Plasma Sources Science and Technology

Detecting trace methane levels with plasma optical emission spectroscopy and supervised machine learning

To cite this article before publication: Jordan Vincent *et al* 2020 *Plasma Sources Sci. Technol.* in press https://doi.org/10.1088/1361-6595/aba488

Detecting trace methane levels with plasma optical emission spectroscopy and supervised machine learning.

Jordan Vincent*, Hui Wang, Omar Nibouche, Paul Maguire

Faculty of Computing, Engineering, and the Built Environment, University of Ulster, Shore Rd, Newtownabbey, BT370QB, Northern Ireland, UK.

* Corresponding author: Vincent-j1@ulster.ac.uk

KEYWORDS cold atmospheric plasma, partial least squares, machine learning, methane detection.

ABSTRACT: Trace methane detection in the parts per million range is reported using a novel detection scheme based on optical emission spectra from low temperature atmospheric pressure microplasmas. These bright low-cost plasma sources were operated under non-equilibrium conditions, producing spectra with a complex and variable sensitivity to trace levels of added gases. A data-driven machine learning approach based on Partial Least Squares Discriminant Analysis (PLS-DA) was implemented for CH_4 concentrations up to 100 ppm in He, to provide binary classification of samples above or below a threshold of 2 ppm. With a low-resolution spectrometer and a custom spectral alignment procedure, a prediction accuracy of 98% was achieved, demonstrating the power of machine learning with otherwise prohibitively complex spectral analysis. This work establishes proof of principle for low cost and high-resolution trace gas detection with the potential for field deployment and autonomous remote monitoring.

1 Introduction

The accurate detection of trace gases and volatiles has important implications for industrial, environmental, atmospheric, and clinical applications. 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. Field deployment of high-resolution detectors and remote monitoring is a major priority yet remains elusive due to the very high system cost. Methane represents an important target gas due to health and safety aspects in landfill regions, water treatment, or in oil and gas industries. Natural gas leakage represents significant greenhouse gas emission source [1, 2] and will become subject to more stringent regulation. Methane and other gases or volatiles are also considered as potential breath biomarkers for disease prediction [3]. Ongoing research in molecular gas detection is focussed on a number of approaches including miniaturisation of tunable laser spectrometers, IR imagers and mass spectrometers, while on-chip optical sensors, microstructured optical fibres, or printable electronic sensors also offer opportunities [4, 5].

Trace element detection using plasma-based atomic emission spectroscopy techniques such as ICP-AES or LIBS, has a long history. Plasma-induced high temperatures can generate characteristic emission lines from a wide range of elements and very low LOD can be achieved. Other low gas temperature plasma sources have been used in chemical analysis, particularly RF glow discharge spectrometry for analysis of solids [6]. Microplasma-based liquid electrode vaporisation spectrometry [7] has been used to detect metals with LOD values down to 10s of μg L^{-1} and minimum LODs around 1 μg L^{-1} with sample preconcentration [8]. While there are many possible microplasma configurations receiving attention [9], detection sensitivity in general is much inferior to ICP-AES and while cooler than the latter, vaporisation

temperatures around 2000 K can still be expected. However, for trace gas detection of species other than atomic gases or high emission metals, microplasma emission spectra are very complex, individual lines are weak and poorly resolved. With the recent advent of low gas temperature, i.e. cold, atmospheric pressure (CAP) microplasmas, miniaturisation of plasma optical emission systems for field detection of trace gases becomes a possibility. Using design constraints and operating parameters that maintain low gas temperatures (< 50 ^{o}C), e.g. for breath analysis or managing safety concerns with flammable gases, adds further noise and complexity to spectra. Therefore, microplasmas have to date been primarily restricted to liquid analysis of trace metals at relatively high temperature. Nevertheless, the low cost, compact size and portability of cold microplasmas along with the possibility of unattended operation would be attractive for field monitoring of trace gases, in a wide range of applications, provided that appropriate complex spectral analysis methods can be developed.

In this work we use a data driven machine learning approach to detect the presence of trace methane at low ppm levels using a custom RF-excited CAP microplasma and a low resolution, hence low cost, broadband (UV-Vis-NIR) spectrometer. Our previous algorithm work has focussed on investigating complex reflectance spectra of samples in the visible and NIR with variants of the common Partial Least Squares Discriminant Analysis algorithm [10, 11, 26, 27]. Using plasma emission leads to higher spectral intensity overall. Nevertheless, important molecular gases, including hydrocarbons such as methane, generally have multiple but weak lines in the UV-Vis-NIR region which often overlap with plasma carrier gases such as helium or argon. 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 this study, we chose CH_4 in helium to represent the challenge of molecular detection. Impurity gas concentrations were minimised as far as possible but not eliminated.

Apart from analysis of individual spectral lines, the use of CAP microplasma emission with full spectra analysis has not previously been attempted. 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. [28, 29] Analysis of the full spectral data is particularly challenging due to the very large number of spectral features of potential interest and the complex overlap of background, impurity and test species features. While atomic and molecular transitions leading to emission, are well known, CAP microplasmas operate under non-equilibrium conditions and emission at any given wavelength cannot be easily attributed to specific chemical pathways. The inherent high spectral dimensionality, where the number of variables (wavelengths) greatly outnumbers the sample count, increases the scope for random correlations and generally obscures valuable relationships. The dimensionality problem in multivariate analysis has led to increasing interest in data driven approaches, which enable more complex interactions to be discovered in otherwise prohibitively large datasets and produces models of detected interactions that can then be investigated further and validated. CAP microplasma spectral data poses additional challenges to machine learning algorithms which can have difficulties when the inputs are not independent – the probability for these difficulties increases as the number of wavelengths to be considered increases [12]. Depending on spectrometer resolution, single peaks may bleed into multiple nearby data points (collinearity) compounding the difficulties further. Additionally, the misalignment of the data between training and test sets, by even one data point, can cause the algorithms to discount them entirely as different signals. The ‘sharp’ peaks associated with emission spectra, in contrast to some other forms of

spectroscopy, make this issue particularly acute as a one or two column mismatch can miss the peak entirely.

Partial Least Squares (PLS) regression and its classification derivative Partial Least Squares Discriminant Analysis (PLS-DA) have been used regularly and hence become a standard tool in the analysis of chemometric and chemistry data [13]. In brief, PLS models the relationship between an input (X) and an output matrix (Y), to develop an N-dimensional hyperplane in the X space that is related as closely possible to the output matrix Y. The input X is projected into a smaller subspace of 'N' dimensions and the regression performed on those instead of on the original data. 'N' is referred to as the number of PLS latent variables. PLS can model data exhibiting noise, collinearity (correlated) variables, and high dimensionality [13]. A detailed description is given in [13]. Since the particularly complex and noisy spectral data produced by CAP's have not previously been analysed using supervised machine learning approaches, obtaining an insight into the underlying relation between the spectra and the model is a prerequisite to any sensor development strategy. Fitting a successful model and gaining insight into the relationship between it and the spectra is the main focus of this study. This also makes some other machine learning methods, such as neural-networks less suited to the task as deciphering the resultant models is much more difficult.

After the experimental setup, operational parameters and data collection protocols are outlined, we detail the statistical characteristics of the observed spectra and how they vary as CH_4 concentration is changed from 0 ppm to 100 ppm. Thereafter the computational pre-processing and machine learning procedures are described. PLS-DA models are developed for the two-class problem of determining whether unseen spectra represent a CH_4 concentration above or below a threshold, set at 2 ppm.

2 Experimental details

The emission spectra were obtained from an RF plasma formed in a 2 mm quartz capillary with an internal diameter of 0.7 mm between two exterior ring electrodes. RF power was supplied via a Plasmatech RF PSU with an inline MFJ320 matching unit. For a nominal 100W PSU set power, the absorbed power in the plasma was measured at ≤ 1.5W using an Impedans Octiv close-coupled RF current-voltage and phase probe. The gas temperature in the plasma during normal operation was estimated using a trace 0.1% N_2 admixture to generate test spectra for comparison with temperature dependent synthetic spectra generated by Specair software. For operation flow rates > 2 slm, average gas temperatures were ~335 K. Estimated confidence limits arising from the fitting are ± 50 K. Other recent measurements using an IR probe on a similar capillary He plasma indicate maximum gas temperature of < 323 K and standard error levels of 2 K. [25] The capillary outlet was a large distance (~100 cm) from the plasma to minimise atmospheric impurity back-diffusion and the system was initially conditioned to remove background impurities, over 21 days, using a 100% He plasma and exterior infra-red heating while monitoring the reduction of spectral impurity bands. Methane concentration values were set to an accuracy of 0.1 ppm using a mass flow-controlled gas dilution network. Methane and helium bottle purity was specified at 99.999%. Emission spectra were obtained using an Ocean Optics HR4000CG-UV-NIR low resolution spectrometer in the wavelength range 194 nm – 1122 nm (interval 0.25 nm), with a slit width of 5 μm and a minimum optical resolution >0.5 nm for a total of 3648 wavelength points recorded. The spectra were collected in multiple runs with different methane concentrations per run. Reference samples with 100% He (0% CH_4) are obtained at regular intervals during the He-CH_4 data collection. The experimental setup is depicted in Figure 1.

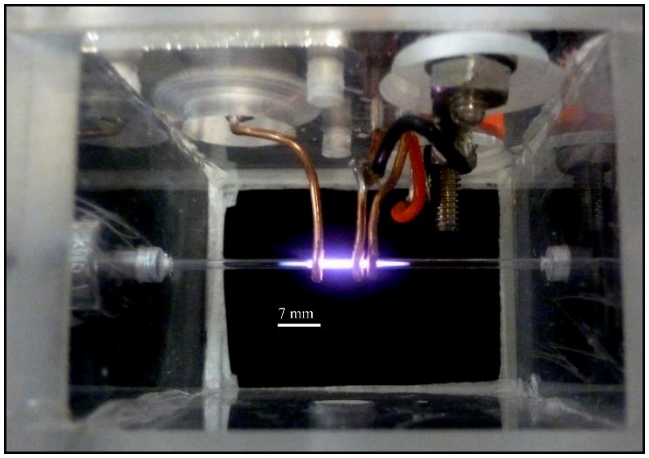

Figure 1. RF-driven (13.56 MHz) low-temperature atmospheric pressure plasma optical emission source. The electrode gap was 5 mm. The quartz capillary was 0.7 mm (id) and the effluent is expelled into laboratory air at 1.5 m downstream of plasma.

Three He-CH_4 emission spectra datasets were collected containing samples with varying methane concentrations, as listed in Table 1. Nine concentration levels, ranging from 0 ppm to 100 ppm were used in two of the datasets, S1 and S2 containing 518 and 918 samples respectively. The third, S3, contained five, from 0 ppm to 20,000 ppm across 1589 samples.

Table 1 Methane ppm levels contained within each collected dataset.

S_1 (CH_4 ppm)	S_2 (CH_4 ppm)	S_3 (CH_4 ppm)
0	0	0
1	1	2
2	2	6
4	4	10000
6	6	20000
12	12	
23	23	
77	77	
100	100	

The lower CH_4 concentration values in S3 (0 ppm, 2 ppm, 6 ppm) allow the model to be tested within its operational (trained) range. The larger concentration values (10,000 ppm, 20,000 ppm) allow the model to be tested for robustness and recovery during and after methane saturation events where the plasma conditions may be significantly altered. The spectra were preprocessed using baseline subtraction and then linearly aligned by setting the main He peak to 587.60 nm [14]. The spectra were then binned to a spectral resolution of 1 nm based on the instrument manufacturers quoted spectral resolution. The aggregation function used within bins is the maximum value of the contained wavelengths. Standard non-linear polynomial alignment is not possible, since the number of unambiguous peaks was insufficient to identify the required number of points to fit a polynomial of the necessary order, due to species overlap and the low resolution of the spectrometer. To allow for system stabilisation, spectral data obtained within 2 minutes after a change in plasma condition, were discarded.

2.1 Data spectral features

Spectral features corresponding to helium (He I, He II), hydrogen (H I, H II) and impurities (N, O, OH/H_2O) were observed. Although He lines are strong, Individual peaks cannot be assigned unambiguously to a single species except possibly those near 516 nm, which may be attributable to the C_2 Swan bands. However, these only appear at ≥ 77 ppm. Pure helium reference spectra exhibited dominant peaks in the wavelength ranges 587.6 nm to 588.2 nm and 707.1 nm to 707.5 nm both of which are within 0.8nm of known He I peaks at 587.6 nm and 706.5 nm [14]. Peak intensity at these wavelengths varied by ~36% (std. dev) across all helium spectra. A small peak at 778.5 nm, possibly O I (777.54), appeared in all spectra with almost constant intensity while other peaks varied arbitrarily with increasing CH_4 concentration. Small peaks, in the range 387 nm to 389.4 nm may include a contribution from CH which along with C2 lines around 516 nm are the only observable fragments of the parent hydrocarbon and become noticeable only at high CH_4 concentrations. The hydrogen (H_α) line at 656.3 nm may be due to H_2O and CH_4 dissociation but the latter is more likely as its intensity is noticeable only at high methane concentrations. Table 2 lists the top 9 peak wavelengths for 0 ppm and 100 ppm CH_4, ranked by intensity, along with possible species assignments. These include atomic, molecular and rotational levels.

Absolute plasma brightness can be expected to vary on introduction of molecular gases as plasma density and electron temperature compensate the change in gas chemistry. In Figure 2 the 587.6 nm peak height mean and range is shown against CH_4 concentration. The variability in intensity, irrespective of CH_4 concentration, is significant and reflects the innate variability of the plasma emission device. At 0 ppm CH_4 (i.e. 100% He), the 587.6 nm peak can be unambiguously attributed to He I. However, with the introduction of CH_4 above ~4 ppm, the peak intensity tends to decrease. A similar situation was observed for the other main wavelengths. It is clear therefore that the intensity variation of any single peak is insufficient to predict methane concentration. Note that to separate each CH_4 sample set, a 100% He plasma was operated for a short period and spectra collected. These were accounted for in the determination of mean and standard deviation for 100% He. An overlay of spectra for 0 ppm, 2 ppm and 100 ppm CH_4 is given in Figure 3(a), illustrating the variability between low and high CH_4 concentration. In figures 3(b) and 3(c), the spectra are subtracted from the 0 ppm references showing the greatest differences lie at the main He peaks (587 nm, 707nm). With the 2 ppm spectra, the other differences are almost imperceptible except for H (656 nm) and with 100 ppm, additional differences become visible at 388 nm, 431 nm and 516 nm.

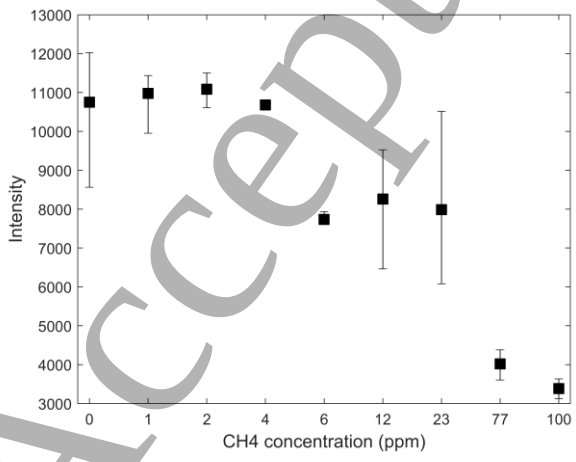

Figure 2 He 587.6 nm peak intensities versus CH4 concentration (ppm) illustrating that the He peak is not sufficient to classify the data on its own. For each concentration, the variability in peak intensities across the sample set is indicated by error bars representing the 10th to 90th percentile.

Table 2 Primary peak wavelengths and relative height for 0 ppm CH_4 and 100 ppm CH_4 spectra, and possible species assignments within observed tolerance. Assignments represent both first (I) and second (II) ionisation levels except where only one level is indicated.

	0 ppm CH_4		100 ppm CH_4		Possible Assignments
λ(nm)	#	$I_\lambda/I_{\lambda 0}$	#	$I_\lambda/I_{\lambda 0}$	
587.7	1	1.00	3	0.51	He, N_2
706.8	2	0.78	6	0.44	He, N_2
668.1	3	0.29	7	0.23	He
777.7	4	0.19	9	0.18	O I, O_2 I
389.4	5	0.18	8	0.23	He, N_2^+, N I, N_2 I, CH I
387.3	6	0.16	1	1.00	CH I, He I
516.5			2	0.65	C_2 Swan bands
656.3			4	0.51	H
336.6			5	0.49	N2rot, O_2 I

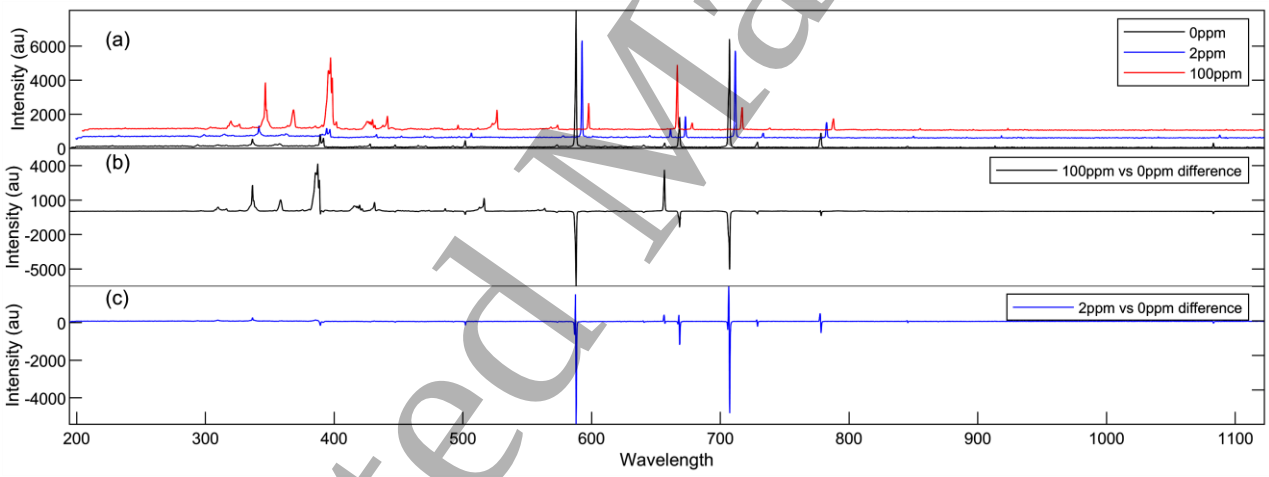

Figure 3 (a) Comparison of the spectra for 0ppm, 2 ppm, and 100 ppm, offset vertically and 0 nm, 5 nm and 10 nm for clarity. (b) The subtractive difference between 0 ppm and 100 ppm spectra. (c) Subtractive difference between 0 ppm and 2 ppm.

3 Experimental Protocols

Analysis of plasma optical emission spectra by multivariate analysis techniques present significant challenges because the number of variables (wavelengths) is very large (~3600) and locally correlated i.e. due to broadening, the variables near an emission line contain no additional information yet can be judged by the model as significant. Using the selected algorithm, multivariate models are built to best fit the input variables to the output concentration value and within the algorithm, different numbers of components (latent variables (LV)) can be chosen. Increasing the number of LV's when generating a model often improves the predictive accuracy within a single dataset but when too many LV's are used, the model performs less well on unseen datasets. This is known as overfitting. The risk of overfitting also increases with the number of variables [26,27]. In order to investigate model

overfitting as well as accuracy, we performed five different analyses using protocols 1 - 5, a summary of which is given in Table 3. Each protocol involves varying the number of LV's to determine the accuracy dependence on LV, for the dataset specific to that protocol. Cross-validation is a standard model assessment technique to estimate how a potential model will generalise to an independent data set and when used with a single dataset, the data is partitioned into complementary train and test subsets. In protocol 1, models built using individual single session datasets are subjected to cross-validation, while in protocol 3, two such datasets are merged and protocol 1 repeated. The cross-validation protocols (P1, P3) are designed to provide an estimate of generalizability and provide an initial assessment of the model. The data is first partitioned using stratified random sampling into 10 subsets. The model is then trained using 9 of these and tested on the remaining 1. This is repeated until all 10 subsets have been used for testing, resulting in an accuracy value for each subset. The entire cross-validation process is then repeated 10 times producing 100 results from which a mean and standard deviation are taken. More rigorous validation of these models is then undertaken via the train/test protocols (P2, P4, P5) which test these models on independent unseen data from entirely different data recording sessions. In P2, models built using dataset S1 (in P1) are tested on dataset S2 and vice versa. In protocols P4 and P5, models from P3 using merged datasets are tested on the unseen dataset, S3. Linear alignment was applied in all protocols except protocol (P5) where a custom non-linear alignment procedure, explained in more detail below, is applied. Linear alignment consists of shifting each spectra left or right so that its maximum peak was placed at the reference wavelength for He at 587.6 nm.

Partial Least Squares Discriminant Analysis (PLS-DA) was used as the classification algorithm in all protocols. PLS-DA models were built to classify spectra into one of the two categories depending on whether the methane concentration of an unknown sample was above or below a threshold value. The threshold was initially set at ≥ 2 ppm such that class A represents 0 or 1 ppm and class B represents the remainder. PLS-DA models were built using a range of latent variables (LVs) from 1 to 15. In order to provide a baseline performance, the Zero Rule algorithm (ZeroR) was used. ZeroR predicts the class that represents the majority of samples in the dataset. Finally, the relative importance of each wavelength in the model decision was determined using the Variable Importance in Projection (VIP) score [13]. The VIP score is defined as the weighted sum of squares of the PLS-weights, where the weights are calculated from the Y-variance explained by each PLS component". [13] When spectrometer resolution is insufficient to resolve all peaks, the possibility exists for a given peak to be allocated to different nearby wavelength variables from sample to sample. This imperfect alignment of samples may be expected to present difficulties for PLS-DA. Therefore, with protocols P1 – P4, each spectrum was wavelength shifted so that its maximum peak was placed at the reference He wavelength, 587.6 nm. However, in certain regions of the spectrum, e.g. 300-400 nm, a large number of peaks exist, and this simple realignment scheme may not be satisfactory. A standard procedure would utilise domain expertise in assigning uniquely identifiable peaks to known locations, which are then used to fit a N-order polynomial to re-evaluate and correct the spectra. However, traditional non-linear alignment methods such as polynomial fitting are hampered in this case by the lack of any additional uniquely identifiable species wavelengths and the difficulties are exacerbated by the peak broadening due to the low spectrometer resolution. An alternative method was therefore implemented using anchor wavelength values selected on the assumption that peaks should be present in spectra at wavelengths that were judged by the model to be of sufficient significance. A VIP – anchor algorithm was developed and is summarised as follows; VIP scores are used to identify the most significant peaks which are then ranked in descending order of importance. For each spectrum, the first N peaks within a predefined tolerance range of the anchor wavelengths are

determined. These are then used to generate a N-order polynomial, from which the non-linear alignment of the spectrum is obtained.

4 Results

Results are presented in figure 5 for protocols 1 and 2 and in figure 6 for protocols 3 – 5 and all results showed a level of improvement over the ZeroR baseline (i.e. guessing the most common class). The within-session accuracy from P1 showed a rising value with the number of PLS (LV) components used in the model and reached the 90% predictive accuracy threshold with 6 LVs. However, the P2 results exhibited a large drop in accuracy (~20%) and highlight the loss of performance when a model is tested with data from an unseen session i.e. one that provided no data for training. Such an impact of sessional differences on the models is not unexpected given the low cost restriction on the plasma source and spectrometer and the observed spectral variability e.g. figure 2. Increasing the number of LV's did not rectify this loss of performance and hence further exploration is warranted. In Protocols P3 we merged two sessions, increasing the overall number of samples and as with P1, this indicated the similar trend in model accuracy with increasing LV numbers. However testing this merged session data using unseen data from S3, again led to a fall in performance, as shown in P4. In this case the maximum accuracy (86%) obtained with 7 LVs is better than that obtained from P2, indicating a beneficial effect of using data from multiple sessions by ~12%. However this improvement is insufficient. Protocol P4 was repeated in P5 but with the application of a custom non-linear alignment and this resulted in the greatest accuracy, 96%, within the margin of error of 10-fold cross-validation in P3. A summary of the results for each protocol can be seen in Table 3 while the variation in accuracy against the number of PLS latent variables in each model is shown in Figure 5 for P1 and P2 and in Figure 6 for P3 – P5.

In order to explore further the source of errors, we observe that all samples containing ≥ 2 ppm CH_4 were classified correctly i.e. 100% accuracy. The incorrect predictions came only from samples with < 2 ppm CH_4. The confusion matrix for P4, which breaks down the model's accuracy above or below the 2 ppm threshold by class, is given in Table 4. Protocol 5 demonstrates the success of this machine learning approach in generating accurate models when exposed to unseen data. The S3 test set contains CH_4 ppm values well above what the model was exposed to during its training (20000 ppm vs. 100 ppm) illustrating it can cope outside the upper end of its range. To test for untrained values within its range, the 6 ppm CH_4 sample data was removed from S1 and S2 before training while S3 data used for testing included 6 ppm samples. If the 6 ppm samples are not removed from the training data, then the maximum accuracy of P4 increases by only a small amount (88%). We observe that all 6 ppm samples are classified correctly as ≥ 2 ppm and thus the removal of 6 ppm samples leads to additional misclassifications of the < 2 ppm classes. The fact that a custom polynomial alignment routine leads to such an improvement in outcome obviously implicates the spectral variation between samples and between sessions. Temporal brightness variation in the plasma, at a constant condition, is clearly observed while instrument jitter can span a number of wavelength bins and the exact position of a peak maximum fluctuates in time and by different amounts across an individual spectrum. Thus the standard linear (shift) alignment is not sufficient to force all of the important wavelengths into the equivalent bin for each spectrum, resulting in bin jitter. This is important for model development as the algorithm treats each bin as of equal weight. Since improvement in source and instrument is not an option for field portable applications, the burden is on the model development to accommodate such variability and hence a more rigorous polynomial alignment routine was investigated.

Table 3 Summary of protocol results, rounded to two significant figures. Accuracy is defined as the percentage of spectra correctly classified. For cross validation tests where multiple accuracies are produced 2 standard deviations are given as a range.

Protocol	Datasets	Method	Accuracy ±2σ	Latent variables	Increase from ZeroR
P1	S1 / S2	10-fold cross validation on S1/S2 individually	98% ±3	12	46%
P2	S1, S2	Train on S1, test on S2	78%	8	13%
P3	S1+S2	10-fold cross validation	97% ±3	8	42%
P4	S1+S2, S3	Train on S1+S2, test on S3	86%	7	44%
P5	S1+S2, S3	Train on S1+S2, test on S3 – custom alignment	96%	9	62%

Table 4 Protocol 4 confusion matrix, rounded to two significant figures.

		Predicted class	
		< 2 ppm	≥ 2 ppm
True class	< 2 ppm	82%	18%
	≥ 2ppm	0%	100%

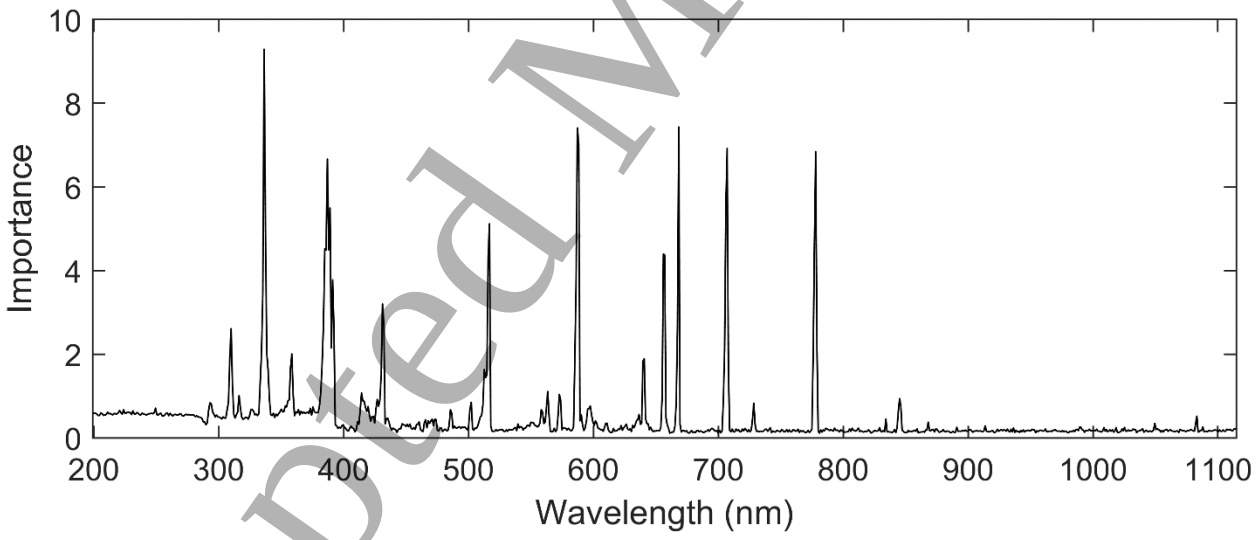

Figure 4 Variable Importance in Projection (VIP) score for each wavelength, obtained from protocol P5. VIP scores provide a relative importance value for each wavelength in the generated model.

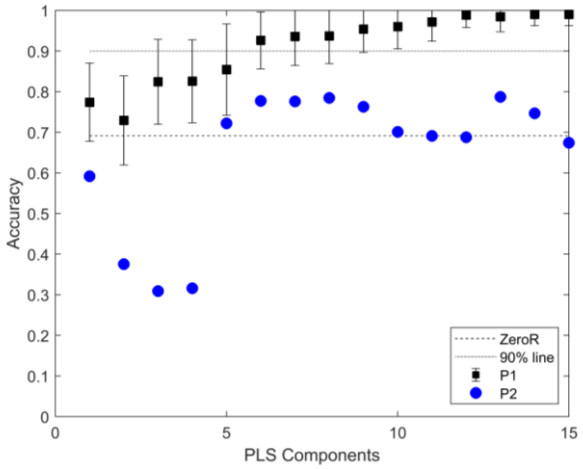

Figure 5 Graph of accuracy for protocols P1 and P2 against number of PLS latent variables in the model. The ZeroR line indicates the performance that could be achieved if the model continuously guessed the most common class; it serves as a baseline for performance.

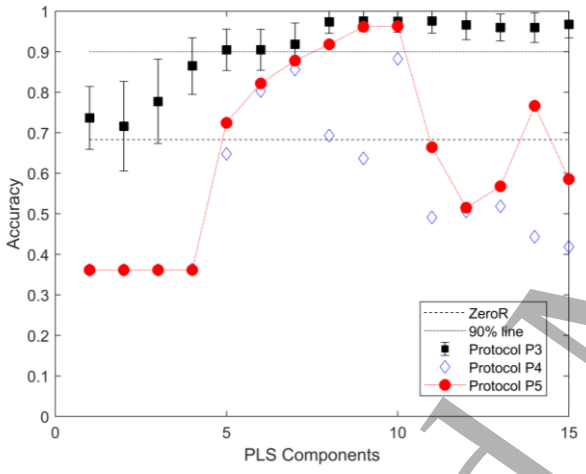

Figure 6 Graph of accuracy for protocols P3, P4, and P5 against number of PLS latent variables in the model. Connecting lines have been included as a guide to the eye.

Using the Variable Importance in projection (VIP) score the PLS-DA model can be broken down to determine which variables (wavelengths) were the most significant for model predictions. As protocol 5 proved the most successful, its model was chosen for analysis. The graph of the VIP scores across the spectrum can be seen in Figure 4, indicating ~336.4 nm as the most important peak, followed by 587.9 nm, 707.8 nm and 669.1 nm, all holding similarly high scores. It also showed a number of peaks in the data such as the 357.6 nm and 316.1 nm peaks to be largely irrelevant. A breakdown of the top 12 significant wavelengths and their possible assignments is presented in Table 5. Given the complexity of the plasma chemistry and the low instrument resolution, multiple species assignment to a given emission line are possible. While knowledge of the underlying chemistry would be valuable, it is not currently available. We therefore confine ourselves to considering only three broad categories of methane-related species (C_xH_y x: $0 \rightarrow 2$,y: $0 \rightarrow 4$), impurities (O_x, N_x, H_xO_y) and helium. From the VIP scores, the impurity peaks in the range 335 nm – 337 nm rank highest after helium peaks at 587 nm and 388 nm. In figure 7(a) the intensity of the highest peak in the impurity range 335 nm – 337 nm and its relative height compared to the highest He peak were both observed to increase linearly with concentration. The impurity intensity

surpassed that of He (587 nm) at 77 ppm. For methane related species, similar plots of atomic H (656 nm) and CH (431 nm) are given in figures 7(b) and 7(c). While the H intensity increases linearly, surpassing the He (587 nm) intensity at a concentration between 23 ppm and 77 ppm, the CH line intensity exhibits the lowest slope. The line at 388 nm, figure 7(c) which may include contributions from He and CH, shows limited change at low concentrations, increasing rapidly at 77 ppm.

Table 5 a summary of Figure 3, listing the top 12 wavelengths of significance in the algorithm as obtained from VIP scores. Neighbouring VIP score values within the indicated wavelength range are summed. Wavelength assignment to chemical species. Where more than one species is listed, this indicates significant peak overlap between species. Impurity species include OH, N_2 and O_2.

Rank	Centre Wavelength	+/- (nm)	Assignment(s)	
1	587.18	1.0	He	
2	388.05	1.1	He	CH
3	336.35	1.1	Impurity	
4	706.57	0.5	He	
5	777.46	0.5	Impurity	
6	516.22	0.5	C_2 Swan	
7	667.82	0.5	Impurity	He
8	656.56	0.5	H	
9	431.64	0.5	CH	
10	384.89	0.0	Impurity	
11	385.94	0.0	Impurity	
12	391.20	0.0	Impurity	

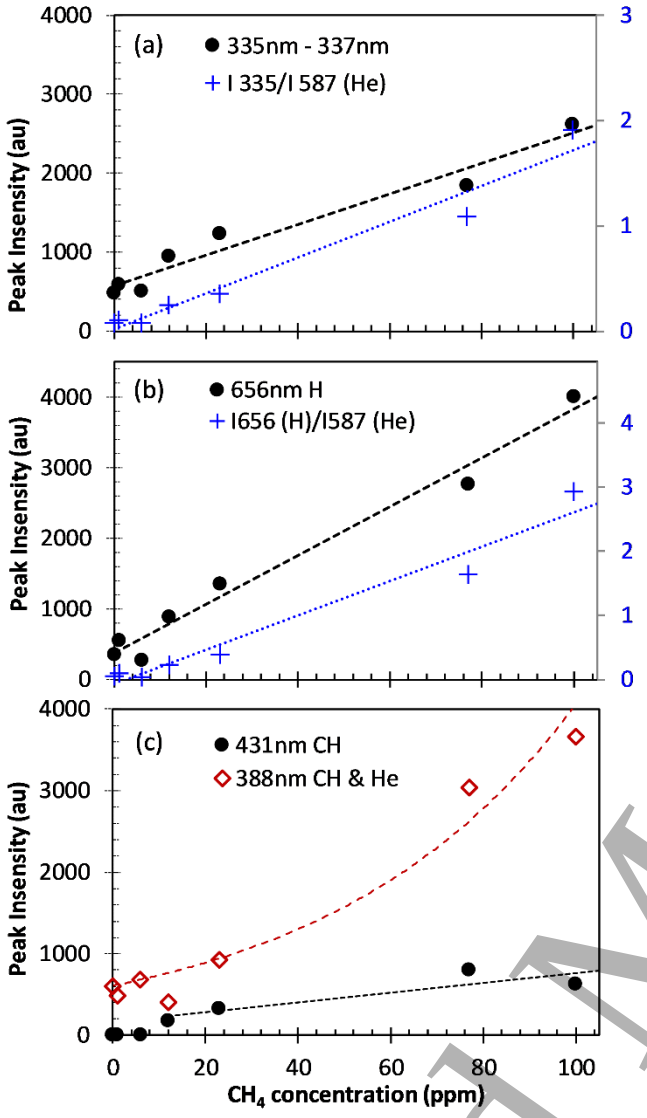

Figure 7 Summary of important spectral features variation with CH4 concentration, obtained from VIP with the highest scores: (a) 335 nm – 337 nm impurity peak, (b) atomic H and (c) peaks at 388 nm and 431 nm that may contain contributions for CH. Also shown (blue +) in panels (a) and (b), is ratio of peak height to that of the He peak height at 587 nm (the maximum peak at 0% CH4) where the H and impurity lines supersede that of the main He line at ≥ 77 ppm. The standard error is below the limits where error bars are visible by comparison to symbols and are therefore not shown.

5 Custom Alignment

Protocols P3 and P4 show a substantial improvement over P2, but a disparity in performance between the cross-validation P3 and the test outcome, P4, remains. We investigated the use of a VIP – anchor wavelength alignment scheme to improve performance. Using aligned data with Protocol 5, the performance improved by 10% over P4, showing an accuracy of 96%, Table 3. This is within the margin of error of the 97% ±3 achieved during cross-validation with the same training data (P4) and suggests that the accuracy is at or near the maximum potential of the model produced during training.

The variation in accuracy with the number of PLS latent variables is given in Figure 6. Additionally, the confusion matrix for this protocol, Table 6, shows that this procedure

increased the accuracy of the system at < 2 ppm without any reduction in the performance at ≥ 2 ppm.

Table 6 Protocol 5 confusion matrix, rounded to two significant figures.

		Predicted class	
		< 2ppm	≥ 2ppm
True class	< 2ppm	94%	6%
	≥ 2ppm	0%	100%

6 Discussion

We have demonstrated the capability of low-cost emission spectroscopy to identify trace methane at parts per million levels. This represents an important step forward in establishing proof of principle for field-deployed trace gas analysis for methane and other gases which is important in many applications from industrial and environmental sensing to clinical diagnostics. The variability of the plasma source is seen in the intensity variation of carrier gas lines without added test gas. While this source has not been optimized for this purpose, the algorithm with the inclusion of a model-based custom alignment scheme has achieved 96% accuracy in identifying trace CH_4 with a LOD of 2 ppm. Future advances will depend on simultaneous complementary development of model and device and this in turn requires more detailed knowledge of the induced plasma chemistry and its underlying impact on model operation.

In order to better understand the algorithm's mechanism of trace gas identification, knowledge of the underlying plasma chemistry would be beneficial. However detailed studies, e.g. via mass spectrometry, have been limited especially with low temperature high pressure plasmas and hence knowledge of the effect of molecular admixtures, particularly hydrocarbons, is lacking. Collision processes, branching ratios and dissociation pathways in plasmas are complex [15]. Methane dissociation in the plasma is initiated by collisions with energetic electrons and results in the formation of CH_x radicals and hydrogen [16], whereupon hydrogen abstraction from CH_x by atomic hydrogen becomes an important channel for the further dissociation of methane, competing with the electron dissociation pathway [17]. Dimerisation reactions between CH and CH_x lead to the formation of C_2H_x, mostly C_2H_2 at high temperature/energy, while at lower temperatures, the low concentration of H favours CH_3 + $CH_3 \rightarrow C_2H_6$. These reactions are often exponentially dependent on gas temperature, with thresholds typically > 1000 °C. Hydrogen abstraction from stable C_2H_x is favoured at high temperature.

We have measured hydrocarbon plasma chemistry, at pressures below atmosphere, using mass spectrometry of neutral and ions. Spectra contain multiple neutral and ionic fragments, with typically ~30 mass spectral lines up to 60 amu. [18]. Of the neutral hydrocarbon species, the parent molecule remained dominant indicating only a small overall dissociation fraction, typically less than 15%. In helium-based plasmas, CH_4 dissociation is reported to decrease with increasing pressures [19]. Also, the measured total fraction of ionised particles, up to atmospheric pressure, is less than 10 ppm. [20]. Liao et al. observed a large number of C_xH_y ($x \leq 10$, $y \leq 12$) fragment mass peaks from a DC Arc Ar-CH_4 plasma operating at a gas temperature > 2800 K and at a high CH_4 concentration (≤ 2%). They also obtained emission spectra but the only molecular derived features observed were weak CH(A-X) and H_β lines and stronger lines representing H_α and C_2 Swan systems [21]. Therefore, we can expect trace methane in our plasma to remain predominantly in molecular form, with dissociated fragments at sub-ppm concentrations. The CH_4 molecule has no emission in the captured spectral range

and it is unlikely, at such low concentrations, that CH_x fragments contribute significantly to the algorithm training.

The VIP scores in Figure 4, and listed in Table 5, show the significance of each peak to the algorithm with 587 nm being the most important, followed by 388 nm, 336 nm, 706 nm and 777 nm. Thus, the top 5 peaks reported by VIP represent He or impurity lines. The 388 nm peak, rank 2, is strongly associated with the He transition (1s2s - 1s3p) at 388.86 nm. However, there is also a nearby transition for CH at 388.90 nm. The peak height is almost constant at low CH_4 concentration but increases significantly at > 77 ppm, which suggests emission is primarily due to CH_x in this range only and that CH transitions play a limited role near the trace threshold. The 336.4 nm peak is observed in all spectra, increasing linearly with CH_4 concentration, surpassing the main He peak (587 nm) at 77 ppm CH_4. There are multiple N_2 rotational and vibrational molecular bands between 335.698 nm to 336.746 nm while O2 I has listed peaks at 335.68 nm and 337.01 nm. However O_2 is a weak emitter and its presence is unlikely under these conditions. The Hα line at 656.28 nm, rank 8, is the most prominent CH_4 fragment peak (although high purity He and CH_4 gases contain trace H_2O) but its intensity is low until 77 ppm CH_4. Low intensity CH_x fragment emission peaks are observed at 431 nm, increasing slightly with CH_4 concentration. Emission due to C2 Swan vibrational states around 516 nm are only visible at ≥ 77 ppm. The C2 Swan vibrational bands correspond to transitions between the $d^3\Pi_g$ (2.48 eV) and $a^3\Pi_u$ (0.09 eV) electronic states. They indicate the final hydrogen abstraction endpoint from CH_4 and normally require high temperature. However, in the presence of noble gases, particularly Ar and to a lesser extent He, metastable Ar* (or He*) induced dissociation of C_2H_x can be important [22].

The VIP scores show that, while CH_4 molecular fragment emission may have some impact, variation in carrier gas emission and impurity emission play the more important role in the PLS-DA model. In essence, the algorithm detects the indirect effect of CH_4 admixtures on these noble gas or impurity features and associates these indirect changes to a concentration threshold. In Figure 2, the height of the main He peak (587.6 nm) varies significantly with CH_4 concentration but the wide error bars prevent the use of this peak as a simple concentration metric. Other large He peaks show similar variability. Note that in non-equilibrium plasmas such as those used here, the intensity of any excited species spectral line does not follow a simple relationship to their ground-state concentration. A change in intensity reflects an underlying change in the electron energy distribution which is responsible for excitation processes. A low concentration molecular admixture into a rare gas significantly changes the electron energy distribution by introducing additional inelastic low energy processes including the excitation of low lying (< 0.1 eV) vibrational and rotational states [23]. An increase in gas temperature is not uncommon with molecular admixtures [24]. In summary, the emission characteristics and underlying plasma chemistry represent a highly complex set of ill-defined interactions which pose an almost insurmountable barrier to obtaining scientific analysis that would be accurate enough for predictive spectral identification. Nevertheless, using a data-driven approach, we have achieved such a predictive capability down to the parts per million range. Future development of this technique will target more complex gas mixtures and enhanced gas constituent discrimination via multi-category algorithm approaches.

From the model perspective, the VIP scores allowed the implementation of a successful wavelength alignment algorithm. This significantly increases the performance in protocol 5, which represents the greatest identification challenge. Traditional alignment methods often utilise best case alignment against reference spectra in the training set. By choosing to align against the computed model, the algorithm focuses on the specific regions that the model deemed important – regardless of their height in the reference spectra. This proved to be

especially important in this dataset, as the peaks in the region of highest importance are not the most significant in the dataset with regards to intensity.

7 Conclusion

This study analysed emission spectra from a helium - methane radio frequency cold atmospheric plasma. Spectra were captured using a low-cost and low-resolution spectrometer over three independent experimental runs containing varying CH_4 concentrations from 0 ppm through 100 ppm. The model was also tested for robustness and recovery during and after methane saturation events with concentrations up to 20,000 ppm. The spectra were analysed using a data driven approach with the goal of detecting methane concentrations at or above 2 parts per million. Overall the method achieved an initial accuracy of 86% which was increased to 96% with the application of a custom alignment procedure. This illustrates that a machine learning/data driven approach presents potential in making use of otherwise prohibitively complex cold atmospheric plasma emission spectra in controlled circumstances with low resolution instruments.

8 AUTHOR INFORMATION

Funding Sources

Engineering and Physical Sciences Research Council (EP/K006088), Invest N. Ireland (RD0714186)
The authors declare no competing financial interest.

9 References

[1] Environmental Protection Agency, “Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2016,” 2018.

[2] R. Alvarez, S. Pacala, J. Winebrake, W. Chameides and S. Hamburg, “Greater focus needed on methane leakage from natural gas infrastructure,” *Proceedings of the National Academy of Sciences of the United States of America,* vol. 109, pp. 6435-6440, 2012.

[3] L. Hwang, K. Low, R. Khoshini, G. Melmed, A. Sahakian, M. Makhani, V. Pokkunuri and M. Pimentel, “Evaluating Breath Methane as a Diagnostic Test for Constipation-Predominant IBS,” *Digestive Diseases and Sciences,* vol. 55, no. 2, pp. 398-403, 2010.

[4] Z. Du, S. Zhang, J. Li, N. Gao and K. Tong, “Mid-Infrared Tunable Laser-Based Broadband Fingerprint Absorption Spectroscopy for Trace Gas Sensing: A Review,” *Applied sciences,* vol. 9, no. 2, p. 338, 2019.

[5] L. Dong, F. Tittel, C. Li, N. Sanchez, H. Wu, C. Zheng, Y. Yu, A. Sampaolo and R. Griffin, “Compact TDLAS based sensor design using interband cascade lasers for mid-IR trace gas sensing,” *Optics Express,* vol. 24, no. 6, pp. 528-535, 2016.

[6] M. Winchestera and R. Payling, “Radio-frequency glow discharge spectrometry:: A critical review,” *Spectrochimica Acta Part B: Atomic Spectroscopy,* vol. 59, no. 5, pp. 607-666, 2004.

[7] S. Liu, Y.-L. Yu and J.-H. Wang, “Advances in discharge-based microplasmas for the analysis of trace species by atomic spectrometry,” *Journal of Analytical Atomic Spectrometry,* vol. 32, no. 11, pp. 2118-2126, 2017.

[8] S. Barua, I. M. Rahman, M. Miyaguchi, A. Mashio, T. Maki and H. Hasegawa, “On-site analysis of gold, palladium, or platinum in acidic aqueous matrix using liquid electrode plasma-optical emission spectrometry combined with ion-selective preconcentration,” *Sensors and Actuators B: Chemical,* vol. 272, pp. 91-99, 2018.

[9] Y. Cai, Y.-L. Yu and J.-H. Wang, “Alternating-Current-Driven Microplasma for Multielement Excitation and Determination by Optical-Emission Spectrometry,” *Analytical Chemistry,* vol. 90, no. 17, p. 10607–10613, 2018.

[10] J. Vincent, H. Wang, O. Nibouche and P. Maguire, “Differentiation of Apple Varieties and Investigation of Organic Status Using Portable Visible Range Reflectance Spectroscopy,” *Sensors,* vol. 18, no. 6, 2018.

[11] W. Song, H. Wang, O. Nibouche and P. Maguire, “Differentiation of organic and non-organic apples using near infrared reflectance spectroscopy — A pattern recognition approach,” in *2016 IEEE SENSORS*, 2016.

[12] B. Clarke, E. Fokoue and H. Helen, “Principles and Theory for Data Mining and Machine Learning,” in *Principles and Theory for Data Mining and Machine Learning*, Springer, 2009, pp. 9-10.

[13] S. Wold, M. Sjöström and L. Eriksson, “PLS-regression: a basic tool of chemometrics,” *Chemometrics and Intelligent Laboratory Systems,* vol. 58, no. 2, pp. 109-130, 2001.

[14] A. Kramida, Y. Ralchenko, J. Reader and NIST ASD Team, *NIST Atomic Spectra Database,* Gaithersburg: National Institute of Standards and Technology, 2019.

[15] D. Reiter and R. K. Janev, “Hydrocarbon Collision Cross Sections for Magnetic Fusion: The Methane, Ethane and Propane Families,” *Contributions to Plasma Physics,* vol. 50, no. 10, pp. 986-1013, 2010.

[16] M. Danko, J. Orszagh, M. Ďurian, J. Kočišek, M. Daxner, S. Zöttl, J. Maljković, J. Fedor, P. Scheier, S. Denifl and Š. Matejčík, "Electron impact excitation of methane: determination of appearance energies for dissociation products," *Journal of Physics B: Atomic, Molecular and Optical Physics,* vol. 46, no. 4, 2013.

[17] M. Heintze, M. Magureanu and M. Kettlitz, "Mechanism of C2 hydrocarbon formation from methane in a pulsed microwave plasma," *Journal of Applied Physics,* vol. 92, no. 12, pp. 7022-7031, 2002.

[18] A. Baby, C. Mahony and P. D. Maguire, "Acetylene–argon plasmas measured at a biased substrate electrode for diamond-like carbon deposition: I. Mass spectrometry," *Plasma Sources Science and Technology,* vol. 20, no. 1, 2011.

[19] K. Katayama, S. Fukada and M. Nishikawa, "Direct decomposition of methane using helium RF plasma," *Fusion Engineering and Design,* vol. 85, no. 7, pp. 1381-1385, 2010.

[20] P. D. Maguire, C. Mahony, C. P. Kelsey, A. J. Bingham, E. P. Montgomery, E. D. Bennety, H. E. Potts, D. Rutherford, D. A. McDowell, D. A. Diver and D. Mariotti, "Controlled microdroplet transport in an atmospheric pressure microplasma," *Applied Physics Letters,* vol. 106, no. 22, 2015.

[21] M. Liao, Y. Wang, H. Wu, H. Li and W. Xia, "Study of Non-Thermal DC Arc Plasma of CH4/Ar at Atmospheric Pressure Using Optical Emission Spectroscopy and Mass Spectrometry," *Plasma Science and Technology,* vol. 17, no. 9, pp. 743-748, 2015.

[22] U. Fantz and S. Meir, "Correlation of the intensity ratio of C2/CH molecular bands with the flux ratio of C2Hy/CH4 particles," *Journal of Nuclear Materials,* vol. 337, pp. 1087-1091, 2005.

[23] A. M. Lietz and M. J. Kushner, "Molecular admixtures and impurities in atmospheric pressure plasma jets," *Journal of Applied Physics,* vol. 124, no. 15, p. 153303, 2018.

[24] J. Voráč, P. Synek, V. Procházka and T. Hoder, "State-by-state emission spectra fitting for non-equilibrium plasmas: OH spectra of surface barrier discharge at argon/water interface," *Journal of Physics D,* vol. 50, no. 29, p. 294002, 2017.

[25] N. Hendawy, H McQuaid, D. Mariotti, P Maguire "Continuous gas temperature measurement of cold plasma jets containing microdroplets, using a focussed spot IR sensor", under review

[26] W Song, H Wang, P Maguire, O Nibouche, "Collaborative representation based classifier with partial least squares regression for the classification of spectral data", *Chemometrics and Intelligent Laboratory Systems,* 182, pp. 79-86

[27] W Song, H Wang, P Maguire, O Nibouche, "Nearest clusters based partial least squares discriminant analysis for the classification of spectral data", *Analytica chimica acta* 1009, pp. 27-38

[28] D. Gidon, X. Pei, A. D. Bonzanini, D. B. Graves and A. Mesbah, "Machine Learning for Real-Time Diagnostics of Cold Atmospheric Plasma Sources," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 5, pp. 597-605, Sept. 2019, doi: 10.1109/TRPMS.2019.2910220.

[29] Ali Mesbah and David B Graves, "Machine learning for modeling, diagnostics, and control of non-equilibrium plasmas", *Journal of Physics D: Applied Physics*, vol. 52, no. 30, pp. 30LT02