, and the spectrometer used. We introduced simulated noise into the spectra as a source of non-ideality, first by randomly assigning deviations from zero to a maximum value of ±0.05 A.U. to the absorbance values at each wavenumber and then multiplying these deviation values by a noise factor, NF, that ranges from 0 (no noise introduced) to 1 (highest noise). NF was used to evaluate the performance of the ML algorithms under different amounts of noise. Hereafter, we refer to computer-generated spectra generated as described above simulated samples or simulated spectra to distinguish them from experimentally measured spectra.

2.1.1. Data preprocessing: dimensionality reduction. Given the large number of predictor variables (2760 absorbance values between 4000–1000 cm^{-1}), we implemented a principal component analysis (PCA) to reduce the dimensionality of the

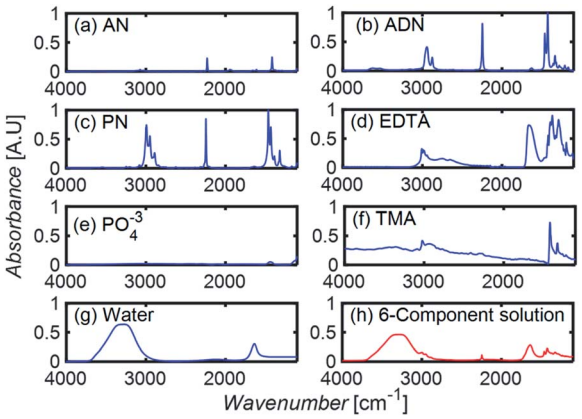

Fig. 2 (a–g) The FTIR absorption spectra of pure components. (h) The spectrum resulting from a linear combination of spectra in (a–g), using a molar concentration of 0.05% for each component.

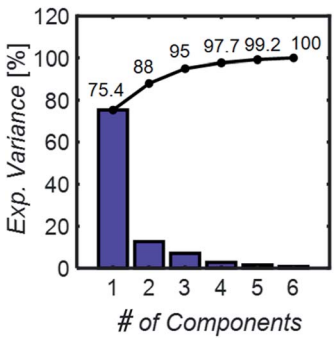

Fig. 3 The principal component analysis explained variance, individual (blue bars), and cumulative (black line). Only 6 principal components are necessary to capture the variance in data sets with spectra from 6 chemical component mixtures.

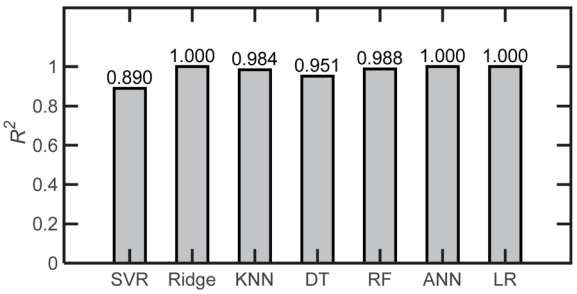

Fig. 4 Coefficient of determination, R^2 averaged over the 6 components for different ML algorithms and for a base case of AN, ADN, and PN in water, applying PCA as a preprocessing step. Two hundred simulated spectra and an 80–20% train-test partition were used in the analysis.

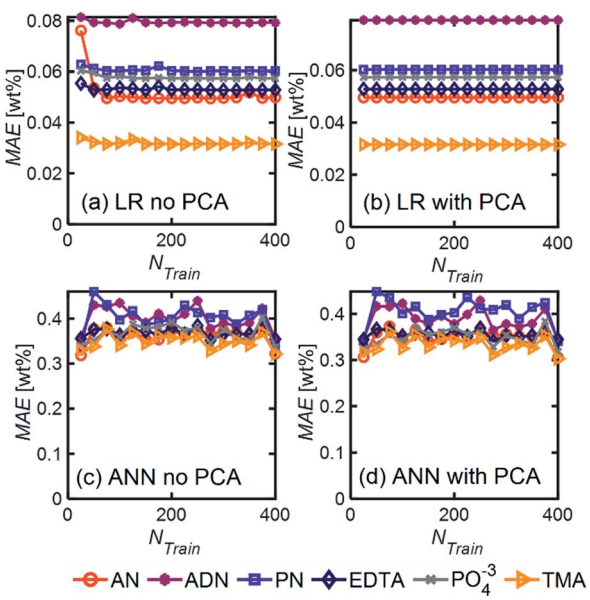

Fig. 5 MAE as a function of training set size for (a, b) LR and (c, d) ANN with and without PCA. These results were obtained using simulated spectra for our model 6-component mixture with NF = 0.5. For ANN, we used 1 hidden layer with 12 neurons with the 'relu' activation function and a batch size of 60. When PCA was applied, only 6 PCs were included in the analysis.

data set, simplifying the model and possibly enhancing its robustness. PCA is a dimensionality reduction technique that groups linearly dependent predictors and outputs a set of linearly uncorrelated principal components (PCs) that represent the directions of the data with the maximum variance.[21] Fig. 3 shows the individual and accumulated explained variance of the PCA for our model 6-component mixture. Six principal components account for nearly all (100%) of the explained variance, consistent with the number of components in the mixtures. Fig. S2 in the ESI† shows the explained variance per component for the other simulated samples.

2.1.2. Model selection. We considered and evaluated seven different regression models to determine the most robust and accurate ML approach. We used a base case of noise-free (NF = 0) 200 simulated ternary solutions of AN, ADN, and PN in water for this evaluation. Fig. 4 shows the mean absolute error (MAE) and the coefficient of determination R^2 for the seven different ML algorithms, including Support Vector Regression (SVR), Ridge Regression, *k*-Nearest Neighbors (*k*-NN), Decision Trees (DT), Random Forests (RF), Linear Regression (LR), and Artificial Neural Networks (ANN). Ridge Regression, ANN, and LR performed the best with MAE ~0.00% and $R^2 \sim 1.00$. LR and ANN were selected for subsequent evaluation based on their simplicity and potential ability to handle non-idealities in experimental data sets, respectively.[22]

2.1.3. Effect of the number of training points. Fig. 5 shows the dependence of the model performance (*i.e.*, MAE) on the number of spectra (N_{train}) in the training set for our model 6-component mixture. For this evaluation, NF = 0.5 was chosen to simulate noise in experimental data. For the case of LR without PCA, performance stabilized for $N_{train} \geq 50$, while for LR with PCA, performance was nearly independent of training size for the datasets with $N_{train} \geq 25$. In the case of ANN, there was no clear trend between training set size and MAE, but there is higher variability between training set sizes. While the application of PCA had no noticeable effects on the performance of ANN, computational time was reduced by a factor of 10 when PCA was used.

Even in the presence of significant noise, LR performed better than ANN, with a smaller MAE between a factor of 5–10, depending on the component of interest. In LR models, TMA had the lowest MAE, which can be attributed to the substantial differences between its spectrum and other components in the range of 4000–1000 cm^{-1}, which results in a simpler differentiation. On the other hand, ADN concentration has the highest MAE given its multiple overlapping peaks with PN and AN and the lower magnitude of the peaks in the fingerprint region, which are more severely affected by noise.

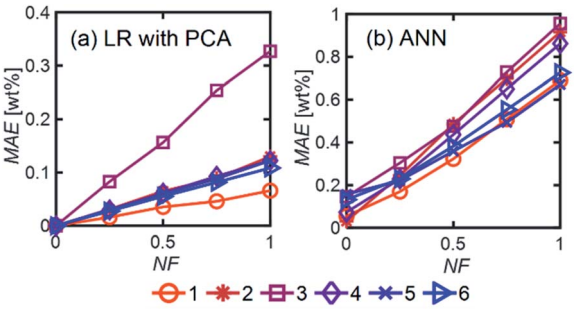

Fig. 7 MAE (averaged over all components) as a function of NF for (a) LR with PCA and (b) ANN over mixtures with different numbers of components and using $N_{train} = 400$. For ANN, we used 1 hidden layer with 12 neurons with the 'relu' activation function and a batch size of 60.

2.1.4. Effect of simulated noise. Noise can reduce the quality of FTIR spectra and complicate analysis. Thus, it is important to determine its impact (*i.e.*, the magnitude of NF) on the ML model prediction accuracy. Fig. 6 shows the effect of NF on MAE for the six chemicals in our model mixture. For LR with PCA and ANN, the prediction accuracy decreased for all six chemical components with increasing noise, but the MAE remained relatively low (<0.15 and 0.8 wt% for LR with PCA and ANN, respectively). For ANN, the dependence of MAE on NF did not vary significantly from component to component. On the other hand, for LR, the increase in MAE with noise was the steepest for ADN and the least steep for TMA.

2.1.5. Effect of the number of chemical components. To determine the robustness of the ML methodology with numbers and identities of the chemical components, we characterized

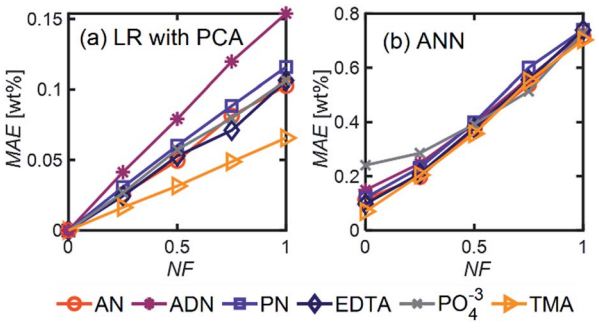

Fig. 6 MAE as a function of NF for (a) LR with PCA and (b) ANN using the model 6-component mixture with $N_{train} = 400$. For ANN, we used 1 hidden layer with 12 neurons, 'relu' activation function, and a batch size of 60.

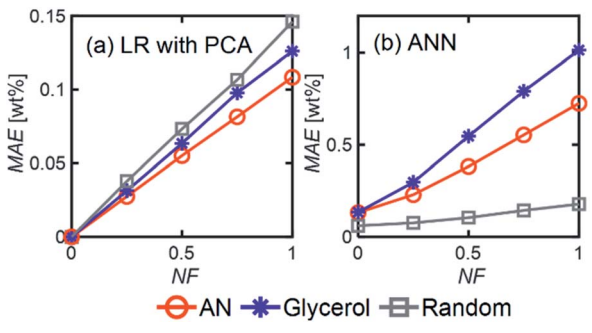

Fig. 8 MAE (averaged over all components) as a function of NF for (a) LR with PCA and (b) ANN models applied to different chemical mixtures and using $N_{train} = 400$. For ANN, 1 hidden layer with 12 neurons was used with an 'relu' activation function and a batch size of 60.

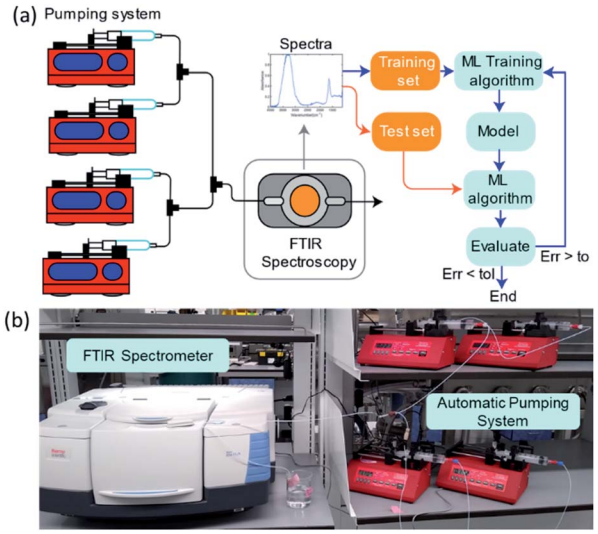

Fig. 9 (a) A schematic illustration of the experimental set-up with four pumps. Components are mixed using T-mixers and then delivered to a transmission flow cell, with ZnSe windows, inside the sample compartment of an FTIR spectrometer. The collected spectra are used to train and test the ML model for concentration prediction. (b) Photograph of the pumping system and FTIR spectrometer.

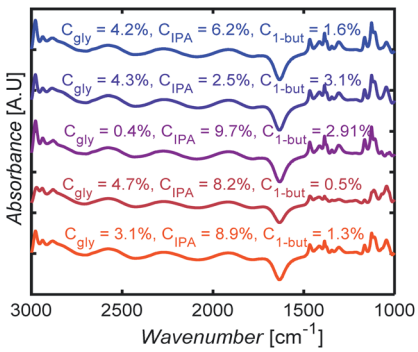

Fig. 10 FTIR spectral measurements for glycerol (gly), isopropanol (IPA), and 1-butanol (1-but) mixtures at different mass concentrations, C_{gly}, C_{IPA}, $C_{1\text{-but}}$, respectively.

the prediction MAE (averaged over all the components in the mixture) of models trained with varying numbers of chemical components and as a function of NF (Fig. 7). Table S1† shows the components used in each one of the mixtures considered. For LR, one component (in addition to water as solvent) showed the least sensitivity to noise, while the 3-component system of AN, ADN, and PN was the most affected by noise due to the similarity of these three components. The averaged MAE is lower in 4–6 component mixtures because the errors associated with EDTA, TMA, and PO_4^{3-} are smaller than those in nitrile-containing components. In ANN models, the sensitivity to the number of components was not as pronounced, but the averaged MAEs were higher than those in LR models.

2.1.6. Effect of type of chemical system. We also studied if the findings from the model 6-component nitrile mixtures were transferable to mixtures containing other molecules and functional groups. To this end, we compared the nitrile-containing mixtures relevant to AN electroreduction with (i) a mixture relevant to glycerol electrooxidation, consisting of glycerol and five possible electrooxidation products, and (ii) a mixture containing six randomly selected molecules. For the "random" case, molecules were selected from a directory containing 21 organic species spectra using random sampling. The species for these cases are shown in Table S2.†

LR MAE as a function of NF behaved similarly for all three types of mixtures, but for ANN, the MAE of the models for the random mixture outperformed the other two, especially at high noise levels (Fig. 8). This is likely because random molecules do not necessarily have similar functional groups (fewer overlapping characteristic peaks), which makes it easier for the algorithm to differentiate between them.

2.2. Experimental implementation of ML methodology

To systematically collect spectra for training the ML models, we used a network of programmable pumps that flowed solutions of selected components with known concentrations into a transmission FTIR flow cell (Fig. 9). A deionized water background was used as a reference. Based on the programmed flowrates and the spectral measurements, we collected ~50 labelled spectra per day, which were then used to obtain LR or ANN regression models. The ML models were developed by partitioning the data randomly into training and testing sets, applying PCA, and then evaluating their performance using the prediction accuracy for the test set. This process was repeated, and new hyperparameters were determined at each iteration until the error was lower than a set tolerance or until the performance stopped improving. The absorbance values at each wn were used as the predictors, and mass concentrations (in wt%) were used as the predictions. The absorbance data

Table 1 Description of types of solutions studied according to species, number of principal components for preprocessing, and total experimental points collected

Mixture label	Species	Number of PC selected	Experimental spectra collected
1-Gly	Glycerol	2	30
2-AN	AN, ADN	3	50
3-Gly	Glycerol, IPA, 1-butanol	5	109
3-AN	AN, ADN, PN	5	67
4-AN	AN, ADN, PN, glycerol	7	50

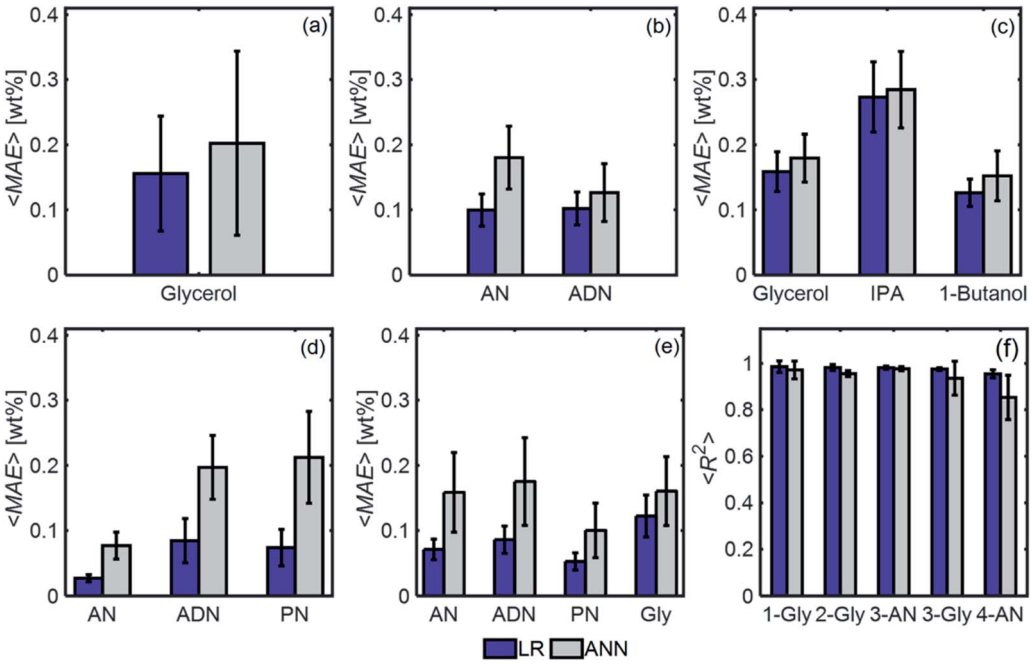

Fig. 11 (a–e) ⟨MAE⟩ [wt%] for 1-, 2- and 3-component mixtures, for ANN and LR models. Over 200 models were evaluated, each with a different train/test subset, and the average MAE is reported. (f) R^2 for the four mixtures considered.

range was limited to between 3000 and 1000 cm^{-1} because absorption saturated outside of this range.

This methodology allowed for the collection of 50 data points per day. An operator was in charge of collecting and labelling samples and refilling the syringes with the single component solutions once they were depleted. This methodology allowed for the autonomous collection of at least 50 data points per day, with human intervention only required to fill the syringes with single component solutions initially. This methodology also allows us to use entire IR spectral measurement as input for our ML models, without needing to select characteristic absorption regions and circumvents the problem of overlapping features of classical approaches.

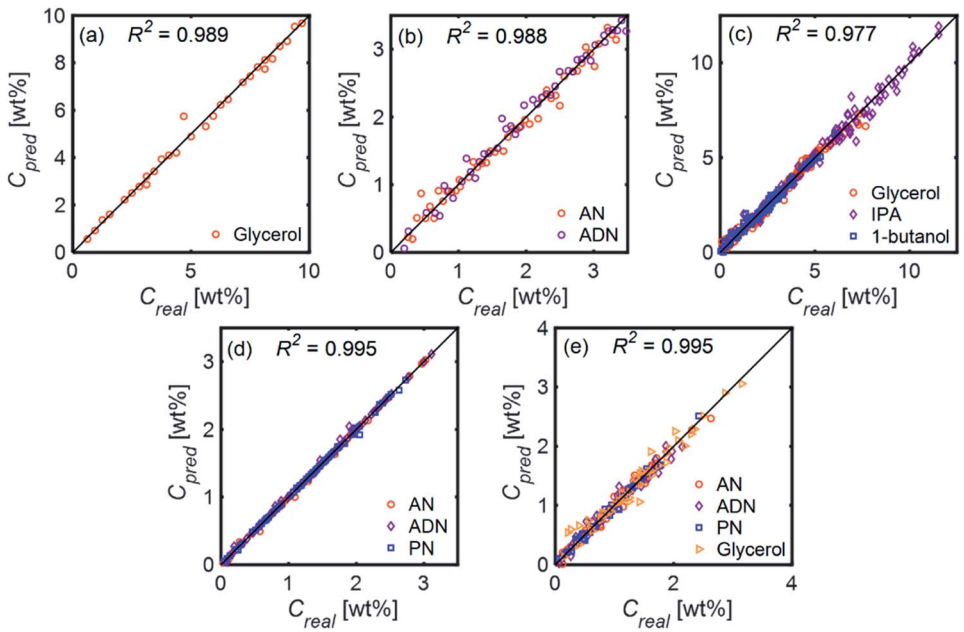

Fig. 12 Predicted (C_{pred}) compared to real (C_{real}) mass concentrations [wt%] for LR models on (a) 1-Gly, (b) 2-AN, (c) 3-Gly, (d) 3-AN, and (e) 4-AN mixture.

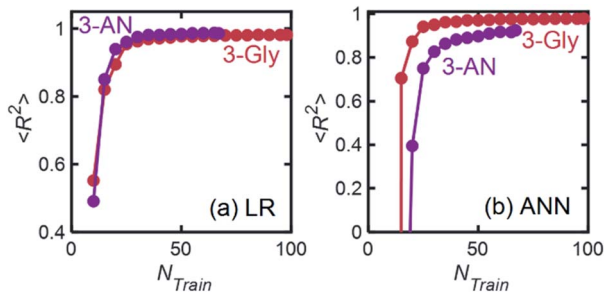

Fig. 13 Model performance in terms of $\langle R^2 \rangle$ as a function of training set size for 3-Gly and 3-AN mixtures. For each point, 200 models were trained, and thus the average R^2 is reported.

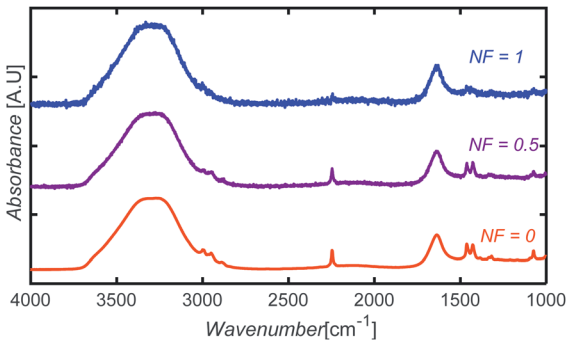

Fig. 14 Illustration of simulated noise introduction to spectra at three different NF levels for an aqueous solution containing AN, ADN, and PN.

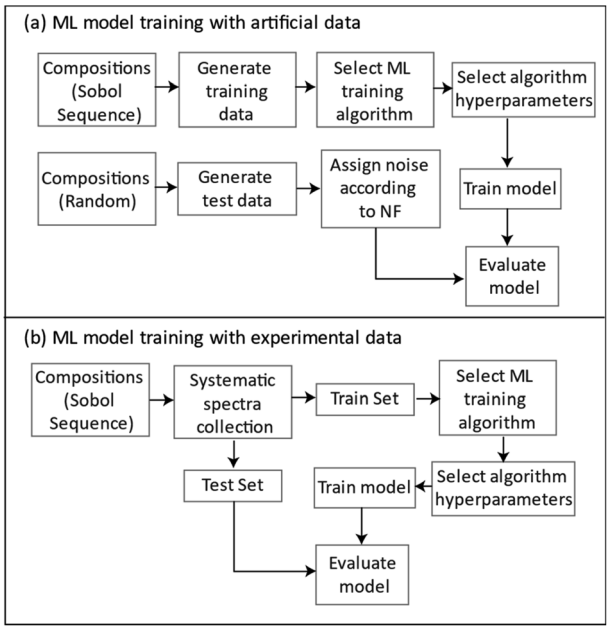

Fig. 15 A flowchart for the general approach to ML model development with (a) simulated generated and (b) experimentally collected data.

Fig. 10 shows selected FTIR spectra of 3-component mixtures with different compositions. The spectra look similar to the eye, with subtle changes in the intensities of some peaks. Without ML models here, one would have to carefully identify peaks for each species, correct for baseline, deconvolute and fit peaks, a nontrivial and arduous task to determine mixture compositions.

We show, however, that ML models with PCA can determine unknown compositions from spectra similar to these. We studied five different aqueous solutions differing in numbers and types of components in the mixture. Table 1 shows the species in the aqueous solution for each of the cases studied.

2.2.1. Linear regression and ANN results. We implemented the LR and ANN algorithms with PCA to analyze the experimentally acquired spectra of mixtures with different compositions because these algorithms performed well when using simulated spectra. Models were trained with 80% of the spectra and then tested with the remaining 20%. We ran the training algorithm 200 times, randomly selecting different sets for training and testing. Here we report the average performance metrics, $\langle MEA \rangle$ and $\langle R^2 \rangle$.

Fig. 11 compares the performances of LR implemented with PCA and ANN for the mixtures in Table 1. The $\langle MAE \rangle$ of the concentrations predicted [wt%] ranged from 0.023% to 0.28%. The $\langle MAE \rangle$ for glycerol-based mixtures did not significantly change between 1-component and 4-component mixtures. The $\langle R^2 \rangle$ values varied between 0.854 and 0.986, decreasing as the number of components increased. LR models had higher accuracy and weaker dependence on train/test subset combinations than ANN models for all mixtures. Fig. 12 shows the predicted and actual concentrations for the mixtures in Table 1. The subplots in this figure depict a model trained with a randomly chosen subset of the entire spectral data set, while the results in Fig. 11 show $\langle MAE \rangle$ averaged over 200 models.

Table 2 Scikit-learn functions used for each ML model algorithm

Model	Function
Linear regression (LR)	sklearn.linear_model
Multilayer perceptron regression or artificial neural networks (ANN)	sklearn.neural_network.MLPRegressor
Decision trees	sklearn.tree.DecisionTreeRegressor
Random forests (RF)	sklearn.ensemble.RandomForestRegressor
Support vector regressor (SVR)	sklearn.svm.SVR
Ridge regression with cross validation (RidgeCV)	sklearn.linear_model.RidgeCV
k-Nearest neighbors (*k*NN)	sklearn.neighbors.KNeighborsRegressor

2.2.2. Effect of number of training points. To understand the training data size requirements to produce accurate ML models, we evaluated the performance of the algorithms in terms of the $\langle R^2 \rangle$ for models trained with different numbers of spectra for two types of ternary aqueous solutions: an AN-based mixture (3-AN) and a glycerol-based mixture (3-Gly). Fig. 13 shows that $\langle R^2 \rangle$ *vs.* N_{train} rapidly increases for these two types of mixtures but eventually saturate at ~40 spectra. ANN is more sensitive to the training data size than LR and requires more training spectra for accurate predictions.

3. Conclusions

We described a general methodology for developing and implementing ML models for quantitatively predicting chemical mixture compositions from their FTIR spectra. For model mixtures chosen from practical applications, we trained linear regression (LR) and artificial neural network (ANN) models with R^2 regression scores ranging from 0.98 to 0.99 and 0.94 to 0.98, respectively. Simpler and less computationally expensive linear regression models were consistently more accurate than ANN models, making them a superior choice for quantitative composition prediction from FTIR spectra. We also studied the relationship between model performance and the number of spectra in the training data set and found that for both LR and ANN, regression scores increased and saturated at approximately 40 spectra for 3-component mixtures. Finally, we showed that trained ML models (Linear Regression with PCA and Neural Networks) maintain their accuracy despite small variations in experimental conditions expected over several days. Our results suggest that this methodology can enhance the analytical capabilities of FTIR spectroscopy for quantitative composition determination and find applications in inline chemical analysis applications that require fast characterization, such as autonomous chemical process development and optimization.

4. Experimental methods

4.1. Materials

Acrylonitrile (AN), adiponitrile (ADN), propionitrile (PN), 1-butanol, and glycerol were purchased from Sigma Aldrich. Isopropanol 70% was purchase from VWR. Stock solutions were prepared with deionized (DI) water.

The pumping system consisted of two NE-1000 Programmable Syringe Pumps and two NE-4000 Programmable 2-Channel Syringe Pumps, manufactured by New Era Pump Systems: 60 ml and 30 ml BD syringes were used to load the stock solutions into the system. A Nicolet iS50 FTIR Spectrometer and OMNIC software were used for spectral data collection. The transmission flow cell was purchased from Harrick Scientific Products and consisted of a demountable liquid cell with Luer lock fittings and a 20 mm diameter clear aperture, equipped with a pair of 25 mm diameter ZnSe transmission windows. For all experiments, the spacing between the transmission windows was 12 μm.

4.2. Simulated data generation

Simulated spectral data for mixtures of selected components were generated using Beer's law (eqn (1)). For the training set, a concentration matrix, $\bar{C}$, with dimensions $p \times (n+1)$, where p is the number of points to generate, and n is the number of different components to consider, was generated according to a Sobol sequence.[23] For the test set, a concentration matrix was created based on random distribution sampling. Compositions of individual solutes were maintained below 10% with water as a solvent. Applying a dot product between $\bar{C}$ and a vertically concatenated matrix of the spectral data of the individual components $\bar{A}_{\text{pure}}$, results in a matrix of spectra, S, where each row is a new spectrum corresponding to a mixture of known concentrations.

$$\begin{bmatrix} C_{11} & \cdots & \cdots & 1-\sum_{i=1}^{n-1} C_{1i} \\ C_{21} & \cdots & \cdots & 1-\sum_{i=1}^{n-1} C_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ C_{p1} & C_{p2} & \cdots & 1-\sum_{i=1}^{n-1} C_{pi} \end{bmatrix} \times \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pm} \end{bmatrix} \overline{C} \times \overline{A}_{\text{pure}} = S \qquad (2)$$

4.3. Simulated noise introduction

To introduce noise to the simulated test data, we defined a variable noise factor, NF, ranging from 0 (no noise assigned) to 1 (maximum noise-to-signal ratio). A number between −0.05 and +0.05 A.U. was randomly selected, multiplied by NF, and then added to each absorbance point of a spectrum. The noise range was selected based on the difference observed between the FTIR spectrum obtained from spectral libraries and the spectrum of a glycerol sample collected experimentally in our equipment using only five scans. Fig. 14 shows sample spectra at three different NF levels for an aqueous solution containing AN, ADN, and PN.

4.4. Data preprocessing: principal component analysis

Principal component analysis (PCA) was used as a dimensionality reduction technique to decrease the number of spectral data points from thousands to up to 10 principal components for the studies conducted with simulated and experimental data. The number of principal components selected depended on the number of chemical components in the solution under study. PCA was implemented using the sklearn.preprocessing.PCA() function from scikit-learn, an ML library for Python.[24]

4.5. Machine learning algorithm training and evaluation

Machine learning models were developed to describe relationships between solution compositions and their FTIR absorbance spectra. Different ML regression algorithms available in the scikit-learn library were initially evaluated for a base case comprising 200 simulated spectra of tertiary mixtures in water, with an NF = 0. The algorithms and respective scikit-learn functions are described in Table 2.

Hyperparameters were optimized using sklearn.model_selection.RandomizedSearchCV. When developing regression models, the predictors or features were the absorbance values at each wavenumber, a matrix denoted S, and the target or predicted variables were the concentrations corresponding to each spectrum, contained in a concentration vector (for 1-component solution) or matrix (for a multicomponent solution) denoted C. For the experimentally collected data, S and C were divided randomly into a training and a test set, with a training : test ratio of 80 : 20%. To avoid model performance dependency on the random training to test partition, each study was repeated 200 times, after which the average metrics were calculated and reported. The infrared wavenumber range for the simulated and experimental data were 4000–1000 cm^{-1} and 3000–1000 cm^{-1}, respectively, the latter omitting the 4000–3000 cm^{-1} range where the noise is very high due to nearly complete absorption by the water O–H stretching vibration.

Fig. 15 summarizes the general approach for developing ML regression models for the simulated and the experimentally collected data.

4.6. FTIR experimental data collection

Spectral measurements of mixtures of known concentrations were pumped into a transmission flow cell placed inside the FTIR spectrometer using a network of programmable pumps, each loaded with a single component aqueous stock solution. Concentrations of the mixture flowing through the cell were changed and controlled by varying the flow rates of the individual single-component solutions. The pumps were programmed to switch flow rates periodically at set intervals, allowing for automated spectra collection while varying compositions. For a two-component mixture, the total flow rates were maintained at 1 ml min^{-1}, 1.5 ml min^{-1}, and 2 ml min^{-1} for two-, three- and four-component mixtures, respectively. The set of compositions to sample were determined using a Sobol sequence. New sampling intervals were determined every time a new component was introduced by pumping a new solution into the flow cell and periodically taking spectral measurements until the resulting spectrum stopped changing over time. All spectra were taken with respect to the water background. Deionized water background was recorded only once at the beginning of each sampling collection session, which typically lasted for about 6 hours at the most. Datasets for one type of mixture were collected during 4 days (3-gly). Performance for the 3-gly mixtures specifically was 0.982 and 0.977 for LR and ANN, which suggests that the same model can be used for experimental campaigns that span several days without the need for recalibration.

The set of compositions to sample were determined using a Sobol sequence.

The code and data for simulated noise generation and introduction, data preprocessing, machine learning algorithm training and evaluation and collected FTIR spectral data are available in a public repository.[25]

Data availability

(1) The code for machine learning regression models development, both for synthetic and experimental data, can be found at https://doi.org/10.5281/zenodo.5498197 with 10.5281/zenodo.5498197. The version of the code employed for this study is version V1.0.0. (2) Data for the generation of the Machine Learning Regression Models, including.csv files for the experimentally collected data for chemical mixtures of known concentration, as well as the corresponding labels, are available at 10.5281/zenodo.5498197 at https://doi.org/10.5281/zenodo.5498197.

Author contributions

The manuscript was written through the contributions of all authors. All authors have approved the final version of the manuscript. Andrea Angulo and Lankun Yang performed the experimental measurements. Andrea Angulo developed the Machine Learning methodology. Eray Aydil and Miguel Modestino conceptualized this work.

Conflicts of interest

MAM is a director and has a financial interest in Sunthetics, Inc., a start-up company in the chemical process optimization space. New York University intends to pursue intellectual property protection for parts of the material presented in this work.

Acknowledgements

The authors acknowledge the financial support provided by the National Science Foundation (Grant # CBET-1943972) and from NYU, Tandon School of Engineering, through the MAM and ESA startup funds. In addition, the collaboration between MAM and ESA is enabled by the Center for Decarbonizing Chemical Manufacturing Using Electrification (DC-MUSE), formed with the help of a generous grant from the Sloan Foundation (Grant # 201-16807) and a center planning grant from the National Science Foundation (Grant # EEC-1936709).

Notes and references

1 F. Häse, L. M. Roch and A. Aspuru-Guzik, *Trends Chem.*, 2019, **1**, 282–291.
2 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2020, **59**, 22858–22893.
3 Z. Zhou, X. Li and R. N. Zare, *ACS Cent. Sci.*, 2017, **3**, 1337–1344.

4 A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen and T. F. Jamison, *Science*, 2018, **361**, 1220–1225.
5 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.
6 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237–241.
7 M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik and J. E. Hein, *Commun. Chem.*, 2021, **4**, 112.
8 D. E. Blanco, B. Lee and M. A. Modestino, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 17683–17689.
9 C. Berthomieu and R. Hienerwadel, *Photosynth. Res.*, 2009, **101**, 157–170.
10 S. Kern, S. Liehr, L. Wander, M. Bornemann-Pfeiffer, S. Müller, M. Maiwald and S. Kowarik, *Anal. Bioanal. Chem.*, 2020, **412**(18), 4447–4459.
11 F. Lussier, V. Thibault, B. Charron, G. Q. Wallace and J.-F. Masson, *TrAC, Trends Anal. Chem.*, 2020, **124**, 115796.
12 S. Kiyohara, M. Tsubaki, K. Liao and T. Mizoguchi, *J. Phys.: Mater.*, 2019, **2**, 024003.
13 M. M. Y. R. Riad, Y. M. Sabry and D. Khalil, *2019 36th National Radio Science Conference (NRSC)*, 2019, pp. 386–392.
14 E. Bona, I. Marquetti, J. V. Link, G. Y. F. Makimori, V. da Costa Arca, A. L. G. Lemes, J. M. G. Ferreira, M. B. dos Santos Scholz, P. Valderrama and R. J. Poppi, *LWT–Food Sci. Technol.*, 2017, **76**, 330–336.
15 Y. Liu, F. Wang, C. Shao, W. You and Q. Chen, in *International Conference on Mechatronics and Intelligent Robotics*, 2019, vol. 856, pp. 784–791.
16 N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson and R. Subramanian, *Atmos. Meas. Tech.*, 2018, **11**, 291–313.
17 Y.-T. Wang, B. Li, X.-J. Xu, H.-B. Ren, J.-Y. Yin, H. Zhu and Y.-H. Zhang, *Food Chem.*, 2020, **303**, 125404.
18 L. D. Ellis, S. Buteau, S. G. Hames, L. M. Thompson, D. S. Hall and J. R. Dahn, *J. Electrochem. Soc.*, 2018, **165**, A256–A262.
19 D. E. Blanco and M. A. Modestino, *Trends Chem.*, 2019, **1**, 8–10.
20 T. G. Mayerhöfer and J. Popp, *Appl. Spectrosc.*, 2020, **74**(10), 1287–1294.
21 M.-L. O'Connell, A. G. Ryder, M. N. Leger and T. Howley, *Appl. Spectrosc.*, 2010, **64**(10), 1109–1121.
22 A. Krogh, *Nat. Biotechnol.*, 2008, **26**, 195–197.
23 bstemper, naught101 and M. Osthege, *GitHub Repository*, https://github.com/naught101/sobol_seq.
24 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
25 aaf431, *aaf431/FTIR_and_Machine_Learning: Machine Learning Enhanced Spectroscopic Analysis*, 2021, DOI: 10.5281/zenodo.549819.