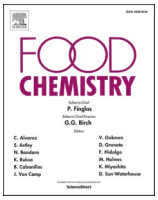
Food Chemistry 404 (2023) 134632

Contents lists available at ScienceDirect

Food Chemistry

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



Multivariate versus machine learning-based classification of rapid evaporative Ionisation mass spectrometry spectra towards industry based large-scale fish speciation

Marilyn De Graeve [a,1], Nicholas Birse [b,1], Yunhe Hong [b], Christopher T. Elliott [b,c], Lieselot Y. Hemeryck [a,2], Lynn Vanhaecke [a,b,*,2]

[a] Laboratory of Integrative Metabolomics, Department of Translational Physiology, Infectiology and Public Health, Ghent University, Salisburylaan 133, B-9820 Merelbeke, Belgium
[b] Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, Belfast BT9 5BN, United Kingdom
[c] School of Food Science and Technology, Faculty of Science and Technology, Thammasat University (Rangsit Campus), Khlong Luang, Pathum Thani 12120, Thailand

ARTICLE INFO

Keywords:
Ambient Ionisation Mass Spectrometry
Multivariate Chemometric Modelling
Machine Learning
Fish Speciation
Real-time Prediction
Metabolomics

ABSTRACT

Detection and prevention of fish food fraud are of ever-increasing importance, prompting the need for rapid, high-throughput fish speciation techniques. Rapid Evaporative Ionisation Mass Spectrometry (REIMS) has quickly established itself as a powerful technique for the instant *in situ* analysis of foodstuffs. In the current study, a total of 1736 samples (2015–2021) - comprising 17 different commercially valuable fish species - were analysed using iKnife-REIMS, followed by classification with various multivariate and machine learning strategies. The results demonstrated that multivariate models, i.e. PCA-LDA and (O)PLS-DA, delivered accuracies from 92.5 to 100.0%, while RF and SVM-based classification generated accuracies from 88.7 to 96.3%. Real-time recognition on a separate test set of 432 samples (2022) generated correct speciation between 89.6 and 99.5% for the multivariate models, while the ML models underperformed (22.3–95.1%), in particular for the white fish species. As such, we propose a real-time validated modelling strategy using directly amenable PCA-LDA for rapid industry-proof large-scale fish speciation.

1. Introduction

Worldwide, there is an increased demand for fish, which is being met by expanding traditional capture fisheries and overfishing, as well as by switching to aquaculture (FAO, 2020). Overfishing has contributed significantly to the global collapse of marine fisheries, although there are also other influencing factors, such as natural predation and climate change (Lima, Canales, Wiff, & Montero, 2020). The drop in stock biomass, increased demand for fish in general, and certain fish species specifically create conditions for a number of different illegal activities to take place, including illegal fishing activities, substitution and mislabelling frauds (Fox et al., 2018). Illegal, unreported and unregulated (IUU) fisheries are believed to be responsible for approximately 26 million tonnes of fish caught each year (FAO, 2020). IUU fisheries undermine the efforts taken by governments and conservationists to protect endangered species and return fisheries to sustainable levels. Mislabelling and substitution moreover threaten food integrity, consumer trust and public health (Fox et al., 2018). Therefore, to be able to

Abbreviations: AIMS, Ambient Ionisation Mass Spectrometry; CV, Cross-Validation; ELISA, Enzyme-Linked Immunosorbent Assays; HRMS, High Resolution Mass Spectrometry; IUU, Illegal, Unreported and Unregulated; LC-MS, Liquid Chromatography coupled to Mass Spectrometry; LDA, Linear Discriminant Analysis; MALDI, Matrix-Assisted Laser Desorption/Ionization; ML, Machine Learning; MS, Mass Spectrometry; OPLS-DA, Orthogonal Partial Least Squares-Discriminant Analysis; PC, Principal Component; PCA, Principal Component Analysis; PCR, Polymerase Chain Reaction; QC, Quality Control; REIMS, Rapid Evaporative Ionisation Mass Spectrometry; RF, Random Forest; RT, Retention time; SVM, Support Vector Machine; TIC, Total Ion Count; UHPLC, UltraHigh Performance Liquid Chromatography; UV, UniVariate scaling.

* Corresponding author at: Laboratory of Integrative Metabolomics, Department of Translational Physiology, Infectiology and Public Health, Ghent University, Salisburylaan 133, B-9820 Merelbeke, Belgium.
E-mail address: Lynn.Vanhaecke@UGent.be (L. Vanhaecke).
[1] Shared first author.
[2] Shared last author.

https://doi.org/10.1016/j.foodchem.2022.134632
Received 5 July 2022; Received in revised form 20 September 2022; Accepted 13 October 2022
Available online 17 October 2022

counteract fraudulent activities, post-harvest and production species identification is required to expose the illegal capture and sale of protected and endangered species, as well as substitution or mislabelling of inferior, lower-value (e.g. more widely available) fish species as higher-value species. For this purpose, accurate and precise as well as high-throughput fish speciation methodologies are required to meet the need of the fisheries sector.

The currently available techniques for industrial fish speciation are varied, ranging from inexpensive rapid testing techniques such as enzyme-linked immunosorbent assays (ELISA), costly and time-consuming molecular biology techniques such as polymerase chain reaction (PCR) and genome sequencing that can be used to accurately determine species and breed, to proteomics and metabolomics approaches making use of mass spectrometry (Cutarelli et al., 2014; Lasch et al., 2019; Rasmussen & Morrissey, 2008; Ruethers et al., 2020; Stahl & Schröder, 2017). Mass spectrometry-based approaches have the specific advantage of being able to simultaneously assess freshness, speciation, pre-catch conditions, fishing period and gear, fish size, physiology and metabolism, geographic and/or farming origin, etc. in both fresh and processed fish. This is a distinct advantage compared to several existing as well as rapidly emerging techniques including e.g. element profiling approaches (limited to geographic origin) (Anderson, Hobbie, & Smith, 2010; Varrà et al., 2021), and artificial imaging strategies (limited to uncut fish, based on e.g. neural networks) (Navotas et al., 2018).

The amount of information provided by an analysis is typically proportional to the time taken to run that analysis, with reliability similarly dependent on the time required per analysis. ELISA is fast but demonstrates higher error rates that can frequently reach 40 %, often due to cross-reactivity in closely related species. PCR, MALDI (Matrix-Assisted Laser Desorption/Ionization) and LC-MS (Liquid Chromatography coupled to Mass Spectrometry) methods, on the other hand, produce more in-depth information and/or have very low error rates, yet entail time and/or cost per-sample penalties (Black et al., 2017; Trotta et al., 2005). Interestingly, Ambient Ionisation Mass Spectrometry (AIMS) is a relatively recent development, which enables the analysis of unprocessed or minimally processed food samples (Birse et al., 2021). The ion sources developed for ambient mass spectrometry are typically mounted to the same models of mass spectrometers used in conventional LC-MS analysis, meaning the techniques can provide rich results with very low error rates, but without the time- and cost-related drawbacks of conventional MS (Black, Chevallier, & Elliott, 2016). Rapid Evaporative Ionisation Mass Spectrometry (REIMS), a specific type of AIMS, was developed to make use of the analyte-rich smoke aspirated from electrosurgical knives used in surgery to assist in rapidly differentiating cancerous from healthy tissue (Schäfer et al., 2009). More specifically, the technique works by generating mass spectra of the lipids released from the cell wall across one or more mass ranges, generating a metabolic fingerprint. The fingerprint of a tissue sample is then compared to reference samples of known provenance or authenticity, such as a specific species or a production system (Birse et al., 2021; Black et al., 2019). This process can be completed in seconds, in the case of REIMS' original cancer surgery application, this speed allows the surgical team to quickly diagnose and more easily remove cancerous tissue during surgery (Phelps et al., 2018). For food testing, the technique was first applied in the aftermath of the horsemeat scandal that swept across Britain and Europe in 2011 and this speed could be used to enable significant numbers of samples to be tested, ensuring representative sampling (Balog et al., 2016; Birse et al., 2022). Samples of mammalian tissue, most typically commercially available meat and fish fillets are cut using the same type of electrosurgical knife, after which the analyte-rich smoke is aspirated into a time-of-flight mass spectrometer. Chemometric modelling is then used to rapidly assess the fingerprint of the sample being analysed, comparing it against the library spectral data of authentic samples to enable classification (Ross et al., 2020).

The aim of the current work was to develop an industry-compliant pipeline that could enable real-time fish species prediction within minutes (Fig. 1). A total of 1736 fish fillets were included in the training and validation set over a period of seven years, while 432 fish fillets received at the end of that period were used to externally test the obtained models and assess the speed of the developed pipeline in light of future at-line implementation. For chemometric modelling, conventional multivariate LDA and (O)PLS-DA approaches were compared to the use of advanced machine learning (ML) algorithms.

REIMS data has traditionally made use of PCA-LDA modelling, and less frequently, (O)PLS-DA modelling (Gredell et al., 2019). These techniques make use of a fixed series of statistical analysis on data to discern patterns within the spectral data based on underlying class data. Multivariate algorithms are essentially dumb and have no capability to adapt to the data in order to improve classification performance. As a result, they tend to perform best with datasets which start with larger differences. Machine learning techniques are considered intelligent as they have the potential to adapt the underlying mathematic algorithms to achieve best classification performance as more data is used to train the system. This makes ML better suited to highly similar data with few discrete differences (Gredell et al., 2019; Morellos et al., 2016).

Machine learning can additionally remove the requirement for data pre-processing, potentially enabling faster analysis results, as such potentially benefitting the overall efficacy of a REIMS-based fish speciation workflow. We hypothesize that multivariate and/or machine learning-based classification of REIMS spectra can be used for directly amenable rapid industry-proof large-scale fish speciation monitoring.

2. Material and methods

2.1. Chemicals

Leucine-enkephalin was obtained from Waters (Millford, MA, USA) and 2-propanol (LC-MS grade) was supplied by Honeywell Riedel-de Haën (Seelze, Germany). Ultra-pure deionised water (18.2 MΩ/cm) was obtained from a Millipore Milli-Q system (Billerica, MA, USA).

2.2. Samples

Authentic fish samples (comprising fillets, tails and unspecified tissue) were received from several trusted suppliers (Tesco, Matis Iceland, unknown sources) and stored at −20 °C upon arrival. In total, there were 17 different fish species, with 11 white and 6 pink fish species (Table S1). Samples were sourced from both producers and wholesale vendors in the United Kingdom, Ireland, Norway, United States and Iceland over a seven-year time period (2015–2022), to ensure neither geographical origin nor dates of harvest or sampling would unduly influence the resulting modelling. For the validation set, Icelandic salmon was included in addition to other fish from Table S1. Samples were defrosted for approximately 2 h at room temperature prior to analysis and then stored at 4 °C ± 2 °C awaiting analysis, having been stored for 48 to 72 h at −20 °C upon delivery. An attempt was made to analyse frozen samples directly, but since iKnife-REIMS analysis requires conductivity of the sample, it proved necessary to thaw the samples immediately before analysis.

2.3. REIMS data acquisition

Samples were analysed using a Waters G2-XS QToF instrument (Waters, Wilmslow, UK) fitted with a Waters REIMS ion source (Waters, Wilmslow, UK). The REIMS system was operated with the following parameters: negative ionisation mode, cone voltage of 60 V, heater bias of 40 V and a collision cell voltage of 15 V. Data acquisition was performed in sensitivity mode with continuum data acquisition, over a mass range of 100–1,200 m/z, with a scan speed of 2 scans per second. Leucine-enkephalin (0.1 ng/μL) in 2-propanol was infused into the

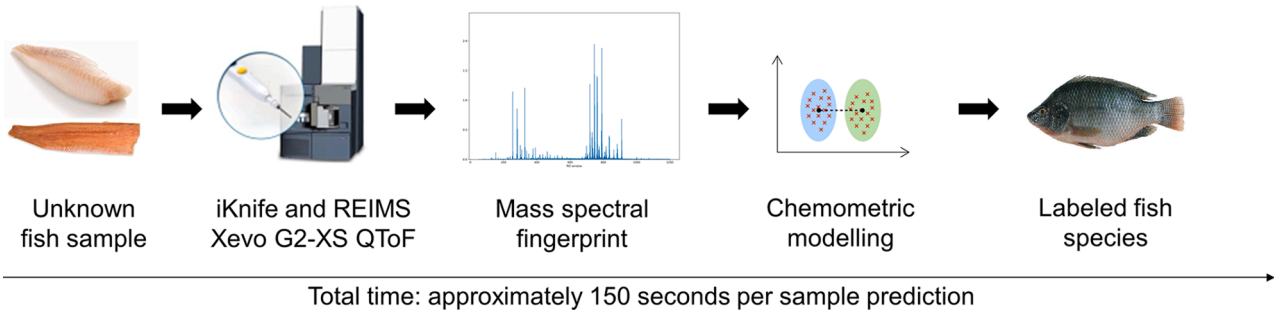


Fig. 1. Workflow for direct prediction of raw REIMS spectra towards industry-proof accurate large-scale fish speciation.

REIMS source at 0.2 mL/min by means of a Waters Acquity I-Class BSM (Waters, Milford, MA, USA) to enable accurate mass correction, whilst the presence of 2-propanol additionally assists ionisation. For priming, non-reoccurring external quality control (QC) samples were analysed prior to the analysis of the actual samples but excluded from further data analysis.

The electrosurgical knife was supplied by Erbe (Erbe Elektromedizin GmbH, Tuebingen, Germany) and connected to the instrument by means of medical grade Tygon 15 mm ID tubing (Saint-Gobain, Solon, OH, USA). The knife was powered by an Erbe VIO50C electrosurgical generator set to 40 W in dry-cut mode. Samples were 'burned' four times using the knife across a representative area of the tissue sample. Each burn lasted approximately 1 s, with each burn evenly separated from the last.

2.4. Chemometric multivariate modelling using PCA-LDA and (O)PLS-DA

REIMS data were acquired using MassLynx v4.2 (SCN 966 & SCN 1010) (Waters, Wilmslow, UK). Data were mass corrected using leucine-enkephalin (554.2615 Da) as a reference ion, followed by background subtraction and total ion count (TIC) normalisation (where each analyte is divided by the TIC for each sample, to reduce sample to sample intensity variation) using the default MassLynx pre-processing algorithms in Abstract Model Builder (AMX) v0.9.2092.0 (Waters Research Centre, Budapest, Hungary). Spectral binning was performed at 0.1 Da. PCA models were generated using AMX, R (version 3.4.3, Vienna, Austria) (Table S2) and SIMCA 17.0 (Sartorius Stedim Biotech, Umea, Sweden) to evaluate instrument stability during sample analysis, identify potential outliers and visualize inherent variation within the dataset prior to supervised analysis (Abdi, 2010). Data were randomized and split into a training and validation set (75/25). Supervised models for "all fish", "white fish" and "pink fish" were generated using AMX for PCA-LDA and SIMCA for PLS-DA and OPLS-DA models, for which data matrices generated in and exported from AMX were used. Specifically, sample matrices were imported into SIMCA using univariate (UV) or Pareto scaling. The obtained PCA-LDA and (O)PLS-DA models from the training set were cross-validated (CV) using a fivefold approach (with a standard deviation of 5 Ω).

2.5. Chemometric machine learning-based modelling using RF and SVM

Both R (version 3.4.3, Vienna, Austria) and Python (version 3.7.4, Fredericksburg, VA) languages were used for data handling and statistical analyses. Data were processed in a virtual computer environment with OS Linux (Ubuntu, v16.04 LTS or v20.04 LTS, Linux) using Oracle VM VirtualBox (version 6.1, Oracle). A comprehensive list of programming languages and packages used is presented in Table S2. Waters raw folders were converted using proprietary software while pre-processing (burn selection, noise removal) was performed using an in-house developed R-based pipeline. Empty (badly acquired) raw files were omitted during this stage. Because of the large number (and size) of data files, all features with an *m/z* value between 100 and 1,200 Da detected in a 10 % weighted randomized sample of the total data (2015–2021) were included in the feature intensity matrix and only the highest burn was retained per REIMS sample spectrum. No spectral binning was performed. Noise removal was based on the distribution of peak intensities (95 % quantile peaks were retained).

Prior to multivariate statistical analysis using random forest (RF) and support vector machine (SVM) algorithms, TIC correction over the MS spectra was performed, and impact of TIC correction (versus no correction) on model predictivities was assessed. Data were randomized and split into a training and validation set (75/25). After splitting, standard scaling (standardizing features by removing the mean and scaling to unit variance) of the data was performed by default. Since no prior research on the best scaling method for this type of data has been published, standard scaling was applied because it is a very commonly used scaling technique that differs little from Pareto scaling (van den Berg et al., 2006). With the training set, ML hyperparameter optimization was performed using fivefold cross-validation, of which applied parameters are listed in Table S3. The remaining part of the training set was used to optimize the hyperparameters based on the accuracy score. At this point, the chosen hyperparameters were fixed and used to predict fish species for the withheld validation set. Variable feature importance rankings were computed from the built-in feature importance (using gini impurity) and correlation coefficients features, for RF and (linear kernel) SVM respectively.

3. Results and discussion

The optimal strategy for the direct sample and data analysis using REIMS depends on the specific classification problem and as such, optimisation of the latter classification strategy per application is of utmost importance (Gredell et al., 2019). Here, five chemometric modelling approaches were compared in the scope of industry-proof fish speciation, for which maximal accuracy, scalability and real-time analysis were the main prerequisites.

3.1. Speciation accuracy using PCA-LDA and (O)PLS-DA

3.1.1. Unsupervised cluster analysis using PCA

PCA modelling demonstrated that PC1, PC2 and PC3 accounted for 66.3 % of the total variability of the REIMS fingerprints (features) with the highest predictive information (Q^2 (cum) = 0.825, with R^2 (cum) = 0.840) achieved from the first 10 PCs (Figure S1). Moreover, an underlying separation between most fish species, particularly between those classified as white and pink (Table S1) could be observed (Fig. 2).

Separation was less obvious though between those species that are phylogenetically closely related (such as cod and coley) and/or can be attributed to comparable taste and flavour attributes (salmon and trout) (Black et al., 2017; Carrera et al., 1999). PCA results also suggest that separation is being driven by discrete differences in chemical profiles of fish species rather than being built on background noise or other factors such as age and storage of samples, bearing in mind that samples were acquired, and analyses run over a time course of seven years, which is to

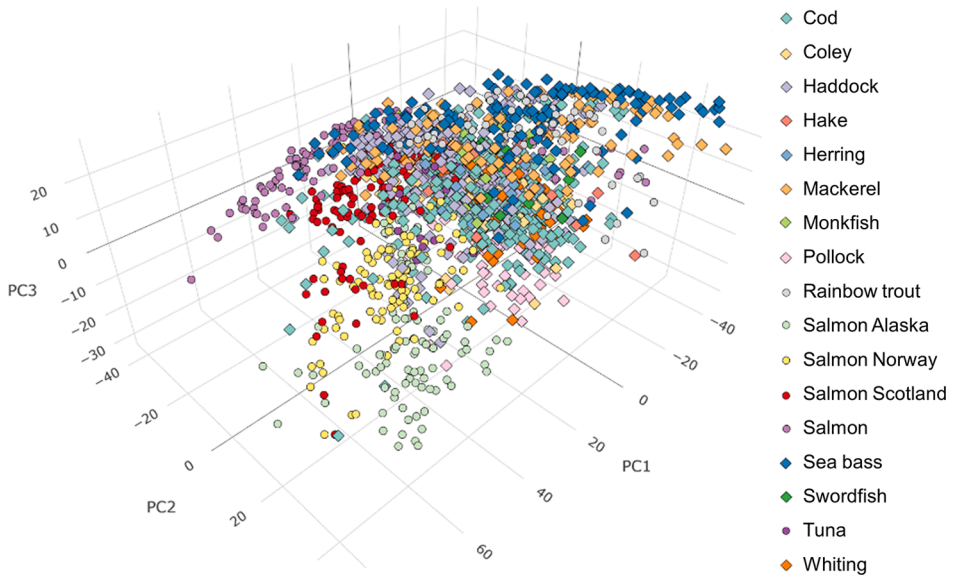


Fig. 2. PCA score plot for all fish samples analysed in the full training and validation data batch (n = 1736) from 2015 to 2021. Diamonds: white fish; circles: pink fish. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the best of our knowledge unprecedented for any food-related AIMS application (Ross et al., 2020). Therefore, we conclude that the instrumental stability is sufficient from an industrial perspective for large-scale fish speciation.

3.1.2. Chemometric modelling using PCA-LDA and (O)PLS-DA

The use of TIC correction (versus no correction) for improvement of model predictivities was assessed and retained because of superior classification accuracy, especially in the all-fish model, which comprised the highest data variability (Table S4). Prior to PCA-LDA modelling, Pareto scaling was performed by default. Prior to (O)PLS-DA modelling, Pareto scaling provided superior performance to UV scaling when assessing $R^2(Y)$ and $R^2(X)$ values for model fit and $Q^2(Y)$ for model predictivity (Table S5).

Except for the all-fish OPLS-DA model, very good to excellent PCA-LDA and (O)-PLS-DA models were obtained with valid R^2X, R^2Y, Q^2 (>0.5), CV-ANOVA p-value ($p < 0.05$) and good permutation testing (n = 100) (Table S6), while accuracies for the latter models ranged between 92.5 and 100.0 % (Table 1). The reported results show the validation parameters of models built using the validation fish batch, ensuring model validity based on untrained data, rendering reliable $Q^2(Y)$ and accuracy. Detailed confusion matrices are provided in Supplementary Tables S7-S15.

On average, the pink fish models demonstrated better accuracy compared to the general all-fish model. Overall, PLS-DA modelling outperformed (O)PLS-DA models, while PCA-LDA generated similar or slightly higher models as compared to PLS-DA as reflected by both the accuracies as well as model validation parameters R^2Y and Q^2Y, which measure goodness-of-fit and predictive ability, respectively. In the model's score plots (Figures S2-S7) significant separation between different classes (species) of fish and very tight clustering within classes (species), could be observed as well. This further provides confidence that the modelling behaviour is an accurate reflection of the ability of these models to correctly predict fish species when challenged with new samples not represented within the models.

The approach of splitting both samples and models into two groups (i.e. white and pink fish) is both easily done visually and highly relevant from an industrial perspective. Moreover, this can have performance benefits – by keeping model size smaller, the time taken to process models when adding new samples or replacing older samples is reduced. From a practical point of view, circumstances, where the use of an all-fish model would potentially be required, are relatively limited, although this would indeed be of added value for the assessment of minced fish products, to determine whether a product that is described as containing only one type of fish has been bulked out with other species, like e.g. minced salmon being bulked with coley (Piredda et al., 2022). The performance of the all-fish model is very high, with classification rates of 97.0 % to 97.9 % which should enable the detection of bulk adulteration products. To be economically viable, adulteration needs to take place at high percentages but in previous work, REIMS was established as a powerful tool to detect adulteration in minced products (Black et al., 2019; Kosek et al., 2019). This level of performance is substantially ahead of current detection technologies such as ELISA and DNA barcoding (Pollack, Kawalek, Williams-Hill, & Hellberg, 2018; Ruethers et al., 2020).

Table 1
Validation results for PCA-LDA and (O)PLS-DA modelling of all-fish (n = 433), white fish (n = 319) and pink fish (n = 125), as obtained from the validation batches.

Model	Model Type	Number of samples	Number of passes	Number of fails	Accuracy
All-fish	PCA-LDA	433	426	10	97.7 %
	OPLS-DA	433	420	13	97.0 %
	PLS-DA	433	424	9	97.9 %
White Fish	PCA-LDA	319	303	16	95.0 %
	OLPS-DA	319	295	24	92.5 %
	PLS-DA	319	301	18	94.4 %
Pink Fish	PCA-LDA	125	125	0	100.0 %
	OPLS-DA	125	124	1	99.2 %
	PLS-DA	125	125	0	100.0 %

3.2. Chemometric modelling using RF and SVM

Prior to ML-based modelling, the number of acquired REIMS features were maximized over the whole data-acquisition period (2015–2022). As such 3602 *m/z* peaks were defined as important features to be implemented in the models. TIC was retained due to superior

classification accuracy (Table S4). Modelling results and obtained accuracies are presented in Table 2. Hyperparameter optimization parameters of RF and SVM algorithms (including different kernels) are given in Table S3 and detailed confusion matrices are provided in Supplementary Tables S16-S21.

For the RF models, four hyperparameters were optimized (Table S3), where 250 trees were selected from the possible maximum of 500 trees per RF for the 'all-fish' model. The individual decision trees were large, 125 layers deep, with a minimum of 3 samples at the leaf node. The selected hyperparameters per model are summarized in Table S22. For the assessed SVM kernels (Table S3), a linear SVM was the best fit throughout as opposed to Gredell et al. (Gredell et al., 2019), where the polynomial kernel outperformed the linear kernel in case the SVM was selected as the best classifier for analysis, providing evidence for their statement that any optimal strategy for REIMS data depends on the classification problem. The SVM C and Gamma were 0.1 and 0.1, respectively, for the all-fish model. The selected hyperparameters per model are summarized in Table S23.

SVM consistently outperformed RF, with a 3.4 % increase on average. Though fewer misclassification errors occurred in the SVM models, the fish species for which the most confusion existed, were the same. Especially for cod, for which samples were gathered from multiple providers, and analysis was performed using different tissues (i.e., tail or neck) and multiple data-acquisition methods (Section 2.2), both false positive and false negative errors were observed (Tables S16-S20). The three different salmon batches, labelled according to their origin (Alaska, Norway, Scotland) also proved relatively challenging as misclassification was observed in the 'all-fish' as well as in the 'pink fish' models. Since no confusion with any of the other fish species was noted, there was no added value in lowering the number of fish species in the models (Tables S16-S18 and S21). Models from both ML algorithms correctly identified monkfish, salmon, sea bass, trout, tuna and mackerel. This can be explained by their unique mass spectral fingerprints (e.g., tuna), but also by the inherently lower variability during sampling and/or data acquisition in one batch (e.g., mackerel) (Fig. 3). While the SVM models were successful in hake and herring recognition, both in the all-fish and white fish models (Tables S19, S20); the identification of these species proved less straightforward in the all-fish versus white fish model using RF (Tables S16, S17).

3.3. Outlook on real-time industry-proof analysis

3.3.1. Real-time performance of multivariate models with a new sample test batch

The multivariate models described above demonstrated to strongly capture fish speciation over an extensive period of time (i.e. over 7 years of data collection), despite the use of different acquisition settings throughout the years (Section 2.2). This proves that by including enough data points, REIMS can overcome inter-batch variability using appropriate multivariate statistics. Adding new timepoints or data is expected to be less straightforward, but what is truly needed to move towards industrial implementation (as opposed to creating models from complete datasets). Therefore, a new external test batch of samples was analysed in 2022 and the raw files from this dataset (n = 444) were used to directly in real-time predict fish species using the models build with the original dataset covering 1736 samples (training and validation data). Analysing new fish species like e.g. Islandic salmon leads to automatic misclassification since this species was not included in the training data. Therefore, correct classification was defined as a prediction of either unknown salmon or salmon from Alaska, Norway or Scotland. The PCA-LDA models for all-fish, white fish, and pink fish enabled to predict fish species in real-time with a very high accuracy of 95.0 % for the all-fish and 99.5 % for the pink fish (Table 3, detailed confusion matrices are provided in Supplementary Tables S24-S26). No direct prediction was however possible with (O)PLS-DA, since the model building can only be executed as a secondary step following pre-processing with AMX. For the ML models, obtained accuracies for the new batch appeared to be relatively low, as summarized in Table 3 (detailed confusion matrices are provided in Tables S27-S32). In the all- and white fish models, most misclassification occurred when other fish species (e.g., haddock) were labelled as cod (Table S27 and Table S28). Cod samples, however, were always predicted correctly, although they did comprise the largest variance in provider, sampling time and instrumental settings. Mislabelling of trout as cod in the all-fish model may however be solved by using the pink fish-specific model. The pink fish models were able to correctly predict fish species with accuracies of 95.1 % and 94.6 % for RF and SVM, respectively (Table 3).

Table 2
Validation results of RF and SVM modelling of all-fish (n = 434), white fish (n = 319) and pink fish (n = 125), as obtained from the validation batches.

Model	Model Type	Number of samples	Number of passes	Number of fails	Accuracy
All-fish	RF	434	384	50	93.6 %
	SVM	434	418	16	96.3 %
White Fish	RF	319	283	36	88.7 %
	SVM	319	302	17	94.7 %
Pink Fish	RF	125	115	10	92.0 %
	SVM	125	117	8	93.6 %

The underperformance of RF and SVM models compared to PCA-LDA was independent of the processing pipeline software used. When PCA-LDA is applied as a classification algorithm in the ML data processing pipeline following the same strategy described in Section 2.5, higher real-time prediction results are obtained compared to RF and SVM for the model for all fish species (Table S33). This implies that the REIMS fingerprints contain linear combinations of features that characterize each fish species, making LDA models better suited to this particular classification problem. We hypothesize that advanced ML algorithms will be better able to understand added layers of complexity, such as e.g. predicting the presence or absence of contaminants in heterogeneous food matrices, rather than a direct classification of species within one type of food (fish). Indeed, RF is said to excel at predicting nonlinear relationships between input characteristics and the target variable (Bengio, 2009). The accuracy of the models can be improved by adding linear combinations of features to the RF and SVM models after a data reduction step that excludes the least significant features (calculated as described in Section 2.5).

3.3.2. Software for real-time fish speciation

The classic REIMS data processing pipeline (Fig. 1) provides considerable flexibility in analysis capabilities and indeed, as demonstrated here, PCA-LDA models generated within AMX can be used directly with the AMX recognition function to provide real-time recognition (45 s/sample). AMX is a powerful software, but its major advantage is its lack of flexibility toward custom pre-processing and splitting train-test data, making the manual creation of new models cumbersome. However, once a model is in place, its application in real-time runs smoothly. The inherent combination of AMX's REIMS-specific data pre-processing (incl. lock mass) and model building capabilities is therefore at present the closest thing to an industry-proof REIMS data analysis pipeline.

Alternatively, the data from AMX can be exported and used to build PLS-DA, OPLS-DA or additional models using third-party software; i.e. by importing the matrices from AMX into e.g. SIMCA 17. With SIMCA real-time analysis is however not feasible since the data needs to be imported and analysed per batch, but it does offer a variety of different data analysis options that are most relevant for metabolomics studies, including e.g. the ability to generate Variable Influence on Projections (VIP) and S-line plots. The latter two outputs can be used to identify the mass bins that contribute most (or least) to the separation of different classes.

The data processing pipeline using ML algorithms provided fast,

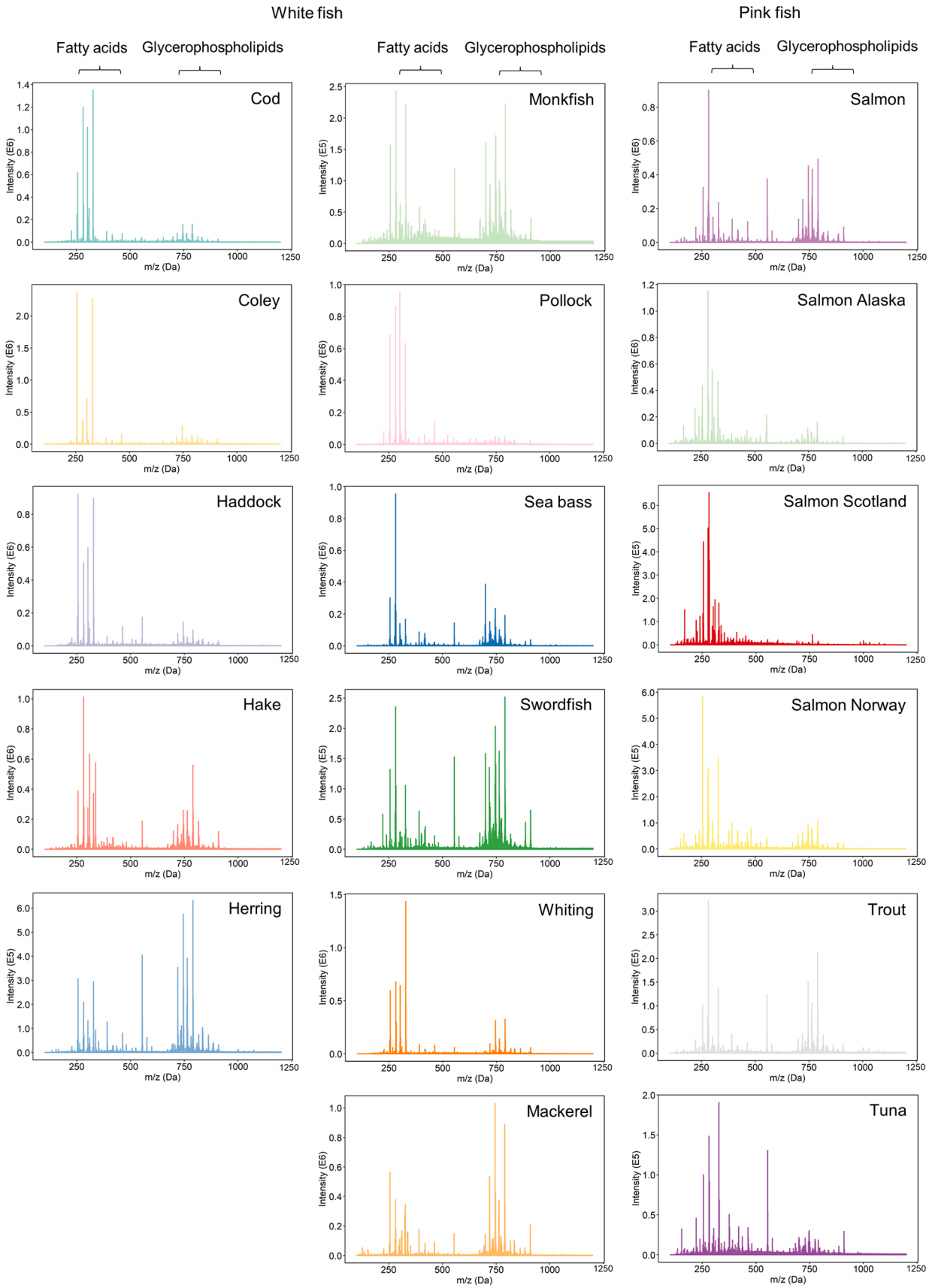


Fig. 3. REIMS mass spectrum per fish species, with molecular class annotation.

Table 3
Real-time prediction results using prior build PCA-LDA, RF and SVM models of all-fish (n = 444), white fish (n = 260) and pink fish (n = 186) for a newly received test batch (2022).

Model	Model Type	Number of samples	Number of passes	Number of fails	Accuracy
All-fish	PCA-LDA	444	422	22	95.0 %
	RF	444	151	293	34.0 %
	SVM	444	236	208	53.2 %
White Fish	PCA-LDA	260	233	27	89.6 %
	RF	260	58	202	22.3 %
	SVM	260	103	157	39.6 %
Pink Fish	PCA-LDA	186	185	1	99.5 %
	RF	184	175	9	95.1 %
	SVM	184	174	10	94.6 %

instantaneous prediction of newly unlabelled samples (150 s/sample), including all sequential steps (file conversion, pre-processing, splitting train-test data, data transformation). Like AMX, real-time classification with these more advanced ML techniques relies on training previously labelled data but can be rolled out more efficiently by including all data processing steps. Unfortunately, the underlying REIMS fish data did not currently allow for sufficiently accurate predictions to move this pipeline to industrial-scale application.

3.4. Underlying molecular features of fish speciation prediction models

REIMS fingerprints obtained from fish tissue are hypothesized to consist of variable fatty acid (100–500 *m/z* range) and glycerophospholipid (500–900 *m/z* range) profiles (Fig. 3) in line with previous findings, where it was observed that the REIMS spectral data of five commercially popular white fish species were indeed dominated by intact phospholipids and fatty acids (Black et al., 2017).

The top 10 most important features (defined as *m/z* values) for each supervised ML model are shown in Supplementary Tables S34-S39. Using AMX and SIMCA, further annotation of the phospholipids – I.e. at the class (e.g. phosphatidylethanolamine) or individual molecular level - is not straightforward, because the features that are most important in fish separation represent mass bins rather than single *m/z* values. For the ML models, it is possible to include the variable feature importance at the *m/z* value level, but this information is still insufficient to annotate the lipid (sub)class or species at the molecular level. The annotation of molecules is best achieved using additional HRMS(/MS) experiments. However, this was beyond the scope of the present work.

3.5. Scalability

Sample analysis using iKnife-REIMS can be performed in less than a minute, rendering the analysis itself to be industry-compliant. An important downside to large-scale application of the classic REIMS data processing pipeline using AMX however, is that data analysis and modelling performance significantly deteriorates when modelling many hundreds to several thousands of samples. The models generated for this study took many hours to build and adding or relabelling just one sample required complete recalculation of each model. In addition, cross-validation of the training set using the fivefold cross-validation approach can take several days to complete. Nevertheless, once models are generated and saved, there is no need for frequent updates. From a company perspective, these could be performed overnight or during weekends, when production stagnates. In addition, such software and hardware-dependent limitations could be easily overcome by increasing computing power, with the result that models could be recalculated with much less effort and more speed (minutes instead of days). The ML-based data processing pipeline does not have this issue of scalability at the current order of magnitude (1000 s samples); i.e. models were created in a few minutes, including hyperparameter optimization, which is highly advantageous for large-scale fish speciation. Automatization of the data processing pipeline would benefit both scalability and real-time classification functionality.

4. Conclusion

The results of this study demonstrate the potential of iKnife-REIMS as an accurate metabolomic fingerprinting tool for fast one-step real-time fish speciation through AMX PCA-LDA modelling. PCA-LDA outperformed (O)PLS-DA, SVM and RF for fish speciation considering the need for high accuracy, scalability, and real-time functionality. PCA-LDA using AMX allows for rapidly analyse of unprocessed fish samples with lower error rates compared to ELISA and with less time- and cost-related drawbacks compared to PCR, MALDI and LC-MS. The technology and industry-compliant data processing pipeline can be implemented to aid in combating fish food fraud but is moreover expected to also be able to help safeguard food safety, quality, and integrity in other high-throughput food production chains. Minor software and hardware optimisations are recommended to reduce the time required to (re)build models, as well as make this step automated.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank Julia Balog (Waters Corporation) for her technical support of the REIMS platform, Saemunder Sveinsson at Matis Iceland for the supply of finfish samples for the project, and Mike Mitchell at Fairseas for his practical insights into the finfish production process.

Funding

This work was supported by EIT Food, the innovation community on Food of the European Institute of Innovation and Technology, a body of the European Union, under Horizon 2020, the EU Framework Programme for Research and Innovation [grant number 20118]. The funding body had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript, and in the decision to submit the article for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2022.134632.

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics, 2*(1), 97–106. https://doi.org/10.1002/wics.51

Anderson, K. A., Hobbie, K. A., & Smith, B. W. (2010). Chemical Profiling with Modeling Differentiates Wild and Farm-Raised Salmon. *Journal of Agricultural and Food Chemistry, 58*(22), 11768–11774. https://doi.org/10.1021/jf102046b

Balog, J., Perenyi, D., Guallar-Hoyas, C., Egri, A., Pringle, S. D., Stead, S., … Takats, Z. (2016). Identification of the Species of Origin for Meat Products by Rapid Evaporative Ionization Mass Spectrometry. *Journal of Agricultural and Food Chemistry, 64*(23), 4793–4800. https://doi.org/10.1021/acs.jafc.6b01041

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends®. Machine Learning, 2*(1), 1–127. https://doi.org/10.1561/2200000006

Birse, N., Chevallier, O., Hrbek, V., Kosek, V., Hajšlová, J., & Elliott, C. (2021). Ambient mass spectrometry as a tool to determine poultry production system history: A comparison of rapid evaporative ionisation mass spectrometry (REIMS) and direct analysis in real time (DART) ambient mass spectrometry platforms. *Food Control, 123*, Article 107740. https://doi.org/10.1016/j.foodcont.2020.107740

Birse, N., McCarron, P., Quinn, B., Fox, K., Chevallier, O., Hong, Y., … Elliott, C. (2022). Authentication of organically grown vegetables by the application of ambient mass spectrometry and inductively coupled plasma (ICP) mass spectrometry. *The leek case study. Food Chemistry, 370*, Article 130851. https://doi.org/10.1016/j.foodchem.2021.130851

Black, C., Chevallier, O. P., Cooper, K. M., Haughey, S. A., Balog, J., Takats, Z., … Cavin, C. (2019). Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry. *Scientific Reports, 9*(1), 6295. https://doi.org/10.1038/s41598-019-42796-5

Black, C., Chevallier, O. P., & Elliott, C. T. (2016). The current and potential applications of Ambient Mass Spectrometry in detecting food fraud. *TrAC Trends in Analytical Chemistry, 82*, 268–278. https://doi.org/10.1016/j.trac.2016.06.005

Black, C., Chevallier, O. P., Haughey, S. A., Balog, J., Stead, S., Pringle, S. D., … Elliott, C. T. (2017). A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. *Metabolomics, 13*(12), 153. https://doi.org/10.1007/s11306-017-1291-y

Carrera, E., Garcia, T., Cespedes, A., Gonzalez, I., Fernandez, A., Hernandez, P. E., & Martin, R. (1999). Salmon and Trout Analysis by PCR-RFLP for Identity Authentication. *Journal of Food Science, 64*(3), 410–413. https://doi.org/10.1111/j.1365-2621.1999.tb15053.x

Cutarelli, A., Amoroso, M. G., De Roma, A., Girardi, S., Galiero, G., Guarino, A., & Corrado, F. (2014). Italian market fish species identification and commercial frauds revealing by DNA sequencing. *Food Control, 37*, 46–50. https://doi.org/10.1016/j.foodcont.2013.08.009

Fao. (2020). The State of World Fisheries and Aquaculture 2020. *FAO*. https://doi.org/10.4060/ca9229en

Fox, M., Mitchell, M., Dean, M., Elliott, C., & Campbell, K. (2018). The seafood supply chain from a fraudulent perspective. *Food Security, 10*(4), 939–963. https://doi.org/10.1007/s12571-018-0826-z

Gredell, D. A., Schroeder, A. R., Belk, K. E., Broeckling, C. D., Heuberger, A. L., Kim, S.-Y., … Prenni, J. E. (2019). Comparison of Machine Learning Algorithms for Predictive Modeling of Beef Attributes Using Rapid Evaporative Ionization Mass Spectrometry (REIMS) Data. *Scientific Reports, 1*, 1. https://doi.org/10.1038/s41598-019-40927-6

Kosek, V., Uttl, L., Jírů, M., Black, C., Chevallier, O., Tomaniová, M., … Hajšlová, J. (2019). Ambient mass spectrometry based on REIMS for the rapid detection of adulteration of minced meats by the use of a range of additives. *Food Control, 104*, 50–56. https://doi.org/10.1016/j.foodcont.2018.10.029

Lasch, P., Uhlig, S., Uhlig, C., Wilhelm, C., Bergmann, N., & Wittke, S. (2019). Development and In-House Validation of an LC–MS and LC–MS/MS Assay for the Determination of Food Fraud for Different Fish Species. *Journal of AOAC INTERNATIONAL, 102*(5), 1330–1338. https://doi.org/10.1093/jaoac/102.5.1330

Lima, M., Canales, T. M., Wiff, R., & Montero, J. (2020). The Interaction Between Stock Dynamics, Fishing and Climate Caused the Collapse of the Jack Mackerel Stock at Humboldt Current Ecosystem. *Frontiers in Marine Science, 7*. https://doi.org/10.3389/fmars.2020.00123

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., … Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering, 152*, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018

Navotas, I. C., Santos, C. N. V., Balderrama, E. J. M., Candido, F. E. B., Villacanas, A. J. E., & Velasco, J. S. (2018). Fish identification and freshness classification through image processing using artificial neural network. *ARPN Journal of Engineering and Applied Sciences, 13*(18).

Phelps, D. L., Balog, J., Gildea, L. F., Bodai, Z., Savage, A., El-Bahrawy, M., … Ghaem-Maghami, S. (2018). The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). *British Journal of Cancer, 118*(10), 1349–1358. https://doi.org/10.1038/s41416-018-0048-3

Piredda, R., Mottola, A., Cipriano, G., Carlucci, R., Ciccarese, G., & Di Pinto, A. (2022). Next Generation Sequencing (NGS) approach applied to species identification in mixed processed seafood products. *Food Control, 133*, Article 108590. https://doi.org/10.1016/j.foodcont.2021.108590

Pollack, S. J., Kawalek, M. D., Williams-Hill, D. M., & Hellberg, R. S. (2018). Evaluation of DNA barcoding methodologies for the identification of fish species in cooked products. *Food Control, 84*, 297–304. https://doi.org/10.1016/j.foodcont.2017.08.013

Rasmussen, R. S., & Morrissey, M. T. (2008). DNA-Based Methods for the Identification of Commercial Fish and Seafood Species. *Comprehensive Reviews in Food Science and Food Safety, 7*(3), 280–295. https://doi.org/10.1111/j.1541-4337.2008.00046.x

Ross, A., Brunius, C., Chevallier, O., Dervilly, G., Elliott, C., Guitton, Y., … Vanhaecke, L. (2020). Making complex measurements of meat composition fast: Application of rapid evaporative ionisation mass spectrometry to measuring meat quality and fraud. *Meat Science, 108333*. https://doi.org/10.1016/j.meatsci.2020.108333

Ruethers, T., Taki, A. C., Khangurha, J., Roberts, J., Buddhadasa, S., Clarke, D., … Koeberl, M. (2020). Commercial fish ELISA kits have a limited capacity to detect different fish species and their products. *Journal of the Science of Food and Agriculture, 100*(12), 4353–4363. https://doi.org/10.1002/jsfa.10451

Schäfer, K.-C., Dénes, J., Albrecht, K., Szaniszló, T., Balog, J., Skoumal, R., … Takáts, Z. (2009). In vivo, in situ tissue analysis using rapid evaporative ionization mass spectrometry. *Angewandte Chemie (International Ed.In English), 48*(44), 8240–8242. https://doi.org/10.1002/anie.200902546

Stahl, A., & Schröder, U. (2017). Development of a MALDI–TOF MS-Based Protein Fingerprint Database of Common Food Fish Allowing Fast and Reliable Identification of Fraud and Substitution. *Journal of Agricultural and Food Chemistry, 65*(34), 7519–7527. https://doi.org/10.1021/acs.jafc.7b02826

Trotta, M., Schönhuth, S., Pepe, T., Cortesi, M. L., Puyet, A., & Bautista, J. M. (2005). Multiplex PCR Method for Use in Real-Time PCR for Identification of Fish Fillets from Grouper (Epinephelus and Mycteroperca Species) and Common Substitute Species. *Journal of Agricultural and Food Chemistry, 53*(6), 2039–2045. https://doi.org/10.1021/jf048542d

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics, 7*(1), 1471–2164. https://doi.org/10.1186/1471-2164-7-142

Varrà, M. O., Ghidini, S., Husáková, L., Ianieri, A., & Zanardi, E. (2021). Advances in Troubleshooting Fish and Seafood Authentication by Inorganic Elemental Composition. *Foods, 10*(2), 270. https://doi.org/10.3390/foods10020270