OPEN ACCESS
IOP Publishing
Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **55** (2022) 225205 (14pp)
https://doi.org/10.1088/1361-6463/ac5770

Methane detection to 1 ppm using machine learning analysis of atmospheric pressure plasma optical emission spectra

Tahereh Shah Mansouri[1,*], Hui Wang[2], Davide Mariotti[1] and Paul Maguire[1]

[1] NIBEC Engineering, Ulster University, Belfast, United Kingdom
[2] Queens University, Belfast, United Kingdom

E-mail: t.shah_mansouri@ulster.ac.uk

Received 16 September 2021, revised 7 February 2022
Accepted for publication 22 February 2022
Published 7 March 2022

Abstract

Optical emission spectroscopy from a small-volume, 5 μl, atmospheric pressure RF-driven helium plasma was used in conjunction with partial least squares-discriminant analysis for the detection of trace concentrations of methane gas. A limit of detection of 1 ppm was obtained and sample concentrations up to 100 ppm CH_4 were classified using a nine-category model. A range of algorithm enhancements were investigated including regularization, simple data segmentation and subset selection, feature selection via Variable Importance in Projection and wavelength variable compression in order to address the high dimensionality and collinearity of spectral emission data. These approaches showed the potential for significant reduction in the number of wavelength variables and the spectral resolution/bandwidth. Wavelength variable compression exhibited reliable predictive performance, with accuracy values >97%, under more challenging multi-session train—test scenarios. Simple modelling of plasma electron energy distribution functions highlights the complex cross-sensitivities between the target methane, its dissociation products and atmospheric impurities and their impact on excitation and emission.

Keywords: methane detection, optical emission spectroscopy, atmospheric pressure plasma, partial least squares, machine learning

(Some figures may appear in colour only in the online journal)

1. Introduction

Gas identification and in particular the detection of trace levels of molecular components in gases has gained increasing attention in many fields from atmospheric pollution and climate change monitoring to industrial safety [1] and breath analysis for clinical diagnosis [2]. There are a number of established techniques including mass spectrometry (MS) [3], gas chromatography [4] optical spectroscopy, electrochemical [5], solid-state and optic fibre [6], that have inspired the development of a wide range of technologies in each category. Laser absorption and spectroscopic detection methods such as non-dispersive IR absorption (NDIR) or Raman have allowed limits of detection (LOD) in the low ppm to ppb range to be achieved. Improvements in, for example, mid-IR quantum cascade laser technology and photoacoustic detectors will enable continued reduction of LOD. Among the

* Author to whom any correspondence should be addressed.

spectroscopic techniques, tuneable diode laser IR absorption spectroscopy (TLDAS) and atomic emission spectroscopy ICP-AES are well-established and routine laboratory techniques. Apart from improving LOD and increasing the number of target species, there is a major drive towards system miniaturisation and cost reduction in order to achieve field deployable gas detection capability e.g. for rapid and continuous environmental monitoring via autonomous distributed networks or point of care clinical breath screening. For example, methane is a high priority greenhouse gas with stringent targets for reduction, including reducing CH_4 emissions from e.g. landfill, oil and natural gas industries [7–10]. Field deployment of high-resolution detectors and remote autonomous monitoring is a major priority yet remains elusive due to the very high system cost [10]. This has inspired the search for high accuracy miniaturised systems. The ARPA-E (US) MONITOR programme has funded development of various technology strands including compact IR spectrometry, compact MS, hollow-core optic fibre and low cost printed nanomaterials [11]. Methane is also an important breath biomarker and detection of trace CH_4 levels is a major challenge. Recently Dong *et al* reported a compact trace CH_4 detection system based on TLDAS with distributed feedback interband cascade lasers in a 5 l volume package [12]. However, these compact systems remain costly when considered for autonomous field deployment. Detector arrays based on high porosity and high surface area nanomaterials have been proposed as a low-cost electronic nose platform for breath and environmental analysis. High sensitivity has been achieved when coupled with machine learning [13] but systems struggle with lifetime, species interference and cross-sensitivity (e.g. temperature and humidity) [14, 15]. The use of plasmas as atomic and molecular sources for optical emission spectroscopy (OES) and MS has a long history, with the ICP-AES technique being the most popular. Samples in the form of liquid or solid particles are introduced into the hot plasma resulting in vaporisation or evaporation, excitation and ionisation and provide either the photons for OES or ions for MS [16]. Low pressure glow discharge plasmas and laser induced breakdown spectroscopy are typically employed to produce OES species from solid surfaces.

Recent progress in the design and control of miniaturised atmospheric pressure plasmas systems has encouraged their application to new research fields such as e.g. plasma-based medicine, agriculture, gas reforming, catalysis, advanced nanomaterials and environmental pollution control [16–18]. The associated plasma devices are of simple construction, small, low cost and operate at atmospheric pressure and at low temperatures. They can also provide high intensity light emission and therefore have the potential to act as an optical emission source for trace gas detection. Hyland *et al* first reported plasma OES with machine learning for trace gas detection and recognition, using spectra from a range of trace volatiles fed into neural networks [19]. Weagant *et al* investigated the use of a low power atmospheric pressure Ar–H_2 microplasma and portable spectrometer to detect trace metal impurities. Liquid samples were dried then electrothermally vaporised into the plasma. Simple spectra resulted, dominated by low excitation energy lines. However maintaining reproducible line intensities and limiting background emission was difficult [20]. Similar trace metal detection in liquid has been demonstrated using RF-excited glow discharge emission spectroscopy at atmospheric pressure [21]. Here the sample is dried and then ablated by a plasma operating at gas temperature up to 1500 °C. The observed spectral variability was up to 28%. Atmospheric pressure plasmas in contact with water and complex liquids have been investigated for rapid, lower cost and low power analytical atomic spectrometry of metals [22–26]. High plasma densities ($>10^{20}$ m^{-3}) [27] and relatively high gas temperatures (600 K–1100 K) are involved [27, 28] and mechanisms depend on liquid evaporation and droplet formation while the inclusion of organic species can enhance emission or produce volatile species containing the elements for detection [29].

While trace metal detection by miniaturised plasmas may offer low cost portable alternatives to ICP-AES, the trace gas detection of molecular and complex volatile constituents represents a much greater challenge since microplasma emission spectra are very complex, individual lines are weak and poorly resolved, especially for non-equilibrium low temperature (NELT) plasma devices. High resolution OES of NELT plasmas containing molecular mixtures is often used to fit observed to simulated spectra in order to determine internal plasma parameters such as gas rotational and vibrational temperatures [30] as well as electron temperature and density [31]. We have carried out such analysis on similar plasma devices to that used here [32, 33]. However with objectives such as portability, low cost, field deployability and possible autonomous operation, the intrinsic complexity of the spectra and the temporal variation in plasma conditions under uncontrolled conditions need to be considered. Using design constraints and operating parameters that maintain low gas temperatures (<50 °C), e.g. for breath analysis or managing safety concerns with flammable gases, adds further noise and complexity to spectra. Knowledge of NELT plasma chemistry is very limited and generating accurate simulated spectra for molecular gases and mixtures, especially at trace concentrations below 100 ppm, is not feasible. This coupled with the use of low-cost limited resolution spectrometers presents a major impediment to accurate detection and to date, the use of OES with NELT plasmas to determine the trace molecular constituents of a gas has not been considered. Kudryavtsev *et al* used a current probe technique integrated into a helium microplasma, for CO_2 gas analysis via collisional electron spectroscopy. This involved measurement of the high energy portion of the electron energy distribution function (EEDF) to determine He metastable reactions with impurities. CO_2 detection at concentrations $\geqslant$500 ppm was achieved. However the plasma was operated below atmospheric pressure [34]. Recently, we demonstrated the feasibility of using OES to detect the presence of methane above threshold values in the low ppm range using a helium NELT plasma jet coupled with spectral analysis via machine learning techniques [35].

Spectral data is often used to help determine the constituents of materials and can consist of, for example, measured values of radiation or mass intensity at fixed discrete wavelengths or mass values, respectively. Chemometric and machine learning techniques have been applied where spectral discrimination is problematic [36]. The interpretation of optical spectra, from UV to far IR, depends on the experimental approach and the instrument resolution. Thus with absorption spectra, the concentration of a target species may be directly related to measured intensity through the Beer–Lambert Law. For emission spectra, a relationship between concentration and intensity, at a specific wavelength, is only possible when the system is in local thermodynamic equilibrium, in which case it is determined from the Boltzmann distribution which relates excited state densities to that of the ground state. High temperatures are therefore required to obtain measurable emission intensities. For the low temperature emission spectra used here, thermodynamic equilibrium is not established and spectral data consist of a large number of lines where there is little *a priori* information about expected line strength and significance. Intensity values will follow a complex non-linear relationship with concentration. Line broadening via intrinsic or instrumental effects will create data values around each peak which may be highly correlated and redundant and/or merge peaks from different excitation states and species. Important molecular gases, including hydrocarbons such as methane, generally have multiple but weak lines in the UV–Vis–NIR region which often overlap with spectral lines from plasma carrier gases, such as helium or argon, and impurities. Furthermore, the introduction of molecular gases into a plasma can affect parameters such as electron density and temperature which in turn modify line intensities of atomic and impurity gases (e.g. O_2, N_2 and H_2O dissociation products).

In order to cope with such complexity, we focus on developing machine learning algorithms for analysis of NELT plasma emission spectra which can handle the challenges of high dimensionality, where the number of variables (wavelengths) greatly outnumbers the sample count, nonlinearity, redundancy, collinearity, where individual peaks bleed into multiple nearby data points, and multimodality [37]. Recently, we developed a number of algorithmic approaches based on partial least squares-discriminant analysis (PLS-DA) to characterise reflectance spectral data from portable optical and infra-red systems under uncontrolled and variable field conditions [37–41]. Algorithm performance was also compared with traditional laboratory-based absorption spectra. Emission spectra, by contrast, display a much larger number of sharp well-defined peaks with a wide range of intensities and thus the algorithmic challenges are heightened. Recently, machine learning approaches have been investigated in an attempt to solve critical unresolved challenges in real-time diagnostics and control of cold atmospheric pressure plasmas. These include monitoring vibrational/rotational temperatures and the effects of changing substrate properties on plasma conditions [42, 43], and to determine the electron energy probability function solely from optical emission spectra [44]. Using a coplanar high-voltage AC plasma and spectral analysis based on convolutional neural networks, Wang *et al* demonstrated detection of methanol and acetone in real-time for concentrations above 1487 ppm and 3439 ppm respectively [45]. In this work, we seek to extend our initial feasibility study [35] to identify impurity species at different and lower concentrations using multi-categorical models. Emission spectra from methane in helium mixtures, with concentrations from 0 to 100 ppm, were obtained from a low cost portable NELT plasma device. The gas mixtures also contained trace impurities from air and H_2O of unknown concentration. Machine learning models based on PLS-DA were investigated, using a range of training and test protocols, along with a number of data manipulation and feature selection approaches in order to maximise performance.

2. Experimental methods

CH_4–He spectra were obtained from an RF-excited (13.56 MHz) plasma formed in a quartz capillary between two exterior ring electrodes (separation 5 mm) while helium was used to sustain the plasma, figure 1. The 0.7 mm (ID) capillary outlet was a large distance ($\sim$100 cm) from the plasma to minimise atmospheric impurity back-diffusion. The system was initially conditioned to remove background impurities from the capillary walls, using repeated daily exposure to a 100% argon plasma, over 21 days, followed by isolation and continuous exterior IR heating of the capillary. A two-stage mass flow-controlled gas network was used to dilute methane gas (purity 99.95%) in the carrier gas (He, purity 99.9995%). Two mass flow controllers (MKS, Model 1179 C, precision 0.05% FSD) were used to deliver up to 0.005 SLM of He–CH_4 mixture at a concentration of 100 ppm into a pure He flow up to 0.05 SLM. The overall specified precision at 1 ppm CH_4 was $\pm$1%. The set concentration error included manufacturer specified flow meter error ($\pm$1% FS) and gas supplier (Buse International) specified CH_4 in He mixing accuracy ($\pm$5%). The maximum deviation from set concentration was $+30.9\%$ / -26.8% and the absolute deviation values are shown in figure 1(b). The gas temperature, measured in a similar plasma system, remained below 30 °C [46].

CH_4–He data was collected in two separate datasets, where dataset A comprised 523 samples in nine CH_4 concentration categories (0, 1, 2, 4, 6, 12, 23, 77, 100 ppm) and dataset B comprises 720 samples in eight (0, 1, 2, 4, 6, 23, 77, 100 ppm) CH_4 concentration categories. Spectra in the wavelength range 194 nm–1122 nm (interval 0.25 nm) were obtained using an Ocean Optics HR4000CG-UV-NIR spectrometer (optical resolution <1.0 nm, slit width 5 μm), with a total of 3648 wavelength points recorded. Spectral mean intensity versus wavelength from 0 ppm, 2 ppm and 100 ppm CH_4 sample sets are shown in figure 2(a). Spectral change with concentration is indicated in difference plots, figure 2(b) while intensity change between samples is indicated by the relative standard deviation at each wavelength, figure 2(c). The main spectral lines are listed in table 1 in rank order of intensity at 0 ppm CH_4 and intensity values relative to the intensity of the largest peak at 588 nm are indicated. Using spectral intensity data, species

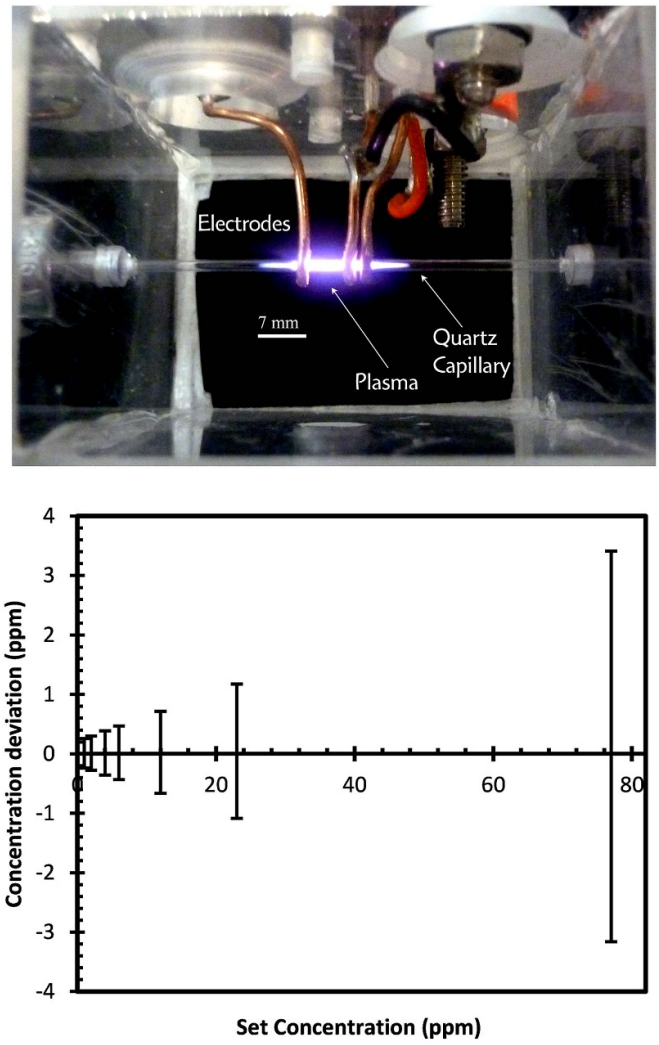

Figure 1. (a) NIBEC RF capillary plasma system operated with He carrier gas at atmospheric pressure. The electrode gap was 5 mm and the internal diameter of the capillary was 0.7 mm, (b) Set concentration deviation due to flow meter specified error and supplied gas mixing accuracy.

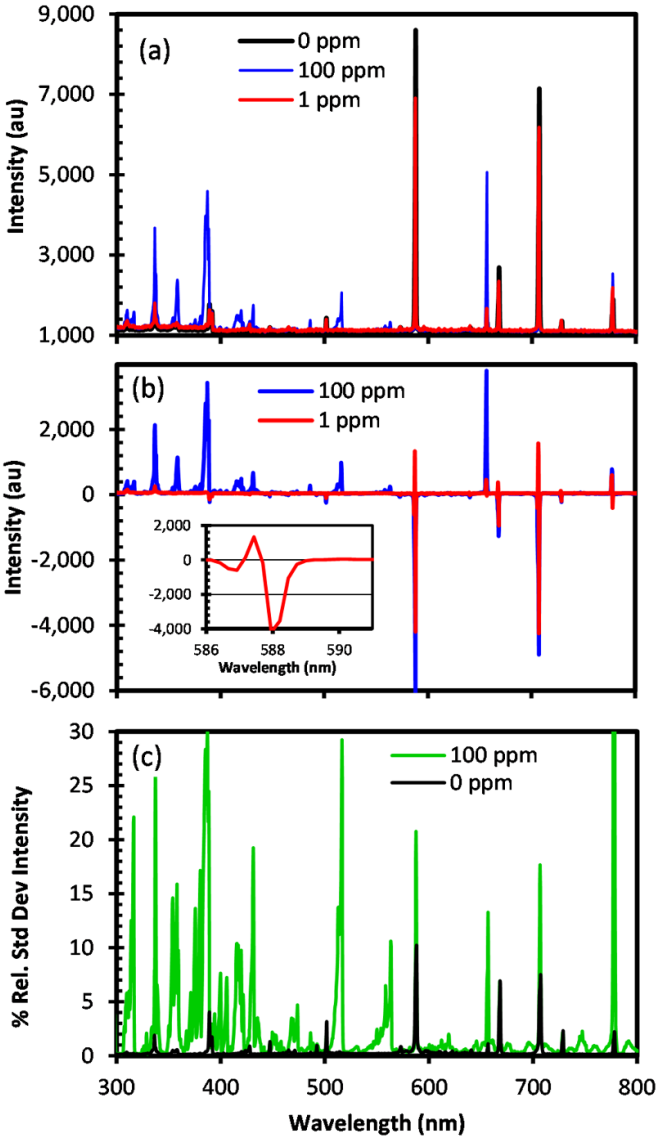

Figure 2. (a) Mean spectra from samples with 0 ppm, 1 ppm and 100 ppm CH_4, truncated to the wavelength range 300 nm–800 nm, (b) Difference plots of 1 ppm CH_4–0 ppm and 100 ppm CH_4–0 ppm. The inset shows the difference around 588 nm for 1 ppm CH_4 and indicates the effect of misalignment as a contributor to the observed difference. (c) % relative standard deviation (SD/mean) for 0 ppm and 100 ppm spectra.

involved in specific transitions are listed [47, 48]. Impurity lines, representing species derived from air (N2, O2, N, O) and water dissociation (OH, H), are noticeable with intensities up to 20% of the maximum. The C_2 (Swan) vibrational bands around 516 nm are only visible at concentrations $\geqslant$77 ppm. The only other detectable lines that may be attributed to CH_4 fragmentation are the CH(A–X) band at 388.90 nm which overlaps with the He line at 388.86 nm and possibly N_2 lines at 389.46 nm. The integrated line intensity taken over the range 388 nm $\pm$3 nm exhibits an approximately constant value at low concentrations, suggesting that CH (A–X) emission may not be significant until 77 ppm, where the intensity is noticeably enhanced [35]. The H_α line intensity at 656.56 nm, which may derive from H_2O dissociation and/or CH_4 fragmentation, varied approximately linearly with CH_4 concentration and at $\geqslant$77 ppm was greater than that of the main He line at 588 nm [35].

3. Computational methods

The overall objective is to develop an algorithmic solution to the task of recognising an unknown spectrum as a member of one category. In an exploratory search, the raw data was subjected to various pre-processing steps. These included Standard Normal Variation, normalization, baseline correction, auto scaling and noise reduction. Initially, we looked briefly at the performance profile of four different algorithmic approaches (PLS-DA, k-nearest neighbour, support vector machine coupled with principal component analysis (SVM-PCA), linear discriminant analysis (LDA)) using the Receiver Operating Characteristic curve, figure 3, for a single category

Table 1. Main OES peaks of CH_4–He listed in rank order of intensity as observed for 0 ppm CH_4 except for features around 516 nm which are only observed at $\geqslant$77 ppm CH_4. The peak wavelengths have been rounded to nearest integer values. The relative intensity column values are calculated with respect to the maximum peak intensity (588 nm) at 0 ppm CH_4. The species column lists the attributed species and for those wavelengths with overlapping species peaks, the species are listed in order of expected intensity.

Rank	Peak wave-length (nm)	Relative Intensity	Species
1	588	1.00	He
2	706	0.79	He
3	667	0.21	Impurity, He
4	778	0.10	Impurity
5	389	0.08	He, CN, N_2, O_2
6	336	0.05	Impurity
7	728	0.04	He, impurity
8	656	0.03	H
9	415	0.02	He I
Selected peaks of intensity rank >9 or which only appear for $CH_4 \geqslant 77$ ppm			
	516	—	C2 Swan
	309	—	OH
	431	—	CH

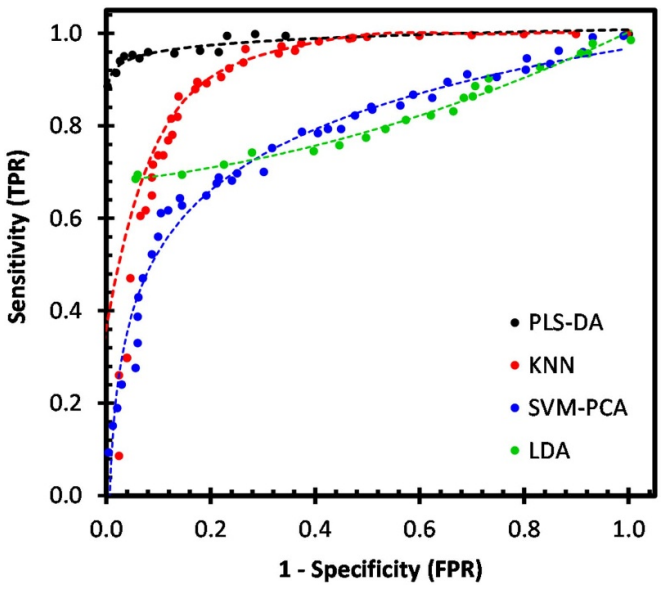

Figure 3. Comparison of Receiver Operating Curves (ROC) for four algorithms (PLS-DA, wKNN, SVM-PCA and LDA) applied to pure He spectra.

(0 ppm CH_4). As we previously found with infra-red spectra [37, 39] the PLS-DA algorithm shows the best ability to distinguish spectral data from any two groups. For example, in this case distinguishing 0 ppm CH_4 concentration from those $\geqslant$1 ppm CH_4, the area under curve was >98% for PLS-DA while LDA [49] showed the poorest classification at 64%. Both weighted k-nearest neighbour [50] and SVM-PCA [51, 52] show intermediate performance.

Table 2. Protocols for implementation of algorithm training, evaluation and testing.

Protocol	Description
1	Model training and cross-validation using dataset A
2	Model training and cross-validation using dataset B
3	Datasets A and B merged. Model training and cross-validation using merged A + B data
4	Model trained and cross-validated using dataset A. Model testing using dataset B

PLS-DA is a classification derivative of PLS regression and is considered a useful algorithm for building predictive models in cases where there is both a large number of parameters and factors which are highly collinear and has been used regularly in the analysis of chemometric data. A Variance Inflation Factor (VIF) greater than ten indicates harmful data collinearity and a reason for concern and almost all CH_4–He spectra data display VIF values >10 [53–56]. PLS models the relationship between an input matrix (X) and an output matrix (Y), to develop an N-dimensional hyperplane in the input X space that is related as closely possible to the output response matrix Y. PLS-DA searches for linear combinations of independent (predictor) variables, namely latent variables (LV), that maximize the covariance between the latent variable and the response. Furthermore, where the Y data measurements can be classified into different independent categories, i.e. trace gas concentrations, the algorithm is capable of setting separate and simpler models for each Y category. PLS-DA is implemented using the SIMPLS algorithm in Matlab [57].

Models were constructed using nine output categories for dataset A and eight for dataset B. Initially, model performance was evaluated using four different protocols as listed in table 2. Protocols 1–3 represent within individual or combined session evaluations while Protocol 4 uses one dataset for training and the other for testing. In each protocol, the ratio between training and validation samples is 50%–50%. With cross validation, different subsets of the data are used for training and testing and the accuracy of model prediction with unseen test data is determined. This procedure is repeated with different data subsets to provide an estimate of average prediction accuracy and the root mean square error (RMSE) [58]. The Leave One Out Cross Validation (LOO-CV) procedure uses all samples but one as the training set, the remaining sample acting as the blind test. This procedure is repeated until all samples are used as test and the mean accuracy and RMSE are returned. The PLS-DA algorithm was tested using a model set, where each individual model was constructed using 1–15 latent variables. LOO-CV approach was applied to each model to acquire an estimate of the model accuracy versus the number of LV used to build the model [58]. We investigated a number of enhancements to the PLS-DA analysis, detailed below, in order to evaluate and improve algorithm prediction accuracy.

3.1. Regularisation

Model overfitting, whereby the high accuracy obtained in model training is not replicated in the test phase can often be reduced using standard regularisation techniques whereby a penalty term is introduced to constrain some of the model regression coefficients [59–61]. Three common algorithms were investigated, namely Lasso, Ridge and Elastic Net. Lasso regularisation (L1 norm) forces the sum of the absolute value of the regression coefficients to be less than a fixed value which in turn forces some coefficients to zero, removing them from the model and its penalty term can be written as $\lambda \sum_{j=1}^{p} |\beta_j|$ where λ is the regularization parameter that determines how much the model's flexibility should be penalized and β_j is the regression coefficient. In contrast, the penalty term for Ridge regularisation (L2 norm) can be defined as $\lambda \sum_{j=1}^{p} \beta_j^2$ resulting in all coefficients being regularised equally but with a much smaller number of coefficients set to zero. Elastic net creates a linear combination of the L1 and L2 regularisation penalties by adding a quadratic, i.e. Ridge, penalty to that of Lasso with a constant α determining the relative weights and is given by $\lambda \left(\frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right)$ [62].

3.2. Data segmentation

To investigate the impact of data redundancy and the large number of data variables on model prediction, each original dataset of 3648 variables (wavelengths) was split into M subsets, each containing N variables. For a given M, N and LV, models were then built for each data subset and accuracy compared. This was an exploratory task with the objective of providing qualitative insight into how spectral characteristics may affect predictions and hence a systematic variation of M, N and LV was not carried out. It is an informal approach which compares different individual subset models unlike interval PLS which performs an exhaustive search and then adds subsets sequentially to the model.

3.3. Variable importance in projection (VIP) Selection

The relative importance of each input variable in modelling the output response can be determined from the VIP scores. These measure the contribution to the model of each predictor variable, j, by accounting for the covariance between X_i and y_i, where i is the ith latent variable, as expressed by the calculated PLS weights $(W_{i,j})^2$ in (1) [63, 64],

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_i^n S^2(y,t_i)\left(\frac{W_{ij}}{w_i}\right)^2}{\left(\frac{1}{m}\right)\sum_i^n S^2(y,t_i)}} \qquad (1)$$

where m is the total number of predictor variables, n is the total number of latent variables and $s^2(y,t_i)$ is fraction of y variance defined by latent variable i. Subsequent revised models can then be generated using a reduced set of input variables whose VIP scores are above a given threshold. A common approach assumes a threshold greater than 1, which is the average of the squared VIP scores, thereby selecting variables with an above average contribution to the model.

3.4. Peak width compression

From [35] we have found that data in regions around spectral peaks makes the most important contribution to the simplified binary classification. Also the model accuracy was found to be very sensitive, in some conditions, to peak measurement misalignment due to spectrometer jitter. This misalignment caused peaks in similar samples to appear up to a few variable units away from its nominally true value and is interpreted by the model as separate variables. To counter this, the variable values of a number of major peaks were established as references and spectra subjected to alignment shifting. However the required shift was non-linear and the existence of reliable reference peaks below 500 nm could not be guaranteed for all conditions. Therefore an alternative approach was investigated. Where the underlying optical transition is expected to be a line transition at a single wavelength subject to instrumental broadening, each measured peak spans a range of wavelengths due to the low resolution and inherent jitter of the spectrometer. Therefore in this approach the observed broad peak, over a wavelength range $\Delta\lambda$, is compressed to a single intensity value by summing the intensities over $\Delta\lambda$. This value is then assigned to a single wavelength variable. The remaining variables within $\Delta\lambda$ are then discarded from the model. This procedure is carried out on each peak of intensity greater than a set threshold (100) for each category using the Savitzky-Golay (SG) algorithm for data smoothing and peak finding. This method aims to remove correlated variables around a peak while assigning their summed intensity to the single peak variable and also avoiding the issue of misalignment.

4. Results

Multi-category PLS-DA models were constructed from He–CH_4 spectra with varying CH_4 concentrations up to 100 ppm, using the full 3668 wavelength variable set over a range of LV values for each of the protocols in table 2. Using cross-validation (CV) the model accuracy range for each LV was determined and results are given in figure 4. As expected, accuracy improves with LV and tends to saturate for LV >8. While the outcomes for protocols 1 and 2 are similar, the accuracy is lower when both datasets are merged, protocol 3. With protocol 4 the test data is obtained from a different session to that of the training data and the outcome is a poor classification accuracy across all LV values. With the addition of regularisation, the penalty factor, λ, was determined from ten-fold CV of the least absolute shrinkage and selection operator (LASSO) algorithm. The variation in mean square error (MSE) with λ is given in figure 5 with $\lambda_{\min} = 10^{-2.25}$ at the lowest MSE. However implementing either LASSO

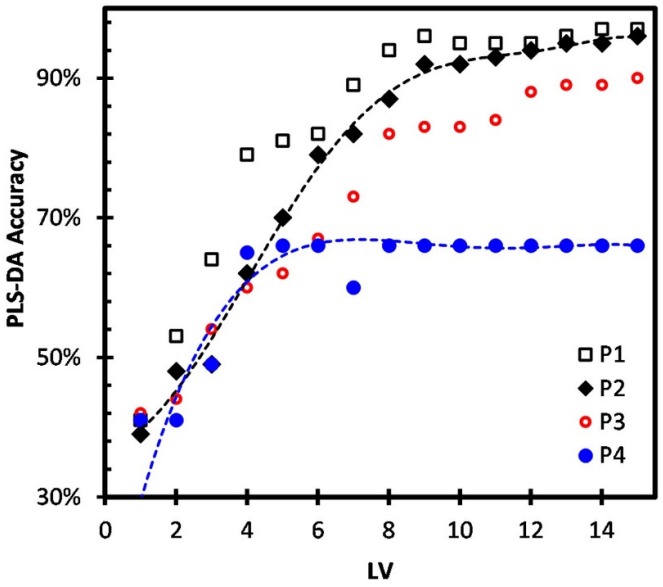

Figure 4. Comparison of PLS-DA accuracy versus the number of model latent variables for four protocols given in table 2. Error bars, representing the RMSE values at each latent variable, are of similar size to symbols for LV >5 and are excluded for clarity.

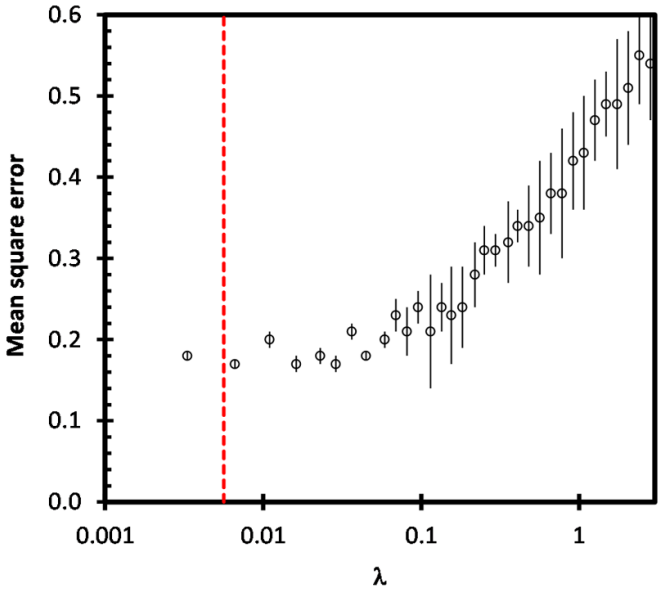

Figure 5. Mean square error (MSE) versus the LASSO penalty λ from ten-fold CV. The minimum cross-validated MSE occurs at $\lambda = 10^{-2.25}$ (red line) while the sparsest model with low MSE occurs at $\lambda = 10^{-2}$.

or Ridge regularisation with λ from $\lambda_{\min}$ to 10^{-2}, where the sparsest models are formed, resulted in limited improvement in accuracy. The LASSO regularisation identified a subset of 275 wavelength variables, from the original ∼3600, for use in the model. While this indicates a high degree of variable redundancy, over 30% of the selected variables were from the long wavelength range (>800 nm) which is featureless and highlights the likelihood of noise amplification as a by-product of the penalty term.

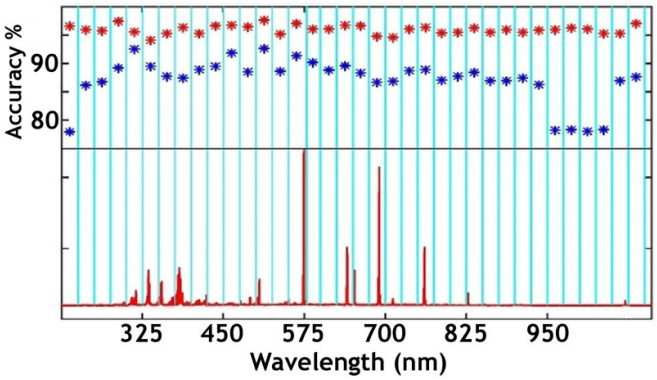

Figure 6. (Upper) Subset model accuracy for 36 subsets with ∼100 wavelength variables each. Vertical lines indicate subset boundaries. Accuracy values are given for training samples (red) and test samples (blue). (Lower) Original spectrum example.

Table 3. Wavelength ranges of the top five test model accuracies and their relation to the spectral peak height ranking.

Subset No.	Wavelength Interval (nm)	Subset Accuracy %	Peak Intensity rank
15	562.64–588.47	91.33	1
13	511.02–536.73	92.58	2
5	300.62–327.10	92.49	3
11	458.86–484.71	91.89	4
16	588.47–614.23	90.15	5

An exploration of the regions of the spectra important to model accuracy was undertaken via data segmentation into M subsets, each containing N wavelength variables. M subset models were constructed using protocol 2 and an LV value of 15. In figure 6 the accuracy is compared for each subset model for both training and test data, where M is 36 and N is ∼100. Overall the outcome shows a degree of overfitting that is most pronounced in the featureless regions at long and short wavelengths, while the lowest degree of overfitting occurs in the wavelength range 300.62 nm–327.10 nm. The best subset model accuracy was observed for subsets in the wavelength range 511.02 nm–536.73 nm with accuracy 92% similar to that achieved from full variable models (95%). The wavelength ranges of the top five test model accuracies correlate well with the highest spectral peaks, table 3.

However this level of accuracy was not maintained when train and test followed protocol 4. The accuracy was found to fall considerably for all subsets with a maximum accuracy of 60% observed, figure 7.

Full wavelength variable models were further analysed by calculating the VIP scores to determine the relative contribution of each variable. For the simplest case of pure He, the scores versus wavelength plot, figure 8, shows the highest VIP scores occur at spectral peaks. However the rank of the scores does not match the peak intensity rank, table 4 i.e. the intensities of spectral peaks are not necessarily a good indicator of value to the model. Note, the VIP wavelengths do not exactly match the original peak values as the latter were rounded to

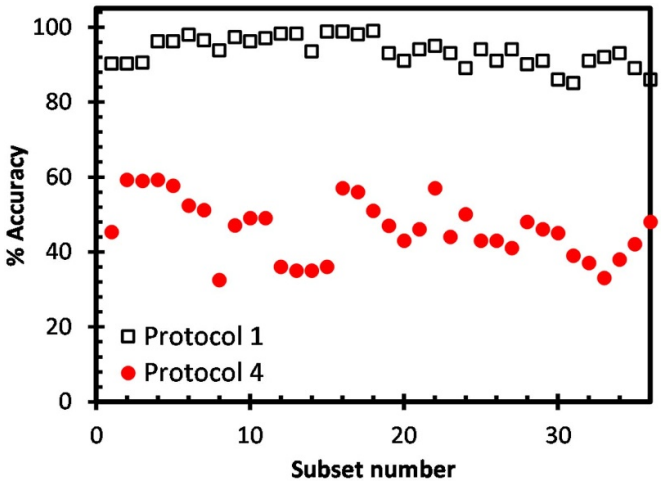

Figure 7. Subset model accuracy for 36 subsets with ∼100 wavelength variables each. Comparison between Protocol 1 and Protocol 4.

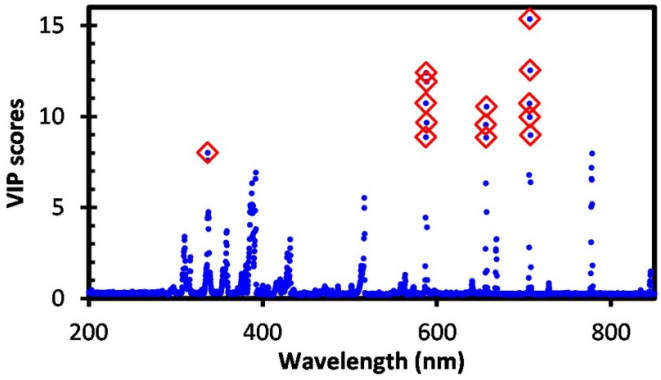

Figure 8. VIP scores for whole spectra in pure He. The red diamonds highlight 14 VIPs with score ⩾8. Each VIP corresponds to a single wavelength value.

indicate spectral variability. Selected VIPs are therefore associated with their nearest wavelength peak reported in table 1. In the case of VIP rank 6 (391.7 nm), this is associated with original peak labelled 389 nm, which is a broad peak likely reflecting a main He line and other, much smaller, impurity features (CN, N_2, O_2).

To create VIP feature selected models, an arbitrary VIP score threshold was chosen to balance the need for a manageable number of scores with the probability of including those most appropriate. For a threshold value of $N_{VIP} > 8$, 14 spectral peaks are selected in four wavelength regions, namely (a) 336.61 nm, (b) 587.18–588.21 nm, (c) 656.30–656.82 nm and (d) 706.31–707.33 nm. A set of 14 reduced feature count PLS-DA models (9 LVs, Protocol 1) was created using ±10 wavelength variables around each of the 14 VIP-selected peaks and their accuracy compared in figure 9 (lower).

With VIP feature selection, an accuracy of 88% is achievable using only 20 wavelength variables over a very restricted wavelength range. This compares to 92% accuracy achieved for the full model comprised of >3600 wavelengths. While there appears to be no relationship between VIP height and resultant model accuracy, figure 9 (inset), a trend of increasing accuracy is apparent for sets containing the higher VIP scores, figure 9 (lower). To further reduce the number of features, we selected the peaks corresponding to the highest scores from each of four sets and built PLS-DA models (2, 3 or 4 peaks, ±10 wavelength variables per peak with 9 LVs). This resulted in an increase in accuracy to 99%, figure 9 (upper). A further set of models were created for a number of VIP selected peaks (9 LVs, ±10 wavelength variables) with the training and testing carried out via Protocol 4. As occurred with data segmentation subset models, the accuracy fell significantly, with a maximum of 45%, figure 10.

While regularisation, data segmentation and VIP-related models reduced the wavelength variable count considerably (to 275, 100, and 20 respectively), only a limited number of variables are removed when using a peak width compression approach. After SG smoothing, all spectral peaks above an arbitrary threshold height (100) are selected for compression, figure 11. Overall, the total number of wavelength variables is reduced from the original 3668 depending on the extent of peak broadening e.g. for 6 ppm 3328 variables remain and 3300 for 100 ppm.

The outcome of peak compression showed slight improvement in accuracy over results for protocols 1 and 2, figure 4, but for protocol 4 a significant increase in accuracy was observed, figure 12, reaching ⩾97% for 8 LVs.

5. Discussion

Using PLS-DA classification applied to low resolution UV—visible range optical emission spectra derived from plasma excitation, we have demonstrated the ability to detect the presence of methane down to concentrations of 1 ppm and to label sample concentrations up to 100 ppm. Simple application of PLS-DA in protocols 1 and 2, with limited pre-processing, shows the capability of this algorithmic approach in developing accurate multi-categorical models based on high dimensionality spectral data. As expected, the accuracy increases with increasing number of latent variables used, levelling off in accuracy (>90%) for LV ⩾ 10. However the potential for overfitting of spectral data is obvious from protocol 3 where the mixed session data shows a fall in accuracy, for a given LV, compared to protocols 1 and 2. In protocol 4, the training and test data were from entirely different sessions presenting a much more realistic and stringent challenge which the algorithm failed to handle satisfactorily. These relatively simple plasma devices along with portable spectrometers inevitably produce highly variable output. Within a single session, relative standard deviation (RSD) values for pure He can be >10%, which increases to ∼30% with added CH_4. Machine learning algorithms and associated pre-processing or enhancements offer the opportunity to directly negate the effect of this variability on predictive accuracy. They also offer the opportunity to gain further insight into underlying mechanisms to

Table 4. Comparison of VIP scores rank with peak intensity rank for pure He spectra.

VIP Rank	VIP wavelength	VIP score	Peak Intensity Rank	Species
1	706.8	15.36	2	He
2	587.7	12.41	1	He
3	656.8	10.55	8	H
4	336.6	8.00	6	Impurity
5	778.2	7.97	4	Impurity
6	391.7	6.92	5	He, CN, N_2, O_2
7	516.5	5.54	>9	C2 Swan
8	309.9	3.41	>9	OH
9	431.4	3.25	>9	CH
10	668.1	3.24	3	Impurity, He

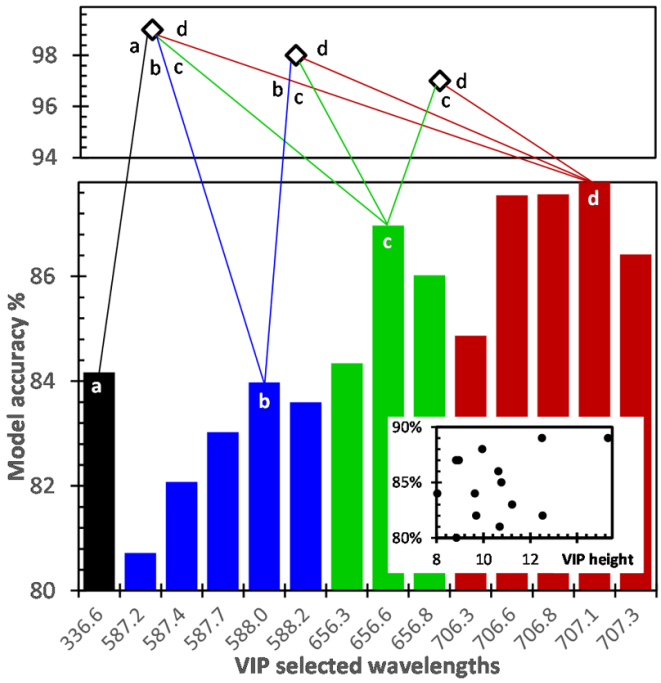

Figure 9. *Lower*: Accuracy of 14 reduced feature count PLS-DA models, each based on ± 10 wavelength variables around single spectral peak centred at wavelengths where VIP scores >8. Models were trained and tested according to Protocol 1 with 9 LVs. *Upper*: Accuracy of three reduced feature count PLS-DA models, each based on ± 10 wavelength variables around 2, 3 or 4 spectral peaks centred at wavelengths where VIP scores >8. Models were trained and tested according to Protocol 1 with 9 LVs. The two-peak model (peaks c, d) uses VIP determined peaks at 707.07 nm and 656.56 nm, while three and four peak models use peaks (b, c and d) and (a, b, c and d) respectively. *Inset:* Model accuracy versus VIP height.

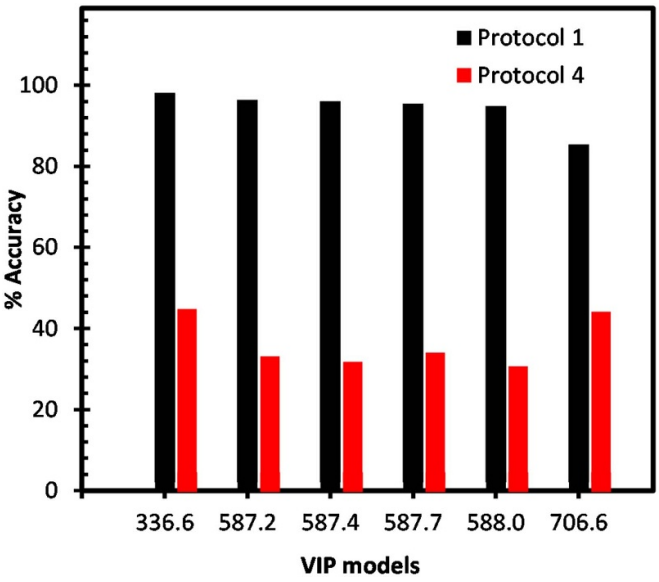

Figure 10. Accuracy of 6 reduced feature count PLS-DA models, each based on ± 10 wavelength variables around single spectral peaks centred at wavelengths (336.6 nm, 587.2 nm, 587.4 nm, 587.7 nm, 588.0 nm, 706.6 nm) where VIP scores >8. Models were trained and tested according to Protocol 4 with 9 LVs and compared with Protocol 1.

help improve issues such a plasma hardware, operating procedures and auto-filtering of data to allow progress to more complex sensing scenarios.

The data challenges faced in this work highlight one of the main practical difficulties with high dimensionality data. With a relatively small number of samples, models overfit to the training data and have reduced generality. The standard regularization approaches used here were unable to overcome the overfitting issue directly and appeared to penalize the data to the extent that featureless regions of the spectrum became primary predictors in the model. Our data segmentation approached showed that by reducing the number of predictor (wavelength) variables from ~3600–100 resulted in limited loss in model accuracy for protocols 1 and 2. This was observed for a number of subset models across the wavelength range 300 nm to 600 nm. However, even with such a reduced variable number, overfitting is still a significant factor and application of the approach using protocol 4 was unsuccessful. Nevertheless, reduction of variable count has been shown to be important not only for data analysis but may also allow use of lower specification and narrower range spectrometers, with implications for reduced cost. There is considerable scope in exploring the data segmentation approach further using multiple segment models and variable window sizes. Using VIP scores to reduce the variable count further, with models containing only 20 variables at single peaks, also resulted in high prediction accuracy for protocol 1 but much lower predictive success for protocol 4. This again indicates the prevalence of

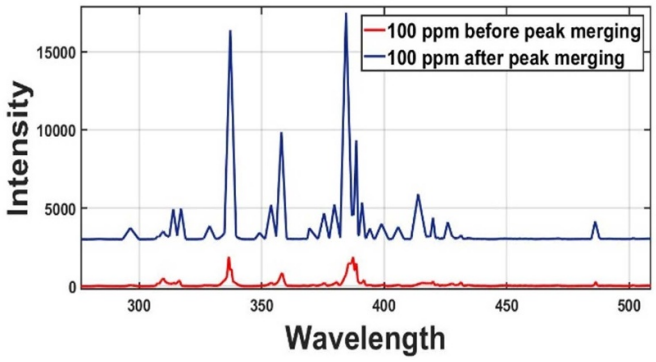

Figure 11. Spectrum of He–CH_4 (100 ppm) sample before (red) and after (blue) peak compression in the wavelength interval 270 nm–500 nm.

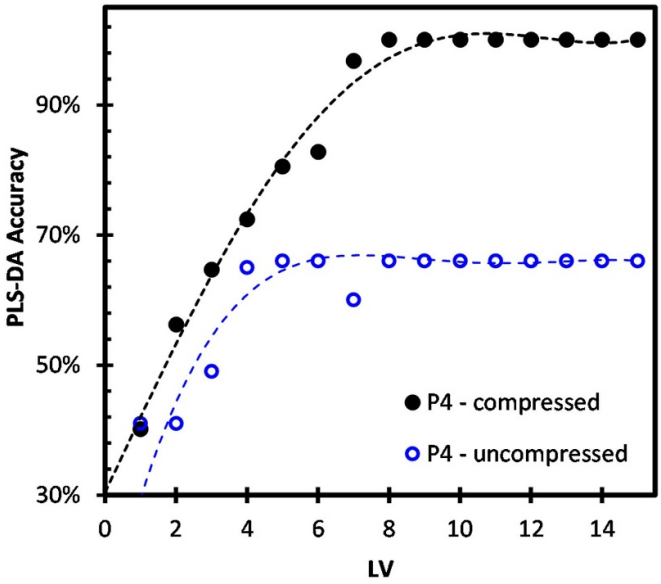

Figure 12. PLS-DA accuracy versus LV using peak compression and trained/tested under Protocol 4, in comparison with accuracy obtained from uncompressed peaks.

overfitting even though the number of wavelength variables is small compared to the number of samples. In contrast, the peak compression procedure has provided the greatest predictive success with regard to protocol 4 with accuracy values >97%, an outcome as good as that obtained from Protocol 1. By reducing the multiple correlated wavelength variables around each spectral peak to a single variable, the effect of overfitting has been minimized. Also, since the compressed peak wavelengths are the same for each sample, the issue of spectrum misalignment is no longer a concern. Including segmentation and/or VIP selection with peak compression offers routes for consideration with more complex gas mixtures.

VIP score calculation is a technique by which the PLS-DA can report the significance of each individual variable (wavelength) to the model predictions and as such provides direct physical insight into the primary plasma factors underlying the model. By definition, the average VIP score is one and is often used as a significance threshold. However, in order to restrict the wavelength priority list to a manageable number we used a high threshold value (VIP > 8), producing a list of 14 wavelengths. Analysis of the full variable (>3600) count models using VIP indicated, as expected, that the primary contributors to models were located at the spectral peaks. However, the spectral peak height ranking did not necessarily follow the ranking of the VIP scores. Within this list of top ten VIP scores, the top two represent the highest peaks from He emission, while of the remainder, five can be attributed to hydrogen, nitrogen or oxygen related impurities, with relatively small peaks, and two to carbon-based species.

The nature and impact of chemical species on the algorithm prediction is important on two counts. If the algorithm were to be dependent solely on He peaks, then the applicability of the technique with other plasma gases, e.g. Ar, N_2 or air, is unclear. A dependence on hydrocarbon impurity peaks may imply a high degree of CH_4 dissociation which could hamper attempts to differentiate different hydrocarbons. With the discovery of non-CH_4 impurity emission as a significant factor in the VIP score list, it is possible however that prediction depends on CH_4—induced changes to the overall plasma. As is observed with molecular gases in general, we would expect collisions with low-energy electrons to result in vibrational excitation of the molecule while the absorption of these electrons may lead to change in the EEDF sufficient to affect the emission of all species. The development of suitable plasma chemistry models is severely hampered by the limited rate coefficient and cross-section data for many of the possible reactions and the lack of experimental plasma parameter values. Nevertheless, it is worthwhile assessing the potential significance of both CH_4 dissociation and non-CH_4 impurity impact on EEDF.

Given the likely complexity of the plasma chemistry along with the limited spectrometer resolution, multiple species assignment to a single emission line is possible and while knowledge of the underlying chemistry would be valuable, it is not currently available. In Vincent *et al*, we discuss the possible He–CH_4 chemistry at trace methane levels and its impact on emission spectra [35]. Molecular CH_4 has no emission lines in the wavelength range 200 nm–1100 nm. However we observe small features around 431 nm and 389 nm which can be attributed to CH emissions from the $A^2\Delta \rightarrow X^2\pi$ system [65] and the (0,0) band of the $B^2\Sigma^- \rightarrow X^2\pi$ system, although the 389 nm peak also represents emission from the He transition (1s2s–1s3p) [48]. These features are weak and the variance is relatively large, nevertheless there is a trend of increasing intensity with CH_4 concentration. The Hα line at 656.28 nm is present in all spectra but becomes the dominant peak at CH_4 concentrations above ~40 ppm. At low CH_4 concentrations the dissociation of H_2O may be the primary source of Hα; observed peaks around 310 nm are likely due to OH(A–X) emission [33]. From high resolution humidity measurements of the plasma source gas, we estimate H_2O content between 10 ppm and 500 ppm in our pure He plasmas. Emission due to C_2 Swan vibrational bands around 516 nm appears at concentrations above 77 ppm. These correspond to transitions between the $d^3\pi_g$ (2.48 eV) and $a^3\pi_u$ (0.09 eV) electronic states and indicate the final hydrogen abstraction

endpoint from CH_4. CH emission normally dominates over C_2 emission in hot methane flames or plasmas [66] since the latter derives from C_2H_y species which in turn are produced by heavy particle collisions between methane radicals, e.g. dimerisation reactions between CH and CH_x [66]. These reactions are often exponentially dependent on gas temperature, with thresholds typically >1000 °C [65]. Therefore CH_4 dissociation and emission from C_xH_y fragments (x: $0 \rightarrow 2$, y: $0 \rightarrow 4$) can be expected to make some contribution to PLS-DA models.

The presence of H_2O, N_2 and O_2 impurities and their associated radicals also lead to additional emission features. For example, a persistent peak around 336 nm can be attributed to N_2 rotational and vibrational molecular bands. To estimate the effect of CH_4 or CH_4 plus some dissociation fragments on pure He emission with or without molecular impurities, we calculated the EEDFs of different mixtures using a Boltzmann solver [67] along with the available cross-sections for CH_4, some related CH_4 dissociation reactions [68] as well as those for H_2O, N_2 and O_2 [69]. Calculated rate equations for CH_4 elastic collisions are similar to those of helium and given the trace level CH_4 concentrations, the calculated electron energy loss due to He elastic collisions remains unchanged on the introduction of trace gases. We observed no change in absorbed power as CH_4 is added to the helium plasma. In figure 13(a), the impact of CH_4 (1 ppm) and impurities (H_2O 500 ppm, O_2 10 ppm, N_2 10 ppm, and H 10 ppm) on the pure He EEDF can be observed. The addition of CH_4 tends to increase the high energy tail of the EEDF with this effect decreasing as the concentration reaches 100 ppm,

Figure 13(b). However, atmospheric impurity species have almost the opposite effect of decreasing the higher energy regions of the EEDF. Nevertheless, the addition of CH_4 (1 ppm) to He with impurities included tends to negate this effect, figure 13(c), and the high energy tail increases. The impact of CH_x dissociation species on EEDF also shows a complex relationship with concentration. While EEDFs for He–CH_4 (100 ppm) and He–CH_4 (0 ppm) are almost indistinguishable, the presence of 10 ppm CH_X in He–CH_4 (100 ppm) leads to a significant decrease in EEDF between mean electron energies of 4–13 eV before rising again at higher energies, figure 13(d). For example, the impact of this change in EEDF on the He emission (1s3d→1s2p, 587.56 nm) is illustrated by the variation in calculated rate coefficients for the He 1s3d excitation, figure 13(e). The rate coefficient is very sensitive to the mean electron energy (ε), falling sharply around 2 eV before increasing up to 10 eV. Using an expected value of $\varepsilon = 2$ eV, the calculated reduction in rate coefficient is ~32%. Experimentally the observed reduction in the He emission (1s3d→1s2p, 587.56 nm) peak is 35%–40% with the addition of 100 ppm CH_4. The analysis of EEDF variation due to the inclusion of trace impurities and hydrocarbons has limited direct predictive capability at this stage due to a lack of information on plasma chemistry at atmospheric pressure and cross-section details for a large number of potential reactions. Nevertheless, even with small quantities of impurities or hydrocarbons, the change in EEDF is noticeable. This is

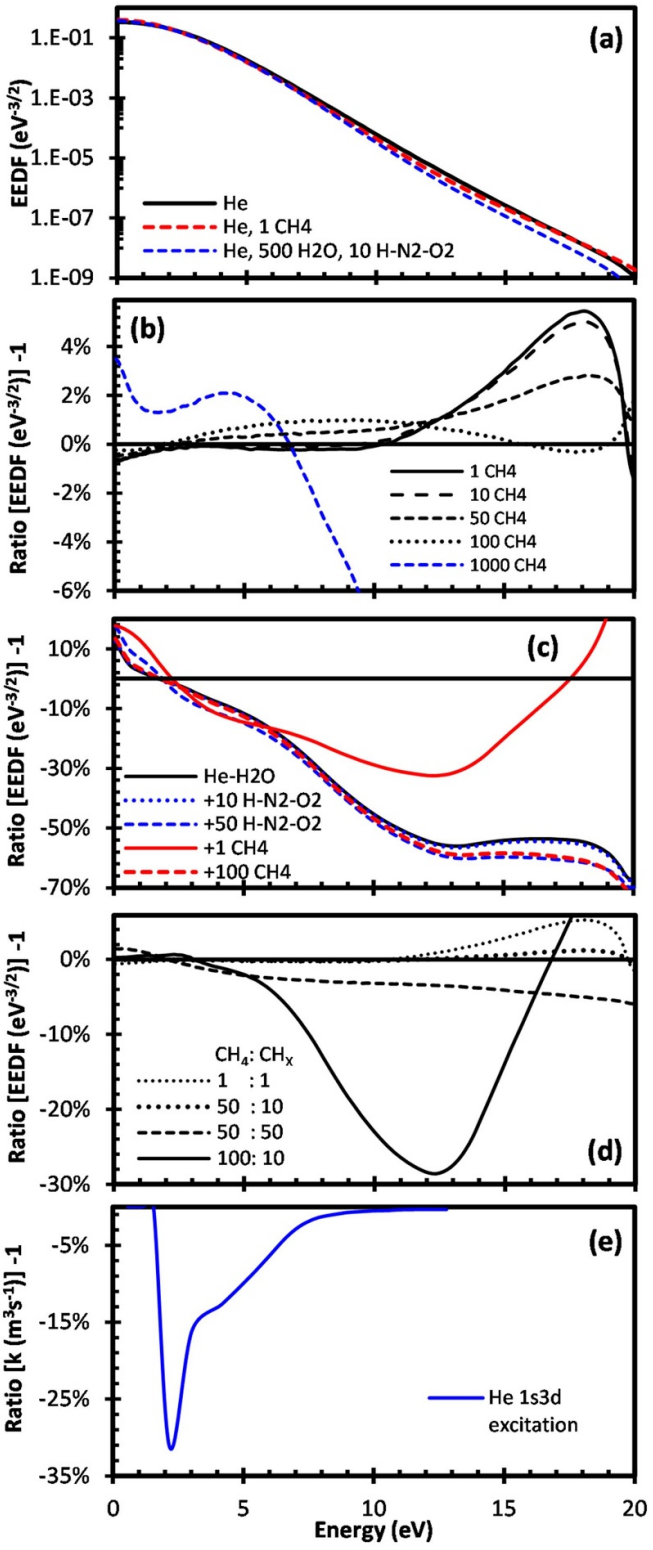

Figure 13. (a) EEDF plots for He, He + CH_4 (1 ppm), He + H_2O (500 ppm) + N_2/O_2 impurities (10 ppm), (b) ratio of EEDF vs energy with addition of CH_4 to the EEDF of He (0 ppm), (c) change in impurity to helium EEDF ratio with the addition H_2O (500 ppm) to He (black), with addition of H, N2 and O2 impurities to He–H_2O (500 ppm) (blue), and with the addition of CH_4 to He–H_2O/H/N_2/O_2 (red), (d) change in EEDF ratio with the addition of CH_4 and CH_x (x: 0→3) to He, (e) variation in the ratio of rate coefficients for the He 1s3d excitation in He and in He + CH_4/CH_x (100 ppm/10 ppm).

expected since a high rate of vibrational/rotational excitation occurs in molecular gases as well as dissociation, at energies well below those of pure noble gases. Overall, with the objective being trace gas detection, we observe spectral changes due to additional impurity and hydrocarbon peaks as well as changes to the primary He peaks, due to impurity and hydrocarbon induced EEDF modification. According to the priority VIP list, the latter is a significant factor in the algorithm operation.

6. Conclusion

We have demonstrated the capability of using OES from a small-volume (5 μl) atmospheric pressure plasma, coupled with PLS-DA spectral classification algorithms, to detect the presence of methane down to concentrations of 1 ppm and to label sample concentrations up to 100 ppm. This compares well with portable NDIR systems [10], which deliver LOD values above 50 ppm, and low cost chemi-resistive sensors which represent the most commonly deployed technology [70]. The ability to detect CH_4 and assign a concentration classification offers scope for higher resolution classifications which will be valuable for diagnostics, online monitoring of trends and developing advanced warning capabilities. Nevertheless, future plasma emission sensor devices will also need to handle increased levels of matrix gases including air and other hydrocarbons and will require further development of algorithms and plasma sources. We have investigated a number of algorithm enhancements including regularization, simple data segmentation and subset selection, VIP feature selection and wavelength variable compression. All these approaches showed the potential for significant reduction in the number of wavelength variables and the spectral resolution/bandwidth—an important technological consideration. However only wavelength variable compression exhibited reliable predictive performance under the more challenging multi-session train—test scenarios. Nevertheless, there is still considerable scope for fine tuning the application of single and multiple enhancements. Gaining some understanding of plasma—gas interactions, their appearance in spectra and their interpretation by classification algorithms is important for algorithm enhancement when faced with a wide array of options. Although knowledge of radical species densities and their cross-sections is very limited, modelling the impact of chemistry on plasma conditions has illustrated the complex cross-sensitivities in the excitation of noble gas, impurities, target CH_4 and its dissociation fractions. The discovery that trace impurity species variation, other than in the target gas, is a significant factor in algorithm prediction indicates that successful operation can be possible independent of the choice of plasma gas.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Tahereh Shah Mansouri https://orcid.org/0000-0003-4710-0546
Davide Mariotti https://orcid.org/0000-0003-1504-4383
Paul Maguire https://orcid.org/0000-0002-2725-4647

References

[1] Kwak D, Lei Y and Maric R 2019 Ammonia gas sensors: a comprehensive review *Talanta* **204** 713–30
[2] Marzorati D, Mainardi L, Sedda G, Gasparri R, Spaggiari L and Cerveri P 2019 A review of exhaled breath: a key role in lung cancer diagnosis *J. Breath Res.* **13** 034001
[3] Casas-Ferreira A M and Nogal-Sánchez M 2019 Non-separative mass spectrometry methods for non-invasive medical diagnostics based on volatile organic compounds: a review *Anal. Chim. Acta* **1045** 10–22
[4] Zoccali M, Tranchida P Q and Mondello L 2019 Fast gas chromatography-mass spectrometry: a review of the last decade *TRAC Trends Anal. Chem.* **118** 444–52
[5] Bulska E and Ruszczyńska A 2017 Analytical techniques for trace element determination *Phys. Sci. Rev.* **2** 20178002
[6] Hübert T, Boon-Brett L, Palmisano V and Bader M A 2014 Developments in gas sensor technology for hydrogen safety *Int. J. Hydrog. Energy* **39** 20474–83
[7] Fox T A, Barchyn T E, Risk D, Ravikumar A P and Hugenholtz C H 2019 A review of close-range and screening technologies for mitigating fugitive methane emissions in upstream oil and gas *Environ. Res. Lett.* **14** 53002
[8] Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt K.B, Tignor M and Miller H.L Overview of Greenhouse Gases (available at: www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks) (Accessed 12 August 2021)
[9] Alvarez R A, Pacala S W, Winebrake J J, Chameides W L and Hamburg S P 2012 Greater focus needed on methane leakage from natural gas infrastructure *Proc. Natl Acad. Sci.* **109** 6435–40
[10] Kamieniak J, Randviir E P and Banks C E 2015 The latest developments in the analytical sensing of methane *TRAC Trends Anal. Chem.* **73** 146–57
[11] King J and Schiff E ARPA-MONITOR Methane observation networks (available at: https://arpa-e.energy.gov/technologies/programs/monitor) (Accessed 12 August 2021)
[12] Dong L, Tittel F K, Li C, Sanchez N P, Wu H, Zheng C, Yu Y, Sampaolo A and Griffin R J 2016 Compact TDLAS based sensor design using interband cascade lasers for mid-IR trace gas sensing *Opt. Express* **24** A528–35
[13] Moon H G *et al* 2016 Chemiresistive electronic nose toward detection of biomarkers in exhaled breath *ACS Appl. Mater. Interfaces* **8** 20969–76
[14] Casey J G, Collier-Oxandale A and Hannigan M 2019 Performance of artificial neural networks and linear models to quantify 4 trace gas species in an oil and gas production region with low-cost sensors *Sens. Actuators* B **283** 504–14
[15] Chen Y, Owyeung R E and Sonkusale S R 2018 Combined optical and electronic paper-nose for detection of volatile gases *Anal. Chim. Acta* **1034** 128–36
[16] Adamovich I *et al* 2017 The 2017 plasma roadmap: low temperature plasma science and technology *J. Phys. D: Appl. Phys.* **50** 323001
[17] Weltmann K *et al* 2019 The future for plasma science and technology *Plasma Process. Polym.* **16** 1800118

[18] Chiang W, Mariotti D, Sankaran R M, Eden J G and Ostrikov K 2020 Microplasmas for advanced materials and devices *Adv. Mater.* **32** 1905508

[19] Hyland M, Mariotti D, Dubitzky W, McLaughlin J A and Maguire P 2000 *Gas recognition using a neural network approach to plasma optical emission spectroscopy* vol 4120 (Bellingham, WA: International Society for Optics and Photonics) p 246–52

[20] Weagant S, Dulai G, Li L and Karanassios V 2015 Characterization of rapidly-prototyped, battery-operated, argon-hydrogen microplasma on a hybrid chip for elemental analysis of microsamples by portable optical emission spectrometry *Spectrochim. Acta* B **106** 75–80

[21] Zheng L and Kulkarni P 2017 Rapid elemental analysis of aerosols using atmospheric glow discharge optical emission spectroscopy *Anal. Chem.* **89** 6551–8

[22] Cserfalvi T, Mezei P and Apai P 1993 Emission studies on a glow discharge in atmospheric pressure air using water as a cathode *J. Phys. D: Appl. Phys.* **26** 2184–8

[23] He Q, Zhu Z and Hu S 2014 Flowing and nonflowing liquid electrode discharge microplasma for metal ion detection by optical emission spectrometry *Appl. Spectrosc. Rev.* **49** 249–69

[24] Doroski T A, King A M, Fritz M P and Webb M R 2013 Solution–cathode glow discharge—optical emission spectrometry of a new design and using a compact spectrograph *J. Anal. At. Spectrom.* **28** 1090–5

[25] Peng X, Guo X, Ge F and Wang Z 2019 Battery-operated portable high-throughput solution cathode glow discharge optical emission spectrometry for environmental metal detection *J. Anal. At. Spectrom.* **34** 394–400

[26] Pohl P, Jamroz P, Greda K, Gorska M, Dzimitrowicz A, Welna M and Szymczycha-Madeja A 2021 Five years of innovations in development of glow discharges generated in contact with liquids for spectrochemical elemental analysis by optical emission spectrometry *Anal. Chim. Acta* **1169** 338399

[27] Wang J, He M, Zheng P, Chen Y and Mao X 2019 Comparison of the plasma temperature and electron number density of the pulsed electrolyte cathode atmospheric pressure discharge and the direct current solution cathode glow discharge *Anal. Lett.* **52** 697–712

[28] Bogaerts A 2020 Modeling plasmas in analytical chemistry-an example of cross-fertilization *Anal. Bioanal. Chem.* **412** 6059–83

[29] Decker C G and Webb M R 2015 Measurement of sample and plasma properties in solution-cathode glow discharge and effects of organic additives on these properties *J. Anal. At. Spectrom.* **31** 311–8

[30] Lu Y, Xu S F, Zhong X X, Ostrikov K, Cvelbar U and Mariotti D 2013 Characterization of a DC-driven microplasma between a capillary tube and water surface *Europhys. Lett.* **102** 15002

[31] Hofmann S, Van Gessel A F H, Verreycken T and Bruggeman P J 2011 Power dissipation, gas temperatures and electron densities of cold atmospheric pressure helium and argon RF plasma jets *Plasma Sources Sci. Technol.* **20** 065010–1/12

[32] Askari S, Levchenko I, Ostrikov K, Maguire P and Mariotti D 2014 Crystalline Si nanoparticles below crystallization threshold: effects of collisional heating in non-thermal atmospheric-pressure microplasmas *Appl. Phys. Lett.* **104** 163103

[33] Maguire P D *et al* 2015 Controlled microdroplet transport in an atmospheric pressure microplasma *Appl. Phys. Lett.* **106** 224101

[34] Kudryavtsev A A, Stefanova M S and Pramatarov P M 2015 Use of nonlocal helium microplasma for gas impurities detection by the collisional electron spectroscopy method *Phys. Plasmas* **22** 103513

[35] Vincent J, Wang H, Nibouche O and Maguire P 2020 Detecting trace methane levels with plasma optical emission spectroscopy and supervised machine learning *Plasma Sources Sci. Technol.* **29** 85018

[36] Wirsz D F and Blades M W 1986 Application of pattern recognition and factor analysis to inductively coupled plasma optical emission spectra *Anal. Chem.* **58** 51–57

[37] Song W, Wang H, Maguire P and Nibouche O 2018 Nearest clusters based partial least squares discriminant analysis for the classification of spectral data *Anal. Chim. Acta* **1009** 27–38

[38] Song W, Wang H, Maguire P and Nibouche O 2017 Local partial least square classifier in high dimensionality classification *Neurocomputing* **234** 126–36

[39] Song W, Wang H, Maguire P and Nibouche O 2018 Collaborative representation based classifier with partial least squares regression for the classification of spectral data *Chemometr. Intell. Lab. Syst.* **182** 79–86

[40] Vincent J, Wang H, Nibouche O and Maguire P 2018 Differentiation of apple varieties and investigation of organic status using portable visible range reflectance spectroscopy *Sensors* **18** 1708

[41] Song W, Wang H, Maguire P and In N O 2017 In Differentiation of organic and non-organic apples using near infrared reflectance spectroscopy—A pattern recognition approach *Proc. IEEE Sens.* **229** 754–60

[42] Gidon D, Pei X, Bonzanini A D, Graves D B and Mesbah A 2019 Machine learning for real-time diagnostics of cold atmospheric plasma sources *IEEE Trans. Radiat. Plasma Med. Sci.* **3** 597–605

[43] Mesbah A and Graves D B 2019 Machine learning for modeling, diagnostics, and control of non-equilibrium plasmas *J. Phys. D: Appl. Phys.* **52** 30LT02

[44] Shojaei K and Mangolini L 2021 Application of machine learning for the estimation of electron energy distribution from optical emission spectra *J. Appl. Phys.* **54** 265202

[45] Wang C, Ko T and Hsu C 2021 Interpreting convolutional neural network for real-time volatile organic compounds detection and classification using optical emission spectroscopy of plasma *Anal. Chim. Acta* **1179** 338822

[46] Hendawy N, McQuaid H, Mariotti D and Maguire P 2020 Continuous gas temperature measurement of cold plasma jets containing microdroplets, using a focussed spot IR sensor *Plasma Sources Sci. Technol.* **29** 085010

[47] Pearse R W B and Gaydon A G 1963 *The Identification of Molecular Spectra, by R.W.B. Pearse and A.G. Gaydon* (London: Chapman & Hall) (https://doi.org/10.1159/000144514)

[48] NIST ASD Team 2021 Atomic spectra database (available at: www.nist.gov/pml/atomic-spectra-database) (Accessed 12 August 2021)

[49] Tharwat A 2016 Linear vs. quadratic discriminant analysis classifier: a tutorial *Int. J. Appl. Pattern Recognit.* **3** 145–80

[50] Hechenbichler K and Schliep K 2004 Weighted k-nearest-neighbor techniques and ordinal classification (https://doi.org/10.5282/ubm/epub.1769)

[51] Sun W and Li S 2007 *PCA-SVM-Based Comprehensive Evaluation for Customer Relationship Management System of Power Supply Enterprise* (IEEE) vol 7 pp 3811–4

[52] Jing C and Hou J 2015 SVM and PCA based fault classification approaches for complicated industrial process *Neurocomputing* **167** 636–42

[53] O'brien R 2007 A caution regarding rules of thumb for variance inflation factors *Qual. Quant.* **41** 673–90

[54] Marquaridt D W 1970 Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation *Technometrics* **12** 591–612

[55] Mason R L, Gunst R F and Hess J L 1989 *Statistical Design and Analysis of Experiments (With Applications to Engineering and Science)* (New York: Wiley) p xvi + 692 1991 *Applied Stochastic Models and Data Analysis* vol *7* p 291
[56] Kennedy P 1992 *A Guide to Econometrics* (Oxford: Blackwell)
[57] de Jong S 1993 SIMPLS: an alternative approach to partial least squares regression *Chemometr. Intell. Lab. Syst.* **18** 251–63
[58] Gowen A A, Downey G, Esquerre C and O'Donnell C P 2011 Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients *J. Chemom.* **25** 375–81
[59] Gromski P S, Muhamadali H, Ellis D I, Xu Y, Correa E, Turner M L and Goodacre R 2015 A tutorial review: metabolomics and partial least squares-discriminant analysis—a marriage of convenience or a shotgun wedding *Anal. Chim. Acta* **879** 10–23
[60] Westerhuis J A, Hoefsloot H C J, Smit S, Vis D J, Smilde A K, Van Velzen E J J, Van Duijnhoven J P M and van Dorsten F A 2008 Assessment of PLSDA cross validation *Metabolomics* **4** 81–89
[61] Brereton R G and Lloyd G R 2014 Partial least squares discriminant analysis: taking the magic away *J. Chemom.* **28** 213–25
[62] Kalivas J H 2012 Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance *J. Chemom.* **26** 218–30
[63] Mehmood T, Liland K H, Snipen L and Sæbø S 2012 A review of variable selection methods in partial least squares regression *Chemometr. Intell. Lab. Syst.* **118** 62–69
[64] Wold S, Sjöström M and Eriksson L 2001 PLS-regression: a basic tool of chemometrics *Chemometr. Intell. Lab. Syst.* **58** 109–30
[65] Danko M, Orszagh J, Ďurian M, Kočišek J, Daxner M, Zöttl S, Maljković J B, Fedor J, Scheier P, Denifl S and Matejčík Š 2013 Electron impact excitation of methane: determination of appearance energies for dissociation products *J. Phys. B: At. Mol. Opt. Phys.* **46** 045203
[66] Fantz U and Meir S 2005 Correlation of the intensity ratio of C_2/CH molecular bands with the flux ratio of C_2Hy/CH_4 particles *J. Nucl. Mater.* **337–339** 1087–91
[67] Hagelaar G J M and Pitchford L C 2005 Solving the Boltzmann equation to obtain electron transport coefficients and rate coefficients for fluid models *Plasma Sources Sci. Technol.* **14** 722–33
[68] Morgan W L Morgan database (available at: www.lxcat.net) (Accessed 01 August 2021)
[69] Kochetov I V Triniti database (available at: www.lxcat.net) (Accessed 01 August 2021)
[70] Hong T, Culp J T, Kim K, Devkota J, Sun C and Ohodnicki P R 2020 State-of-the-art of methane sensing materials: a review and perspectives *TRAC Trends Anal. Chem.* **125** 115820