European Food Research and Technology (2026) 252:149
https://doi.org/10.1007/s00217-026-05062-3

ORIGINAL PAPER

Species identification of shrimps, lobsters and crayfish using peptide-based liquid chromatography low-resolution mass spectrometry and random forest model

Tim Roggensack[1] · Christian Brenn[1] · Maik Döring[2] · Wolfgang Jira[3] · Christian Treitz[4] · Reinhold Hanel[5] · Maria Blažina[6] · Ahmed Yahyaoui[7] · Andreas Tholey[4] · Ingrid Clawin-Rädecker[1] · Ute Schröder[1]

Received: 8 August 2025 / Revised: 17 December 2025 / Accepted: 16 January 2026 / Published online: 20 February 2026

Abstract

Seafood, including crustaceans, provides a high-quality animal-origin protein source with globally increasing consumption rates. However, international seafood trade is frequently confronted with issues like commercial fraud and safety risks for human health, partly due to incorrect species identification of aquatic food products. To distinguish between crustacean species through the identification of species-specific peptides, an untargeted liquid chromatography low-resolution mass spectrometry (LC-LRMS) method was developed and validated. Programmed algorithms were applied to identify marker candidates from peptide profiles. A targeted multiple reaction monitoring (MRM) method was established to verify the suitability of selected candidates as crustacean biomarkers related to the used dataset. An additional random forest model was built to determine unknown crustacean species based on an LC-LRMS raw data training set. In total, 49 out of 150 selected peptides were identified as species-specific biomarkers based on the monitored dataset and those are able to differentiate 14 crustacean species. To further evaluate specificity, additional commercial samples from several crustacean, mussel, insect and fish species were tested. *De novo* sequencing of LC-LRMS and comparing high-resolution mass spectrometry (HRMS) data mostly showed similar results concerning the proposed amino acid sequence and average local confidence score. Selected peptide markers were synthesized and experimentally confirmed. The random forest model exemplarily demonstrated the correct identification in an unseen test dataset based on plurality vote. The results show the feasibility of solving authenticity questions, e.g. to identify unknown crustacean species, by applying alternative procedures without using HRMS instruments, requiring a higher effort. At the same time, the results prove that certain limitations cannot be overcome using LRMS devices.

Keywords Food authenticity · Species identification · Proteomics · Crustaceans · Liquid chromatography mass spectrometry · Random forest

✉ Ute Schröder
ute.schroeder@mri.bund.de

[1] Department of Safety and Quality of Milk and Fish, Max Rubner-Institut, Kiel, Germany

[2] National Reference Center for Authentic Food, Max Rubner-Institut, Kulmbach, Germany

[3] Department of Safety and Quality of Meat, Max Rubner-Institut, Kulmbach, Germany

[4] Institute for Experimental Medicine, Christian-Albrechts-University, Kiel, Germany

[5] Thünen Institute of Fisheries Ecology, Bremerhaven, Germany

[6] Center for Marine Research, Ruđer Bošković Institute, Rovinj, Croatia

[7] Faculty of Sciences, Mohammed V University, Rabat, Morocco

Introduction

Seafood, including crustaceans, provides a highly demanded and essential animal originated protein source, which might become even more relevant in the future [1, 2]. Crustaceans are globally consumed on a large scale due to their nutritional value [3], originating from the combination of high-quality protein content with polyunsaturated fatty acids, vitamins and minerals [2, 4]. Besides high-quality nutritional properties, crustaceans are popular due to their texture and taste [2]. The greater demand for food and the ongoing globalisation of trade increased the adulteration risks alongside more complex food supply chains [5].

Due to the high phenotypic similarity, the unambiguous species identification of crustaceans is a challenging issue for the fishery industry and control authorities to fulfill today´s requirements concerning labeling and traceability [5, 6]. By means of identification methods on a molecular level, illegal or undocumented fishery could be better detected, and therefore fish and shellfish stocks be protected [7]. However, the purpose of crustacean authenticity is not only to prevent commercial fraud but also to protect consumer health [6].

The development of reliable analytical tools for species identification is mandatory to achieve this goal. Besides deoxyribonucleic (DNA)-based [8] methods, mass spectrometry (MS) applications could also be used for this purpose. In the meantime, multiple MS-based methods are available for food authentication [9]. For certain species, choosing an MS approach (e.g., proteomics) may even be advantageous compared to DNA-based methods, particularly when DNA analysis is not suitable [10] or a confirmation of DNA results is desirable. For instance, it is challenging to differentiate closely related tuna species of the *Thunnus* genus [11], especially when they are canned, and distinguishing domestic pig from wild boar [12] is also difficult using DNA-based techniques. Moreover, even when using a quadruplex real-time polymerase chain reaction (PCR) approach targeting four species, it becomes increasingly difficult to design a balanced primer-probe system, and the risk of cross-reactions might also rise [13]. For food control laboratories, DNA meta-barcoding is becoming an increasingly considered analytical approach because it enables to characterize the composition of complex seafood products in a comprehensive untargeted manner across a broad taxonomy [14–16]. However, this approach requires the selection of universal primer sets, optimised analytical protocols and a trustworthy sequence database. Due to the high specificity of mass transitions applying a targeted multiple reaction monitoring (MRM) method, a larger number of species can also be simultaneously detected using MS – even with, low-resolution mass spectrometry (LRMS).

The current literature concerning food authenticity using organic-based MS methods (including proteomics) dealt with species identification and differentiation, certain ingredients, production methods, or geographic origin of different seafood [9]. In the field of crustacean species authentication, Ortea et al. [17] also used a liquid chromatography (LC)-LRMS proteomics approach based on species-specific peptides to distinguish seven closely related shrimp species. Proteomics-based LC-MS methods for distinguishing three shrimp species were also developed by Hu et al. [18], who applied LC-HRMS (high-resolution mass spectrometry) for biomarker identification and targeted LC-LRMS MRM for biomarker detection, with biomarker selection performed by chemometric analysis. Chatterjee et al. [19] combined untargeted LC-HRMS for biomarker identification with targeted LC-LRMS for detecting species-specific metabolomics analytes, enabling the identification of five shrimp species. Furthermore, various MS techniques without chromatographic separation were employed for crustacean species identification. Salla and Murray [20] created a reference database of mass spectra to distinguish six shrimp species using matrix-assisted laser desorption ionization (MALDI) MS combined with mass spectral fingerprint matching. As an in-situ real-time application, Lu et al. [21] introduced a rapid evaporative ionization mass spectrometry (REIMS) approach for identifying seven shrimp species in minced form, based on lipidomic analytes. However, independent to MS resolution or analytical technique, currently available publications exclusively focused on the verification of shrimp species. No recently published LC-MS based studies could be found that targeted species identification of lobsters, crabs, and crayfish [9].

Considering further MS based seafood authentication investigating mollusc or fish species, more publications are available using proteomic workflows. For instance, Gu et al. [22] developed a proteomic approach including a biomarker peptide for quantification to detect the substitution of Atlantic salmon by cheaper rainbow trout and Li et al. [23] identified signature peptides to prevent the adulteration of oyster powder. Both articles used the combination of an untargeted and targeted LC-MS based proteomics workflow, but still the identification of marker peptides was performed using HRMS. Due to its high mass accuracy, HRMS data is preferred for untargeted species identification [19, 24]. Despite the growing number of HRMS publications in recent years, this study employed the more cost-effective LRMS as an alternative [24]. Regarding the authenticity of seafood species, meat or other protein-based foods e.g. eggs, no proteomic-based literature sources were found, which used LRMS for peptide marker identification, which is the advance of innovation of this study.

This article aimed at expanding the diversity of crustaceans for species identification with selected lobster, crab, and crayfish species besides shrimp species with an LC-MS based application. A further methodic aim was to prove the feasibility of untargeted and targeted LRMS data in combination with a statistical tool to solve species differentiation issues including the *de novo* sequencing of crustacean peptide markers. To achieve this goal using predominantly LRMS data, a stronger focus was placed on the bioinformatic side. On the basis of programmed algorithms, no database entries or any kind of peptide spectral library are necessary to identify species-specific peptide markers inside the given data, which is an advantage of this approach. In this study, those algorithms were able to simultaneously identify peptide marker candidates for up to eight different crustacean species. The number of species included in biomarker identification was further expanded by evaluating two biological groups. Furthermore, an extensive manual specificity validation was performed with the aid of various marine and insect samples belonging to 23 crustacean species (genus level for *Metapenaeus* species) and two mussel, insect and fish species each. Additional HRMS data were measured for data comparison. To achieve the objectives, proteomics-based analytical methods and statistical applications were developed and validated for the reliable identification of selected crustacean species with LRMS data.

Material and methods

Figure 1 shows for the example of crustaceans the different stages of the analytical and statistical approaches based on untargeted and targeted data. For LC-MS the milestones were coloured in light red and for the random forest model in light blue. The bullet points of the workflow were chronological ordered as sequential steps.

Sample material

For method development and validation of the untargeted LC-LRMS, nine commercial crustacean specimens of four different species (*Nephrops norvegicus*, *Homarus americanus*, *Penaeus monodon*, *Penaeus vannamei*) were purchased at a local market or retail store. For the identification of possible species-specific peptides and their validation with targeted LC-LRMS, 71 raw reference samples were

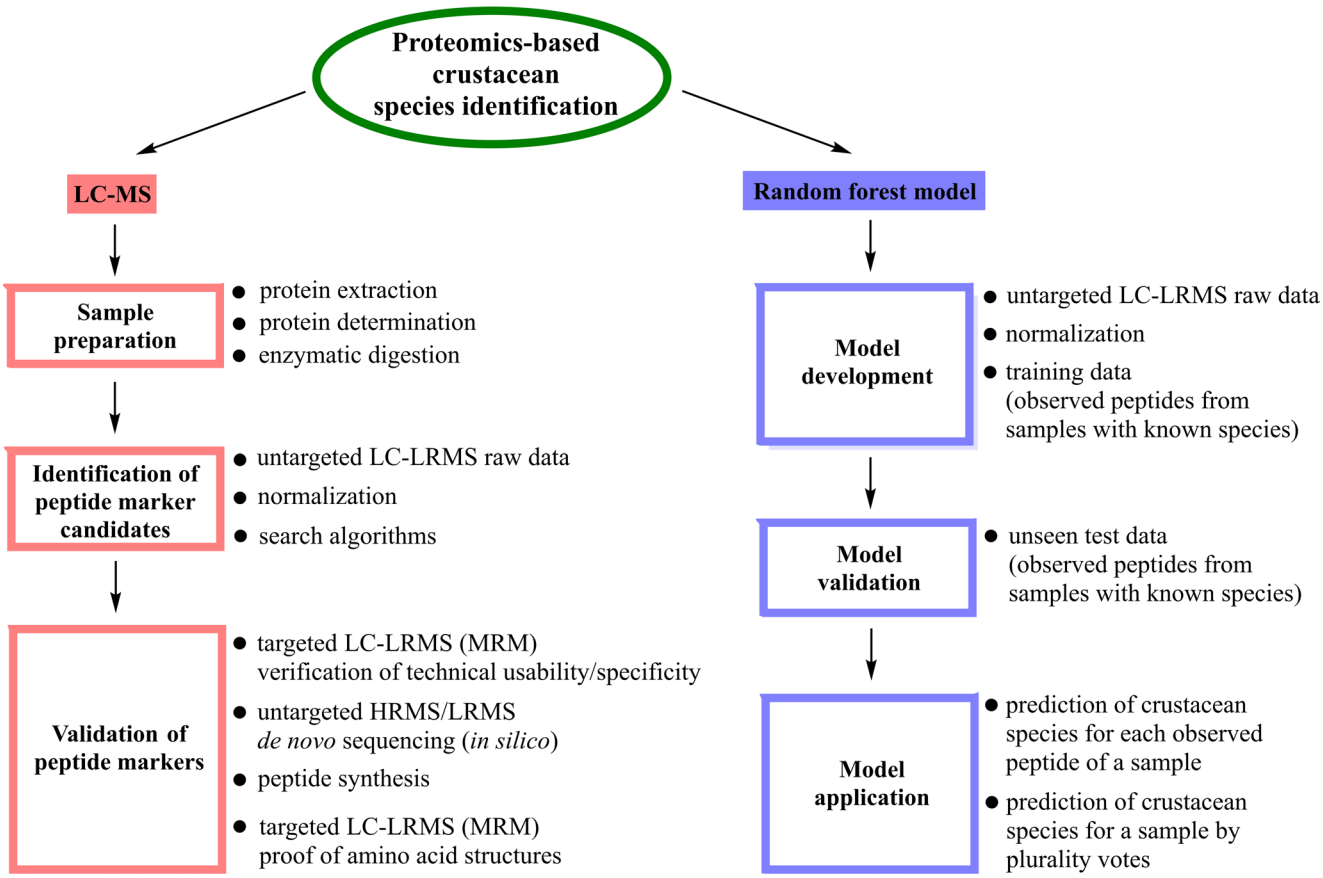

Fig. 1 Alternative proteomics study design combined with the workflow at each stage for the species identification using predominantly LC-LRMS data

used which belonged to 15 different crustacean species (crab, crayfish, lobster, shrimp) verified by DNA analysis [7]. To increase the number of species using LC-LRMS, the reference sample material was further divided into group A and B. Group A consisted of eight shrimp species (*Aristeus antennatus*, *Crangon allmannii*, *Crangon crangon*, *Penaeus vannamei*, *Melicertus kerathurus*, *Parapenaeus longirostris*, *Plesionika martia*, *Squilla mantis*) whereas group B was composed of one crab (*Cancer pagurus*), three crayfish (*Astacus astacus*, *Pacifastacus leniusculus*, *Pontastacus leptodactylus*) and three lobster (*Homarus americanus*, *Homarus gammarus*, *Nephrops norvegicus*) species. Additional 45 commercial crustacean samples including the same and further crustacean species (*Metapenaeus affinis*, *Metapenaeus dobsoni*, *Metapenaeus ensis*, *Metapenaeus monoceros*, *Pandalus borealis*, *Penaeus monodon*, *Penaeus stylirostris*, *Penaeus vannamei*, *Pleoticus muelleri*, *Squilla mantis*) and six individuals of two blue mussel species (*Mytilus edulis*, *Mytilus trossulus*), six insect specimens of two insect species (*Schistocerca gregaria*, *Zophobas morio*) as well as two fish samples (*Salmo salar*, *Thunnus albacares*) were used for a further validation of identified peptide marker candidates. Some of the commercial samples were already processed at the time of purchase and several muscle tissues were cooked in the laboratory for validation issues. Further information about the sample material is provided in supplement Table 1.

Preparations before LC-MS

Sample preparation

A few whole purchased crustacean trade samples were pooled on ice by removing the organs and shell parts. The reference sample material and most of the trade samples were already provided as muscle tissue. Before analysis most of the tissues and extracts for method development and validation were stored at -20 °C, whereas reference sample materials were frozen at -80 °C until further use.

Protein extraction

Approximately 250 mg of muscle tissue was employed for the protein extraction in 2.5 ml extraction buffer (tris(hydroxymethyl)-aminomethane (TRIS) 200 mM, 2 M NaCl, pH 9.0) according to Korte et al. [25]. Subsequently, the sample homogenisation (Bead Ruptor Elite, OMNI International, Kennesaw, Georgia, United States) was applied by 3 cycles of 15 s homogenisation with 15 ceramic balls and 4 m/s speed. Between the extraction cycles, the samples were cooled down on ice bath for 45 s. After centrifugation (13000 x g; 15 min; 4 °C) the supernatant was split into two parts. One proportion of the protein extract was immediately frozen mainly at -80 °C, whereas the second part was previously heat denatured for 4 min in boiling water.

Protein determination

After extraction, the total protein content of the native part from the centrifuged supernatant was determined with nanodrop microvolume spectrophotometry at 280 nm (3 µl; Lid-factor: 10) for screening and with Bradford assay for accurate quantification. For Bradford, the binding of the protein to the dye changes the maximum of absorption from 465 nm to 595 nm [26]. Both wavelengths were measured but the value of 595 nm was used for quantification. To define the protein content of the unknown samples a bovine serum albumin standard as an external calibration (nanodrop: 5 mg/ml-25 mg/ml; Bradford: 0 mg/ml-1.4 mg/ml) was generated at every measurement.

Tryptic digestion (in solution)

For LC-MS analysis, denatured protein extracts were digested with dimethylated trypsin from porcine pancreas (Sigma Aldrich, St. Louis, Missouri, United States) according to von Oesen et al. [27] with modifications. Previously, the extracts were diluted with urea reagent (6 M urea, 100 mM TRIS) depending on their protein content. The solution was reduced with 200 mM dithiothreitol and carbamidomethylated with 200 mM iodoacetamide before tryptic digestion for 16 h.

LC-MS measurements

Untargeted low-resolution MS (LC-LRMS)

A nano-high performance liquid chromatography (nHPLC) separation of tryptic peptides was carried out on an UltiMate 3000 (Thermo Scientific, Germering, Germany) with an analytical Pep swift monolithic column (200 µm x 25 cm; Thermo Fisher Scientific, San José, California, United States). An additional precolumn Pep swift monolithic trap (200 µm x 5 mm; Thermo Fisher Scientific, San José, California, United States) was used in trap mode to bind the analytes and purify them from matrix components [28]. Loading of the precolumn was run in backflush mode for 2 min with a flowrate of 5 µl/min (total volume 10 µl). The loading buffer contained 2% acetonitrile and 0.1% heptafluorobutyric acid (HFBA). After valve switching and 1 min break the gradient elution started with 1% of eluent B which contained 80% acetonitrile and 0.08% formic acid. The aqueous eluent A consisted of 0.1% formic acid. During linear gradient the proportion of eluent B increases until

50% within 60 min. During the rinse phase the eluent B enhanced to 90% within 2 min and maintained at 90% for 5 min, followed by an equilibration of barely 8 min returning to 1% eluent B and resulting in an analysis time of 75 min. The sample injection volume was 1 µl and a flow rate of 1 µl/min was used during separation. The peptide solutions were acidified with trifluoro acetic acid (TFA; concentration about 0.1%) before measurement. LRMS analysis was performed on LTQ XL (Thermo Fisher Scientific, San José, California, United States) instrument with electrospray ionization (ESI) in the positive mode. The mass spectrometer was run with a source voltage of 2 kV and the transfer capillary was set to 200 °C. For the identification of peptide marker candidates, the mass spectrometer was operated in the data-dependent mode. In full scan mode, MS spectra between m/z = 220–2000 were acquired. The three most intense ions at a timepoint captured from the ion trap were isolated and fragmented with collision induced dissociation. The normalized collision energy was set to 35.0 eV. All LC-LRMS measurements were performed with three technical replicates.

Untargeted high-resolution MS (LC-HRMS)

Untargeted HRMS peptide profiles were additionally measured to evaluate the developed search algorithm for peptide marker identification and also to compare LRMS results of *de novo* sequencing (in silico). Prior to the LC-HRMS measurement, the peptide hydrolysates were purified with ZipTip material (Thermo Fisher Scientific, Rockford, Illinois, United States) by slightly modified standardized protocol. For chromatography an ultra-high performance liquid chromatography (uHPLC; Dionex U3000) was used. In back-flush mode with the precolumn C18 PepMap 100 (300 µm x 5 mm), samples were trapped and desalted with a flowrate of 30 µl/min for 2 min. Afterwards, the separation was performed on an analytical Acclaim PepMap RSLC (75 µm x 50 cm) column with a flowrate of 0.3 µl/min. Sample volume of 1 µl was injected and eluted over a gradient. Each sample was analysed with two different MS instrument setups to either focus on the acquisition of high-resolution MS1 scans or acquire a high number of MS2 identifications. Details of the gradient and the MS detection parameter are stated in supplement Tables 9, 10.

Targeted multiple reaction monitoring (MRM) method (LC-LRMS)

In total, 49 species-specific peptide markers of 14 different crustacean species with available reference sample material were acquired simultaneously with a developed targeted LC-LRMS MRM method. This method was used to verify the suitability of selected peptide marker candidates as a crustacean biomarker related to the applied dataset. The HPLC separation of tryptic peptides was carried out on a 1290 Infinity II (Agilent Technologies, Waldbronn, Germany) with a Hypersil GOLD analytical column (5 µm; 2.1 mm x 15 cm; Thermo Scientific Baltics, Vilnius, Lithuania) and a Hypersil GOLD precolumn (5 µm; 2.1 mm x 10 mm; Thermo Scientific Baltics, Vilnius, Lithuania) in front. The aqueous eluent A consisted of 0.1% formic acid whereas eluent B contained 80% acetonitrile and 0.08% formic acid. The linear gradient elution in guard mode started after 1 min and the proportion of eluent B increased from 1% until 70% within 11.5 min. During rinse phase the eluent B immediately enhanced to 90% and maintained at 90% for almost 1 min, followed by an equilibration of approximately 2.5 min returning to 1% eluent B, resulting in an LC-LRMS MRM method run time of 15 min. Sample injection volume was 10 µl and a flow rate of 1 ml/min was applied. The peptide solutions were acidified with TFA (concentration about 0.1%) before measurement. LRMS analyses were performed on a QTRAP 6500^{+} (Sciex Germany GmbH, Darmstadt, Germany) instrument with ESI in the positive mode. MS spectra of three specific precursor fragment mass transitions were acquired for each peptide marker or marker candidate. The collision energy (CE) was individually optimized for each mass transition. The dwell time was increased though the use of measuring time windows of 60 s for each peptide and is dependent on the frequency of signals in a certain section of the gradient. All LC-MS measurements were performed with two technical replications.

Method validation

Untargeted low-resolution MS (LC-LRMS)

The untargeted LC-LRMS method was validated by the parameters of recovery, matrix effects, linearity, precision, stability and reproducibility. Further information for all validation parameters is stated in the supplement Tables 2, 3, 4, 5, 6, 7 and 8.

Recovery (as trapping efficiency) For recovery, the loss of the online-linked purification on the precolumn (trap mode) was determined in comparison to a guard measurement due to the fact that no further purification was performed after the modification of analytes during tryptic digestion. The mean peak values of three technical replicates of 10 selected peptides from a β-casein protein standard (validation peptides 1–10) were compared between the trap and guard measurement mode (Supplement Fig. 1). The statistical parameter of the standard deviation (SD) and the coefficient of variation (CV) were given for each peptide. Further information

about those peptides from β-casein protein were previously provided by Altmann et al. [29].

Matrix effects The influence of the biological crustacean matrix was determined by comparing the mean peak values of three technical replicates of 10 selected peptides from a pure β-casein protein standard (validation peptides 1–10) in eluent A to the same protein standard in crustacean matrix of *H. americanus*. The statistical parameters of SD and CV were given for each peptide.

Linearity The linearity was measured for the β-casein protein standard (validation peptides 1–10) and furthermore for a lobster (*H. americanus*) validation sample (validation peptides 11–20) in four different concentrations whereas the other component was kept constant to exclude a matrix dependence. The different concentration levels were measured with one technical replicate. The strength of linear correlation was given by the coefficient of determination R^2.

Precision (daily precision / precision on different days) The daily precision of the instrument was measured by consecutive technical replicates ($n = 10$) from one biological replicate from the lobster validation sample for 10 selected peptides (validation peptides 11–20) within approximately 12.5 h. For the precision at different days the same validation peptides 11–20 were determined with three technical replicates each included in 10 different measurements distributed to various measuring days in a time period of approximately three weeks. For those measurements the same lobster validation sample was used as a pooled mix derived from three independent protein extracts, which were digested three times each. The total of nine peptide hydrolysates were merged together after enzymatic digestion to generate this standard. The peptide mixture was used as a quality control to monitor the long-term conditions of untargeted LC-LRMS measurements. The mean values of the replicate peak values were given with the statistical parameter CV for each peptide and both precision evaluations. For the precision on different days mean values of the three technical replicates were formed initially which were summarized to an overall mean value of the total amount of technical replicates ($n = 30$).

Stability The sample stability and the effect of storage conditions were determined by measuring 10 selected peptides from the lobster validation sample (validation peptides 11–20) in dependence of storage-duration, -temperature, and a previous heat denaturation of the protein extract. For all stability experiments, peptide digests as ready to measure analytes were stored under various conditions. In detail, three biological replicates ($n = 3$) with one technical replicate each from the same sample were determined under different storage conditions. Samples were measured after three hours at elevated room temperature (approximately 20–25 °C) and after one week of storage in a refrigerator at 6–7 °C. Those samples were kept in a freezer at -20 °C until measurement. Further samples were measured after two weeks in the freezer and three thawing-freezing cycles (approximately 20–25 °C) within one week. For long-term storage another sample set was stored at -20 °C for 10 months. All parameters were determined with and without a denaturation of the protein extract before enzymatic digestion and subsequent storage. The mean values of the replicates were compared to reference values of samples stored at -80 °C, which were determined during the same measurement with the same denaturation status.

Reproducibility As an additional validation parameter, the comparability of peptide spectra and intensities was monitored. In this analysis, the biological variability of samples [30] also influenced the results besides sample preparation and instrument performance. For this purpose, 10 biological replicates from six purchased shrimp specimens were examined, which were verified as *P. vannamei* by DNA-analysis [7]. For intraindividual differences muscle tissue from five different parts derived from one specimen were taken and individually investigated as five biological replicates ($n = 5$). For the interindividual differences one biological replicate ($n = 5$) from each of the five further shrimp validation samples was used. To monitor the reproducibility of peptide spectra as a qualitative approach, three merged technical replicates of each biological replicate were submitted to the basic search algorithm explained below. The data sets were tested for individual occurring peptides in the biological replicate samples portraying the intra- or interindividual differences and furthermore the combination of both effects was also studied. As a quantitative validation of reproducibility, the peak intensities from 10 selected peptides (validation peptides 21–30) from shrimp validation samples were determined. In detail, at first the normalized peak values of three technical replicates were summarized as a mean for each biological replicate ($n = 10$) and afterwards the means of all biological replicates were combined to receive an overall mean value of the total amount of replicates ($n = 30$) for the 10 validation peptides given with the statistical parameter CV. Exemplarily for validation peptide 29, the quantitative reproducibility was further evaluated by split-

ting the intra- and interindividual factors into two separate graphs consisting of five biological replicates ($n = 5$) each.

Targeted low-resolution MRM (LC-LRMS)

The targeted MRM LC-LRMS method was validated with regard to the technical suitability and specificity of peptide marker candidates. Further information for both parameters is provided in the supplement section.

Technical usability At first, the three most intense fragments were selected for the precursors of those peptides, which were identified as a potential biological marker by the search algorithm filled with untargeted LC-LRMS data of the biological groups A or B. For the fragment selection, isotopes and known adducts were avoided. For the first validation step, the mass transitions were proved towards a suitable signal with the established acquisition parameters of the targeted MRM method. The fragment ion intensities of suitable peptides, which were also specific inside their biological group (validation step 2), were further optimized by the CE (validation level 3). Starting from a CE-value of 20, three transitions of each peptide marker candidate were individually monitored with a CE-value of 10 and 30. Further information on the individual marker candidates and examples of the CE optimization is provided in the supplement Tables 16, 17 and supplement Fig. 6a, b.

Specificity Suitable peptide marker candidates (after validation level 1 respectively 3) were further validated towards species specificity with the aid of the Analyst software (version 1.7.2) based on the acquired dataset. The peptide signals of non-target species were compared to the target species reference material in terms of retention times, number and order of the mass transitions, and peak intensities. The exact procedure for the exclusion of peptide marker candidates are given in the supplement Table 18. In a first step (validation level 2), the specificity validation was related to the specificity inside the biological group A or B. An expanded specificity testing (validation level 4 and 5; Table 1) included the other biological group, trade crustacean species, and additional organisms. For the second

Table 1 Number of peptide marker candidates per crustacean species for the biological groups A and B after testing different validation criteria

Crustacean species	Technical suitability (MRM-measuring)	Specificity (biological group)	CE 10 20 30	Specificity (trade samples) crustaceans	Specificity (trade samples) mussels/ insects/fish	Suitability peptide-synthesis	
						LRMS	HRMS
Aristeus antennatus	9	8	7	3	3	1	2
Crangon almannii	7	1	1	1	1	0	1
Crangon crangon	7	6	6	4	4	0	4
Melicertus kerathurus	9	8	7	4	4	1	3
Parapenaeus longirostris	8	7	7	3	3	0	1
Penaeus vannamei	5	4	4	0	0	0	0
Plesionika martia	8	8	7	5	5	3	4
Squilla mantis	10	10	9	5	4	2	3
Sum shrimp species	63	52	48	25	24	7	18
Astacus astacus	10	6	6	4	4	2	3
Cancer pagurus	9	9	7	2	1	0	0
Homarus americanus	8	5	5	4	4	3	3
Homarus gammarus	8	6	6	4	4	1	1
Nephrops norvegicus	9	7	5	3	3	2	3
Pacifastacus leniusculus	7	6	6	3	3	1	2
Pontastacus leptodactylus	9	6	6	6	6	3	2
Sum crab/crayfish/ lobster species	60	45	41	26	25	12	14
Total	123	97	89	51	49	19	32

Initially, ten peptides per species were tested. Amount of peptide markers candidates were given after general proof of technical suitability (validation level 1) monitoring with the established targeted MRM-method and also after optimization of the collision energy (CE) testing values of 10, 20 and 30 (validation level 3). The specificity was evaluated inside the biological group with crustacean reference samples (validation level 2) and further verification was performed with commercial crustacean, mussel, insect and fish species sample material (validation level 4/5). Lastly, the suitability of the peptides for synthesis (validation level 6) was listed. The evaluation was split between LRMS and HRMS data

specificity validation approximately 27.000 spectra were manually evaluated.

Software and statistical applications

Search algorithm for species-specific peptide marker identification

At first, mean values from precursors of untargeted LC-MS raw data (LRMS: peptide marker candidate identification; HRMS: data comparison with LRMS to show methodic possibilities/limitations) were calculated with BioPharma Finder software (version 2.0). For those values, all measured peptide spectra (technical/biological replicates) of each species with available reference material were combined (Supplement Table 12). A basic search algorithm was programmed based on the mean values of the precursor retention times Rt, mass-to-charge ratios m/z and their peak areas, which were partially intensity normalized in accordance to the dilution during enzymatic digestion. Indicators were added for Rt and m/z, which indicates if the first nearest neighbor is bigger than the given threshold value. Running the search algorithm with LRMS data, the closest measured data points had to differ at least more than Rt > 2 min and precursor m/z ratio > 0.2 (both absolute values) to be regarded as specific signals related to the certain data set. The threshold values can be adjusted depending on instrument parameters and biological samples. To evaluate a higher amount of species with LRMS, the data were split into biological groups A and B, which were treated independently to identify biological markers. Furthermore, two modifications of the basic algorithm were developed with a lower grade of specificity. In modified version 1, only the 40 closest data points compared to the monitored target peptides were tested for specificity and the variables of Rt and m/z were sorted and monitored separately. In detail, at first the peptides were sorted based on m/z and then, up to twenty peptides (if available) closest greater and up to twenty peptides (if available) closest smaller m/z were monitored. The target peptides were used for further investigation if all of the 40 peptides differ at least Rt > 2 min compared to the target peptide. The remaining peptides were additionally evaluated by the proof of the m/z threshold sorted by Rt. In modified version 2, the basic algorithm version was used, but peak intensity signals < 50,000 were excluded. The functionality of the algorithm versions, the number of detected peptide marker candidates per algorithm modification, advantages, and possible applications of the certain algorithm versions with examples as well as a comparison with HRMS data are provided in the supplement Tables 11, 12 and 13 and supplement Figs. 2a, b, c, 4a, b, c.

PEAKS (DENOVO algorithm)

The PEAKS Studio software (version 10.6) was used to identify MS^2 fragment spectra from untargeted LC-LRMS/HRMS data by *de novo* sequencing to elucidate the amino acid sequence of identified and validated biomarkers in an unbiases search without the use of a database. The PEAKS DENOVO algorithm was run with the following search parameters. For LRMS data, the mass error tolerance for the parent and fragment ion was both set to 0.5 Da. For HRMS data, the mass error tolerance of the parent ion was set to 8.0 ppm and for the fragment ion 0.02 Da was chosen. For both LRMS/HRMS data trypsin was chosen as the enzyme, carbamidomethylation was selected as a fixed, and an oxidation of methionine as a variable modification.

Random forest model

Species classification was additionally established by a new random forest model working with untargeted LC-LRMS peptide profiles. Each observed peptide of a sample described by the three variables retention time, mass-to-charge ratio m/z and intensity are assigned to a species by a random forest classification. The predicted species of the sample is then given by majority vote over all observed peptides of this sample. At first, raw data of technical and biological replicates were averaged and summarized (BioPharma Finder) per species inside the training and test data sets. Before running the model, pooled raw data peak values were in few cases intensity normalized to account for deviating sample protein content and subsequent volume adjustments during enzymatic digestion to standardize the intensity values between LC-MS runs. The training data set A contained untargeted LC-LRMS data from 3 to 9 specimens of shrimp species of biological group A and the training data set B consisted of untargeted LC-LRMS data from 2 to 8 specimens of one crab, three crayfish, and three lobster species of biological group B (all reference material). The random forest model to classify the crustacean species of a peptide based on the three numerical variables retention time, mass-to-charge m/z ratio and intensity of the precursor was performed with 2001 decision trees (ntree = 2001) and with all three variables as candidates at each decision tree split (mtry = 3). The final species prediction is based on a plurality vote over all peptides of the sample. Different to the typical random forest approach [31], this model was trained on peptides as observations instead of samples. Applying this approach, the random forest was taught a species relation to peptides and thus multiple predictions per sample were used to determine the species based on a prediction of majority vote. The model ruggedness was proven by an exemplary test data set filled with crustacean species

verified by DNA analysis. Non of the test data were part of the training data. The test data set was composed of four commercially derived samples including *N. norvegicus*, *H. americanus*, *P. monodon* and *P. vannamei* (Table 4). As a post hoc test, the authors would recommend to apply the chi-square goodness-of-fit test to the result of plurality vote to determine significant differences from the expected absolute frequencies of the top two species based on a significance level of 99%. A case of the absence of a significant difference indicates that the species prediction needs to be interpreted with caution. Further, a chi-square goodness-of-fit test to the result of plurality vote to determine significant differences from the discrete uniform distribution was tried. Data management, data visualisation and the statistical approach were performed with the following R packages: base/stats/utils [32], ggplot2 [33], pacman [34], readxl [35], randomForest [36] and sanzo [37] inside the R Studio platform. The programmed R-script of the statistical model (Supplement Table 26) and the training data of biological group A and B (Supplement Fig. 7a, b; Supplement_02) were stated in supplement section.

Peptide synthesis

The results of the *de novo* sequencing were analysed by a developed system based on adverse features compared to an optimized structure [25, 38, 39] for a biological marker (Supplement Table 19) to find peptides, which were most appropriate for synthesis. The exact evaluation criteria including the average local confidence (ALC) score and their impact on the final result were listed in supplement Table 19. Besides structural criteria of the proposed peptide sequences, adverse features were also assigned for non-structural criteria as missing peptide signals (specimens/technical replicates) or low signal intensity (Supplement Tables 19, 24). In general, peptides with an adverse feature value less than 10 were rated as suitable for peptide synthesis (Table 1). The peptide synthesis was performed with a Liberty Blue Microwave Peptide Synthesizer (CEM, Kamp-Lintfort, Germany) using solid-phase 9-fluorenylmethoxycarbonyl (Fmoc-) chemistry with subsequent purification as described earlier [40]. The identities of the synthesized peptides were verified by targeted LC-LRMS MRM.

Results and discussion

Validation of the untargeted LC-LRMS method

Additional information about validation peptides 1–30 including their retention times and m/z-values are provided in the supplement section.

Validation peptides with different retention times over the course of the gradient showed a predominantly good recovery between 66.3 and 102.5% in trap measuring mode with an average CV of 5.1%. Except for validation peptide 5 the online-linked purification (trap mode) led to a low to moderate loss compared to guard mode measurements. Especially more hydrophobic validation peptides (8–10) with higher retention times tended to have weaker recovery rates of about 65–80%. Besides validation peptides 9/10, all monitored peptides were inside the proposed recovery range of 70–120% based on validation guidelines summarized by Kruve et al. [41].

Low to moderate matrix effects between 70.2 and 105.1% and an average CV of 6.9% were determined. Except for validation peptide 10 the peak intensity was lower with matrix compared to the pure standard. Matrix influence caused by adduct formation can be attributed to the mineral content of crustaceans [4] which was even enriched by the sodium chloride containing sample buffer [25]. The exclusion of a more complex sample preparation e.g., solid phase extraction [42] also increased the probability of occurring matrix effects, which on the other hand should not considerably disturb qualitative measurements. Therefore, the occurring matrix effects were lower than expected and tolerable.

A linear relationship in quantification was found for all selected validation peptides of the protein standard (R^2 between 0.982 and 0.999) and lobster sample (R^2 between 0.992 and 0.999). Therefore, the linear relationship between the peak values and the sample concentration was not substantially affected by the crustacean matrix background. Some more hydrophobic peptides showed a lower range of linearity which resulted in a lower strength of correlation.

For the lobster validation peptides, the CVs varied between 4.5 and 23.5% (average CV: 11.1%) indicating an acceptable daily precision of the instrument. More hydrophobic peptides with a higher retention time showed a weaker daily precision with a constantly decreasing peak area in two cases (validation peptide 19/20). Referring to Kruve et al. [41], solely validation peptide 20 exceeded the limit of 15–20% for a within-run CV. Figure 2 shows exemplarily the course of technical replicates for validation peptide 12. The good daily precision of this peptide (CV: 4.5%) is confirmed by the normalized double (± 0.09) and triple (± 0.14) SD which was not exceeded by any measuring point. Furthermore, no continuous trend was recognized (Fig. 2).

The same validation peptides showed a higher CV range 12.1–38.2% (average 20.7%) at various measuring days which was in most cases a tolerable precision for a longer measurement period. The precision on different days is further influenced by additional factors e.g., temperature and impurity of the analytical column [43]. Concordant to daily

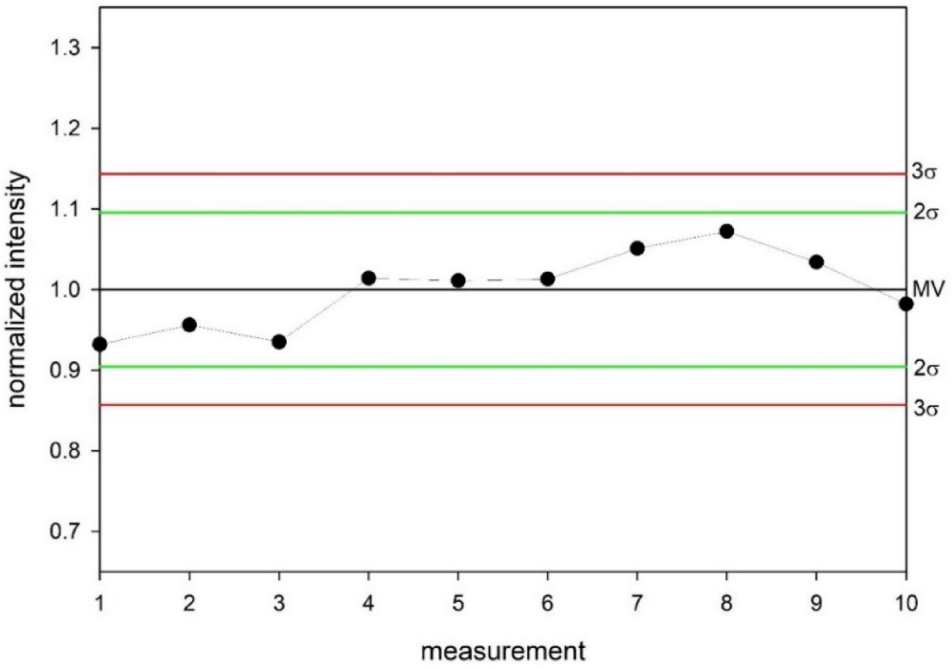

Fig. 2 Daily precision of validation peptide 12 in consecutive technical replicates ($n = 10$) from one biological replicate of the lobster validation sample (*Homarus americanus*)

precision, a weaker precision was observed for more hydrophobic peptides.

Figure 2 shows the normalized peak values in relation to the normalized mean value about the course of measurements. The normalized mean value (MV) at 1 is given by the black line. The green lines visualise the double (±0,09) and the red lines the triple (±0,14) normalized standard deviation (SD).

A high stability of the validation peptides was recognized independent of storage conditions. In 14 of 80 cases (17.5%) a decrease greater than 10% compared to the reference value was determined, including six cases in the group of refrigerator storage. The highest percentual loss was measured for the hydrophobic peptides 19/20 (-35.9%/-21.0%, both not denatured) in the freezing group of long-term storage. This was attributed to slightly more unstable measuring signals for peptides with higher retention times at the end of the gradient. Heat denaturation of the protein extracts to inactivate potential proteolytic degradation of the samples [44] did not result in noticeable stability differences for the validation peptides. Comparing to the validation guidelines, the proofed peptides showed a good stability due to the fact that 81.3% (65 of 80 cases) of the stability data were in the range of 85–115%, which is more than the recommended 2/3 proportion of monitored samples [41].

The reproducibility of peptide spectra was monitored by qualitative and quantitative validation. For the qualitative approach, the basic search algorithm determined that 87 of 4609 (1.9%) peptides were detected individually in one of five biological replicates (each composed of three technical measurements) originating from different parts of one shrimp specimen. Concerning interindividual differences, 68 of 3191 (2.1%) detected peptides were individual for one of the five biological replicates from five different specimens of the same species. Combining both effects of intra- and interindividual differences considering all 10 biological replicates from six shrimp validation samples, only 91 out of 7800 peptides (1.2%) were specific in one biological replicate of this dataset. Those percentual values were rated as relatively low and it was therefore estimated that the biological variability would if at all, only slightly influence the identification of species-specific biomarkers and their validation towards species specificity. Based on those results, it becomes clear that a higher number of peptide spectra (Supplement Table 12) could reduce differences within a species. However, those differences should not only be due to biological variability, but also due to measurement inaccuracies as a different number and variety of peptides occurs in every technical replicate generated by untargeted LC-LRMS. For quantitative reproducibility the shrimp validation peptides showed CVs between 7.5 and 18.4% (average CV: 13.2%). When considering the various relevant factors (among others sample preparation), it is remarkable that the minimum, maximum and average CV is lower than the results from the lobster validation sample examining the precision on different days. This insight indicates a huge influence of the instrument conditions for untargeted LC-LRMS measurements. However, the additional evaluation of validation peptide 29 demonstrated a higher CV of 14.7% for the interindividual differences of five specimens of the same species compared to the CV of 10.3% for the intraindividual difference of five samples from the same specimen (Fig. 3a, b).

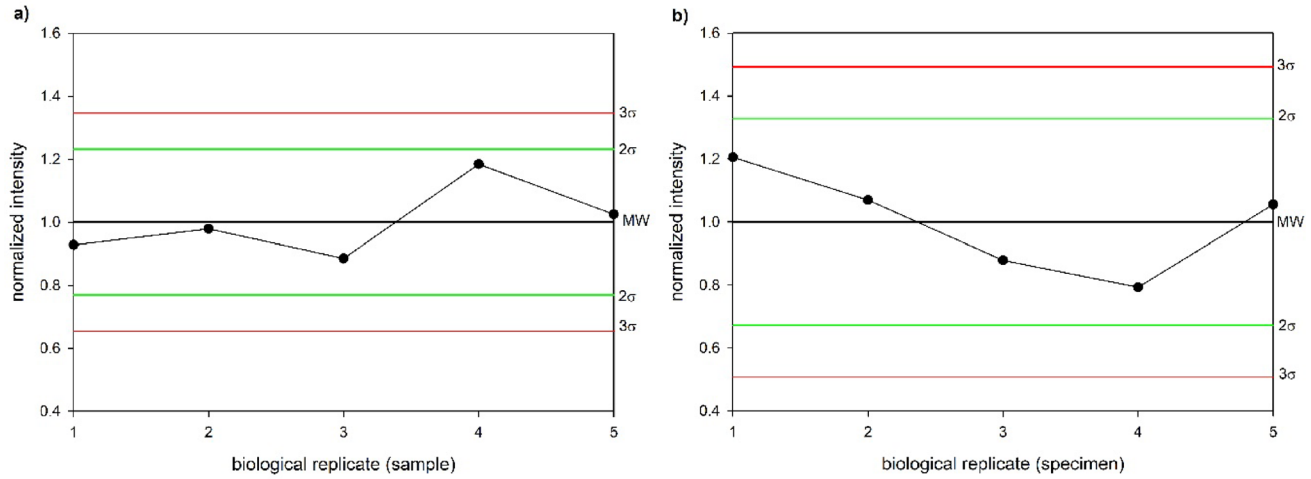

Fig. 3 **a**, **b** Quantitative reproducibility of validation peptide 29 for five biological replicates of one sample (**a**) and for five biological replicates of five specimen (**b**) each from the same shrimp species (*Penaeus vannamei*)

Therefore, there is also an existing impact deriving from the biological variability [30] of shrimps.

Figure 3a, b shows the normalized peak values about the course of measurements ($n=3$ technical replicates per biological replicate; total $n=15$ for both figure parts 3a, b). The normalized mean value (MV) of five biological replicates each is given by the black line. The green lines visualise the double (Fig. 3a: ± 0.21 / 3b: ± 0.29) and the red lines the triple (Fig. 3a: ± 0.31 / 3b: ± 0.44) normalized standard deviation (SD). The coefficient of variation (CV) is 10.3% for intraindividual differences of the same sample (Fig. 3a) and 14.7% for interindividual differences of five specimens (Fig. 3b).

The interindividual reproducibility was also high for the evaluated species-specific peptides measured by the targeted LC-LRMS method (Sects. "Selection and evaluation of species-specific peptides", " *De novo* sequencing (in silico)", "Experimental proof of synthesized peptides"). For 46 out of 51 (90.2%) biomarkers, the peptide analytes were detectable in all specimens and across all technical replicates of the respective species. Only one lobster leg sample (*H. gammarus*) could not be detected using the four *H. gammarus* biomarkers (B-32, B-33, B-34, B-37), and the signal was missing for one technical replicate of a single individual for one *S. mantis* biomarker (A-80).

All described validation parameters showed sufficient to very good results, which were further provided in more detail (Supplement Tables 2, 3, 4, 5, 6, 7 and 8; Supplement Fig. 3). Consistent measuring conditions of the untargeted detected full tryptic hydrolysates were the basis for the further use of those datasets to identify possible suitable peptide markers.

Selection and evaluation of species-specific peptides

The five intense peptide marker candidates (Supplement Table 14, 15) were chosen for each crustacean species, which were identified and selected by the modified versions 1 and 2 of the programmed search algorithms inside the biological groups A and B. Those peptide marker candidates were validated for technical suitability and specificity using the targeted LC-LRMS method.

Figure 4a, b shows extracted ion chromatograms as an example for monitoring specificity inside the biological group (evaluated at validation level 2). Figure 4a displays the mass transitions for the target species *H. americanus*, whereas Fig. 4b clearly demonstrates no signals at the relevant retention time for each transition of the same peptide marker candidate monitoring the different investigated lobster species *H. gammarus*. The absence of the signal for Fig. 4b was further emphasized though a zoom in of the relative intensity of factor 10. This peptide marker candidate also showed to be a suitable biological marker for *H. americanus* after validation level 5, fulfilling all further conditions.

Figure 4a, b represents the three mass transitions of the peptide marker candidate B-21 from the lobster species *Homarus americanus* (retention time Rt: 4.24 min; mass-to-charge ratio m/z: 565.6 → 884.5 (y8; blue) / 656.3 (y6; red) / 983.5 (b10; green) as a peptide marker candidate stage. In Fig. 4a the mass transitions were shown for the target species *Homarus americanus* and in Fig. 4b for *Homarus gammarus*. The data used for visualization were taken at validation level 4 and therefore after intensity optimization (validation level 3) of the mass transitions. The evaluation

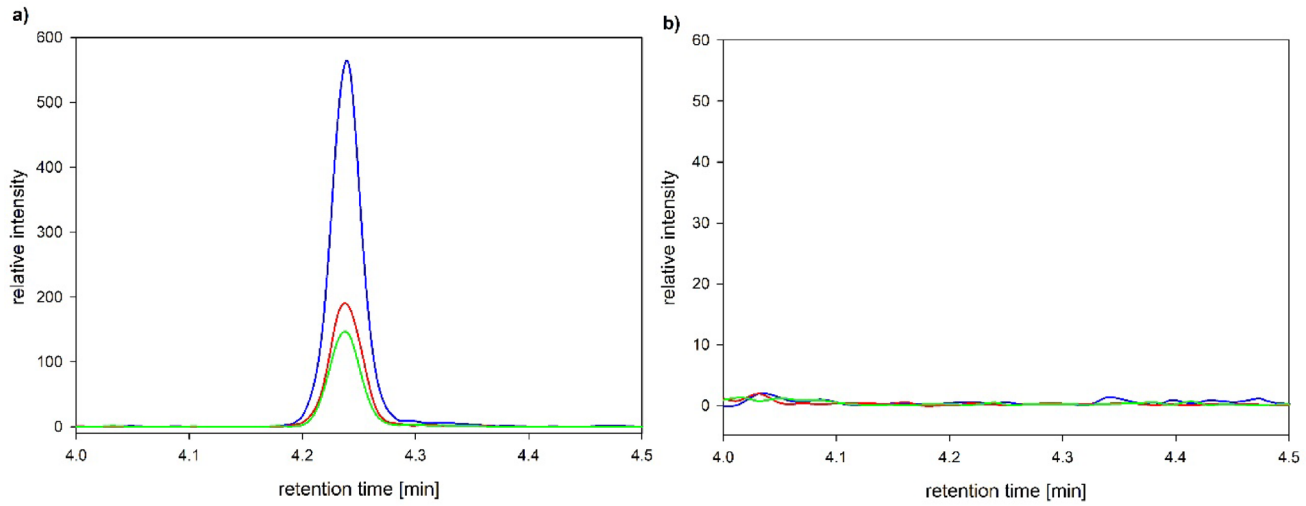

Fig. 4 **a**, **b** Comparison of extracted ion chromatograms inside biological group B

of specificity inside the two biological groups were performed at validation level 2.

All in all, the experimental and software-evaluated results demonstrated the functionality and the effectiveness of the modified versions of the search algorithm for biomarker identification with LRMS data.

Table 1 summarizes the number of remaining peptides for the different validation levels. In total, 49 of 150 (32.7%) peptide marker candidates were identified as species-specific biomarkers based on the investigated data set (Table 1). Table 1 shows that the highest number of peptide markers (6) was achieved for the crayfish species *P. leptodactytus*, whereas only one biological marker was found for the shrimp species *C. allmannii* and the crab species *C. pagurus*.

The targeted MRM-method enabled a possible differentiation of 14 crustacean species. However, no biological markers were found for the white tiger prawn (*P. vannamei*) due to the missing reference sample material of the closely related shrimp species *P. stylirostris*. In consequence, marker candidates for *P. vannamei* were not specific during evaluation against commercial crustacean samples of *P. stylirostris* and were therefore excluded. However, a different example related to the developed methods using reference material of mussel species (*M. edulis*/*M. trossulus*) showed the possibility of identifying species-specific markers of closely related species [45] by LRMS peptide profiles (Supplement Fig. 5a, b). Further evaluation of the specificity of crustacean peptide marker candidates was performed by analyzing mussel, insect, and fish species. Two peptide marker candidates were excluded after validation level 5, which both became unspecific due to fish species. The mass transitions (Supplement Tables 14, 15) of the identified biomarkers could be adapted from different laboratories even though different chromatographic conditions were selected. These peptides could be applied for the identification of one of the 14 crustacean species without the need for HRMS and with a reviewed species specificity inside the known sample material (Supplement Table 1).

Since the three mass transitions of the biomarkers are highly specific, with no or negligible interference from non-target species observed during validation – including lobster species from the same genus (Fig. 4a, b) – the targeted MRM method may be suitable of detecting crustacean mixtures. The high multiplexity of the MRM method arises from low dwell times (Supplement Tables 16, 17) for the measurement of mass transitions. Furthermore, the possibility of the scheduled mode to focus on selected peptides in predefined time windows (Sect. "LC-MS measurements") was carried out, in which particular peptides elute from the column. This would offer an advantage over DNA-based methods, as mixtures can typically only be identified using next generation sequencing (NGS). However, NGS-approaches are currently still time-consuming, both in terms of experimental procedures and the required bioinformatic analysis [11]. Nevertheless, the applicability of the LC-MS method would depend on the proportion, concentration, and number of species in the mixture, and would ultimately be limited by the detection thresholds of the peptide markers.

De novo sequencing (in silico)

Species-specific biomarkers after validation level 5 were submitted to *in silico de novo* sequencing to possibly elucidate some amino acid structures. Table 2 shows selected results of software-based proposed sequences compared between LRMS and HRMS data.

Table 2 Selected results for peptide sequences of species-specific biomarkers (after validation level 5) from biological group A based on *de Novo* sequencing (in silico)

Peptide marker	Crustacean species	Resolution	Biomarker peptide sequence	ALC-score	Adverse feature value
A-03	*Aristeus antennatus*	LRMS	SSMEENLSQLDNLR	92	5
		HRMS	AAYEENLSQLDNLR	92	3
A-61	*Plesionika martia*	LRMS	SSWSDDVMVVAAALR	79	5
		HRMS	SSGESDDDDVVAAALR	98	0
A-79	*Squilla mantis*	LRMS	AVFPSLVGPR	81	6
		HRMS	AVFPSLVGPR	72	7

The Elucidation of the peptide sequences by *de Novo* sequencing was performed using PEAKS-Software with untargeted LC-LRMS and LC-HRMS raw data. the results of both instruments were given regarding the proposed amino acid sequence, the average local confidence (ALC) score and the adverse feature value

Exemplary, for the specific peptide marker A-03, similar peptide sequences with the same ALC-score were identified by the PEAKS-software from both LRMS and HRMS data. Still with HRMS data a lower adverse feature value (3) compared to an optimal peptide sequence was obtained due to the oxidative amino acid methionine given for LRMS data proposal (5). Again, for peptide marker A-61 comparable structures were computed, but the considerably higher ALC-value and lower adverse feature value were acquired from HRMS data. In contrast to peptide marker A-79, which showed a higher ALC-value and a lower adverse feature value for LRMS data for the same peptide sequence. Considering Table 2 and the complete table of both biological groups (Supplement Tables 20, 21, 22 and 23), comparable results between LC-LRMS and LC-HRMS data were obtained in many cases in terms of the proposed amino acid sequence and for the ALC score. However, with HRMS more peptide sequences could be matched to biomarkers based on the measured data (Supplement Tables 20, 22), and more biomarkers were rated to be suitable for peptide synthesis (Table 1). Based on those results and to proven feasibility of LRMS data for peptide structure elucidation, proposed amino acid structures of both LRMS and HRMS data were submitted to peptide synthesis.

Experimental proof of synthesized peptides

A total of 39 peptides were synthesized, belonging to 11 of the 49 identified biomarkers, including seven crustacean species. The higher number of synthetized peptides derived from the approach to take *de novo* sequencing structure proposals of both LRMS and HRMS data into account for structure elucidation of selected peptide markers. Furthermore, the isomeric amino acids leucin and isoleucine cannot be differentiated based on the mass-to-charge ratio resulting in multiple possible structures, if more leucin or isoleucine amino acids are present in the peptide structure. Without using ion mobility as a third dimension [46], leucin and isoleucine have to be exclusively separated based on the chromatography. The verification of the synthesized peptides was performed using the targeted LC-LRMS MRM method.

Figure 5a, b shows the detection of the marker peptide SSGESDDDDVVAAA**I**R (A-61), which was based on HRMS data. The chromatograms show that the peaks of this peptide were detected in the crustacean matrix (Fig. 5a) and in the synthesized peptide (Fig. 5b) with the same order of transitions, the same intensity ratio and almost identical retention time. This is an evident proof of the structure proposal, even though different concentrations were used due to the food matrix influence [47] of the crustacean sample (Fig. 5a, b).

Unambiguous proof for the sequence SSGESDDDDVVAAA**I**R was provided by spiking the crustacean matrix with the synthesized isomeric peptides SSGESDDDDVVAAA**I**R and SSGESDDDDVVAAA**L**R. The sample with the peptide containing isoleucine (Fig. 5c) resulted in a signal amplification of the peak observed in the crustacean matrix. For the corresponding leucin version (Fig. 5d), however, two peaks occurred.

Besides peptide SSGESDDDDVVAAA**I**R, four more biological markers were structurally elucidated as peptide marker (Table 3). Two further peptides both for crayfish species were fully structurally elucidated, whereas one peptide marker for the shrimp species *M. kerathurus* and one for the lobster species *H. americanus* were partly elucidated with the exception of the leucin respectively isoleucine presence or order. This position could be clarified by a slightly modified gradient to generate a higher separation between both versions. For peptide marker A-33 (Table 3), the versions **L**Y**L**TEVC(+57.02)QAVEK and **I**Y**I**TEVC(+57.02)QAVEK could be excluded with the separation of the developed MRM method, so that the sequence of this peptide marker contains one leucine and one isoleucine position each.

For exemplary biomarkers, the final evidence of peptide analytes was achieved based on complete respectively almost complete structure elucidation and those peptide markers could be used from different laboratories as standards. More information on the experimental verification of synthesized peptides is provided in supplement Table 25.

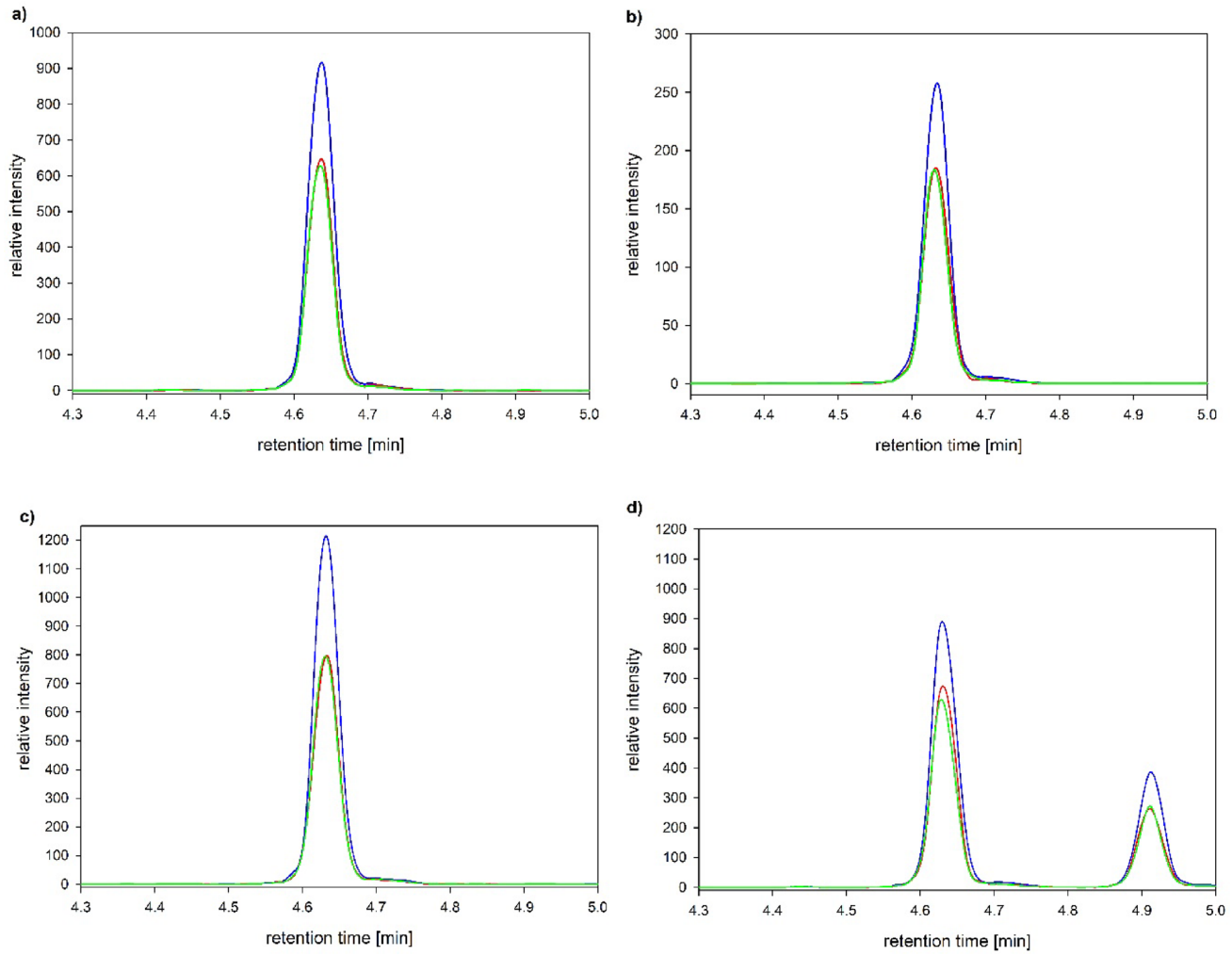

Fig. 5 **a**, **b**, **c**, **d** Detection of marker peptide A-61 (SSGESDDDDVVAAA**I**R; m/z: 804.2 → 501.4 (y5; blue) / 359.3 (y3; red) / 600.5 (y6; green)) from shrimp species *Plesionika martia* in crustacean matrix (**a**; Rt: 4.64 min, c = 290.0 µg/ml) and matrix-free reference standard (**b**; Rt: 4.63 min, c = 0.52 µg/ml). The synthesized sequences SSGESDDDDVVAAA**I**R (**c**; Rt: 4.63 min) and SSGESDDDDVVAAA**L**R (**d**; Rt crustacean matrix: 4.63 min, Rt reference standard: 4.91, c reference standard = 0.50 µg/ml) were spiked in the crustacean matrix (proportion 1:1) to elucidate the amino acid presence of leucin/isoleucine

Random forest model

In addition to peptide marker identification, a statistical screening approach was developed for classification which enables species identification by the numeric distribution of peptides. Figure 6 exemplarily demonstrates the correct identification of an *N. norvegicus* trade sample by the random forest model.

In total, 1212 of 2954 (41.0%) peptides from the *N. norvegicus* test sample (Table 4) were correctly assigned by the trained model. This result is an unambiguous verification that clearly separates from a random percentage distribution of approximately 14.3% considering seven groups. In addition, it is comprehensible that the closely related lobster species *H. gammarus* and *H. americanus* (16.2%/10.5%) received more peptides than the crayfish species (4.0-7.6%). However, 482 peptides and thus the second most amount (16.3%) were distributed to the crab species *C. pagurus* which is genetically more distant compared to the lobster species [13]. A highly significant p-value of the chi-square goodness-of-fit test using only the top two species matches to the clear result of this test sample 4 (Table 5).

To evaluate the accuracy and robustness of the model, the test data set was even extended by crustacean species that were not part of the training data set, e.g., *P. monodon* and with peptide spectra from processed samples e.g., *H. americanus* as interference. Applying the processed lobster as test sample with only six peptide spectra (raw state) included in the training data set (Supplement Table 12), merely 53 of 596 peptides (8.9%) of the *H. americanus* test data set peptides were correctly assigned to the lobster species. This contrasts with the closely related species *N. norvegicus* and

Table 3 Further fully or partly structurally elucidated species-specific biomarkers from biological group A and B

The elucidation of the peptide sequences by *de Novo* sequencing and the assignment to b- or y-ions were performed using PEAKS-software based on untargeted LC-LRMS and LC-HRMS Raw data. the confirmation of the synthesized peptides was applied by targeted LC-LRMS. Unknown leucine or isoleucine positions were highlighted in bold and underlined

Peptide marker	Crustacean species	Confirmed amino acid sequence	Mass-to-charge ratio molecular ion [m/z]	Mass-to-charge ratio fragment ions [m/z]	Fragment ion
A-33	*Metapenaeus kerathurus*	**I**Y**L**TEVC(+57.02)QAVEK / **L**Y**I**TEVC(+57.02)QAVEK	727.1	1063.5	y9
				833.5	y7
				1177.5	y10
B-21	*Homarus americanus*	R**I**TVGEVEVK R**L**TVGEVEVK	565.6	884.5	b8
				656.3	b6
				983.5	b9
B-58	*Pacifastacus leniusculus*	NFGDVNQFVNVDPDGK	883.7	531.3	y5
				744.3	y7
				1235.3	y11
B-62	*Pontastacus leptodactylus*	QFVTEVC(+57.02)QEVEK	748.6	1121.4	y9
				792.3	y6
				891.3	y7

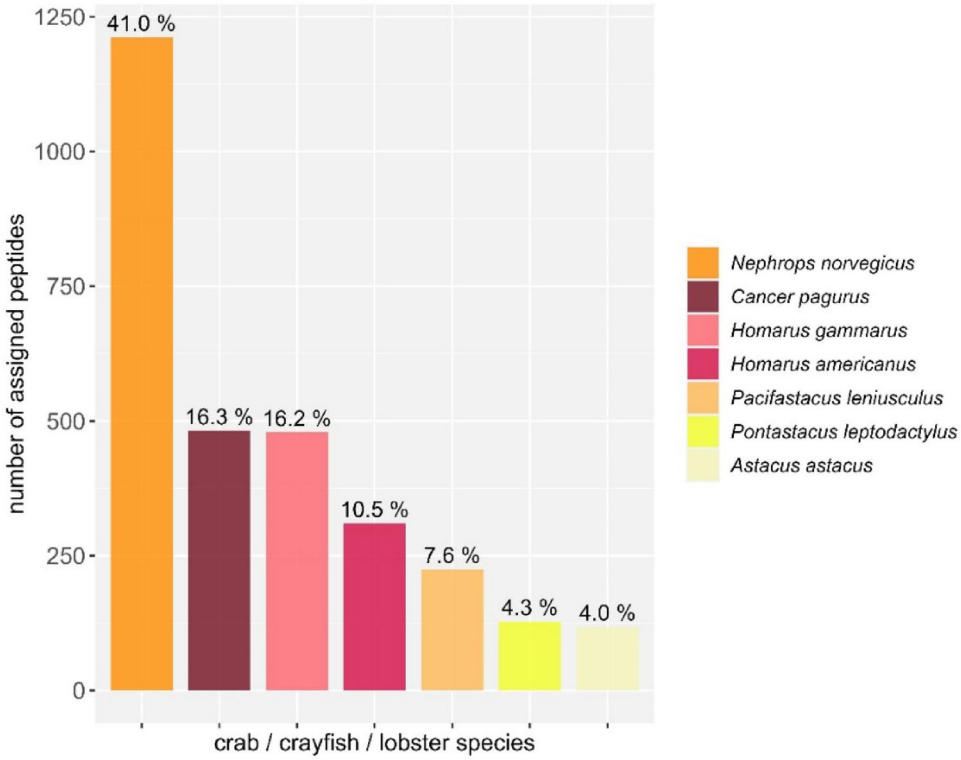

Fig. 6 Peptide distribution of the unseen test data from commercial *Nephrops norvegicus* after applying on the trained random forest model

Table 4 Exemplary test data set containing untargeted LC-LRMS data from samples with different conditions, along with final predictions from a trained random forest model

Test sample	Crustacean species	Species included in training data of biological group B	Processed sample material	Correct prediction based on plurality vote
1	*Homarus americanus*	Yes	Yes	No
2	*Penaeus monodon*	No	No	Not possible
3	*Penaeus vannamei*	No	Yes	Not possible
4	*Nephrops norvegicus*	Yes	No	Yes

H. gammarus, for which the highest and second-highest numbers of peptides were assigned in this test sample 1. For these two species, more spectra from a larger number of individuals (all in the raw state) were available in the training data compared to *H. americanus*. Therefore, a high number of peptide spectra from the same species with the same processing state is necessary to use the random forest model as a screening approach for species identification. However,

Table 5 Evaluation of the random forest model based on a test data set with different sample conditions

Test sample	Number of peptides in test data set	Highest result [%]	Difference to random result of 14,3% for 7 classes [%]	Difference to second- highest result [%]	Quotient of the highest result to the sum of the maximum + second-highest result [%]	Sum of peptides for the two top species	*p*-value of chi-square goodness-of-fit test maximum + second-highest result
1	596	33.6%	19.3%	11.8%	60.6%	330	0.0001165
2	2389	25.1%	10.8%	3.4%	53.6%	1118	0.01673
3	1886	48.0%	33.7%	30.8%	73.6%	1230	< 0.0000001
4	2954	41.0%	26.7%	24.7%	71.5%	1694	< 0.0000001

Detailed results of the plurality vote for each sample of the test data set were summarized in this table, also including the number of peptides in the certain test data set, the sum of peptides for the two top species and p-values of the chi-square goodness-of-fit test used as post hoc test to prove the percentage distribution (maximum compared to the second-highest result) of the species prediction

the size and condition of the training data are independent of the performance and choice of a post hoc test, as a statistical test alone cannot provide a complete interpretation of the prediction. Still, the post hoc test indicates that the false positive results of test sample 1 and also for the processed non-class test sample 3 are significant, which underlines the need of the target species in the training data and the importance of a comparable sample state between training and test data (Table 4). In case of the non or other class at raw state, the shrimp species *P. monodon* shows the expected, more uniform distribution of a non-class test sample, resulting in a non-significant p-value of 0.017 (significance level: 99%), when comparing the maximum with the second-highest result (Table 5).

The chi-square goodness-of-fit test was also performed using the given percentage distribution of all seven classes from biological group B. In this test, all p-values for the four crustacean samples contained in the test data set were <0.01, indicating a non-uniform distribution based on a significance level of 99%. This is comprehensible, since at least one species of the training data differs to the equal frequency.

The purpose of this small test dataset was primarily to evaluate different conditions and scenarios applied to the random forest model (Table 4). The usability of the model and subsequent post hoc test for detection of unknown crustacean species should be further proved in an interlaboratory trial.

Comparison of developed methods and possible applications

In this study an untargeted and a targeted proteomics-based analytical method as well as a statistical screening approach were introduced. The LC-LRMS MRM and the random forest model could both be used for species identification of the selected species. However, the LC-LRMS MRM approach is more rugged and reliable due to the comprehensive validation of species-specific peptides whereas the random forest model can only provide a prediction based on training data originating from the untargeted LC-LRMS method. The targeted MRM method can be used as a standalone authentication method provided that species-specific biomarkers were identified for the listed crustacean species (exception: *P. vannamei*). However, the statistical screening approach could be applied to confirm the result of biomarkers and it is also advantageous to receive a species prediction (e.g., *Penaeus genus*) without the need of specific markers, but depended to the highest number of assigned peptides from the full tryptic digest. Moreover, it was previously discussed (Sect. "Selection and evaluation of species-specific peptides") that the high specificity of the selected mass transitions using targeted LC-MS MRM might be suitable to measure mixtures. Concerning random forest, the developed screening model based on untargeted LRMS data and a prediction based on plurality vote would not be suitable for the detection of multiple species. It might be even difficult in case of a more uniform distribution to justify the test result for one species with the aid of an objective statistical post hoc test. One concern would be that several peptides with the same sequence would be included in multiple closely related species.

The adoption of those methods is easy possible for a larger analytical community or control authorities. Concerning the developed LC-LRMS MRM method, the m/z-values of the mass transitions could be transferred to a LRMS device even with a different chromatographic set up (e.g., columns). In principle, the alternative proteomics-based workflow (Fig. 1) using the cheaper and thus more affordable LRMS could be applied for other emerging issues, but it requires bioinformatics expertise and appropriate software. The random forest model could be also used monitoring different test data due to the fact that the R code (Supplement Table 26) of the screening model and the training data of biological group A and B were provided in the supplement section. In detail, the developed statistical approach offers several advantages in terms of transferability. Due to the new peptide-to-feature relationship, it is not necessary to group features collectively by sample (bucketing). As a result, no prior alignment within the statistical model

is required, owing to the intrinsic data-splitting capabilities of the random forest algorithm. This means that untargeted LC-LRMS data only need to be intensity-normalized as a preparation step. Another advantage of a non-bucketed approach is the avoidance of missing values. Consequently, training datasets can be easily expanded or modified using newly acquired data from comparable untargeted LC-LRMS conditions, and predictions for subsequent generated test samples remain feasible.

Advances over previous MS-based approaches for shrimp species identification and other instrumental analytical methods in food authentication

The evaluated data showed the possibility to use LRMS for species authentication even though a greater effort is necessary. In contrast to for instance Chatterjee et al. [19], Hu et al. [18] or Lu et al. [21] using HRMS, in this study the identification of peptide marker candidates was performed based on untargeted LC-LRMS data. Identified species-specific peptide markers were subsequently detected by targeted LC-LRMS and they enable the secure identification of various crustacean species. Still, not only the bioinformatic evaluation is time-consuming, but so is the experimental specificity validation based on targeted LC-LRMS MRM. An alternative approach involves the use of spectral libraries, which rely on a pre-constructed MS spectral database used for comparison with test samples [20]. MALDI-MS fingerprinting approaches, such as those by Salla and Murray [20] or Stahl and Schroder [48], offer more rapid sample preparation. However, the development and validation required to establish a reliable MS spectral library are also labor-intensive. Although sample preparation for LC-LRMS MRM is more time-consuming – particularly due to the overnight digestion – the final method allows for the detection of 14 different crustacean species within 15 min. As in this study, unambiguous and high-quality reference materials would form the basis for building a valid spectral library. However, the inclusion of closely related species in the dataset is even more critical, as the absence of a close relative with a similar protein pattern could result in false-positive identifications, despite achieving a high matching score [48]. Varunjikar et al. [10] established a proteomics approach based on spectral library matching using LC-MS to differentiate fish species. Similar to the approach described here, it does not rely on protein identification. However, although species-specific proteomes are used, verification is based on a matching score. A limited number of target peptides provide three species-specific mass transitions (according to the evaluated dataset). The evaluation of peptide markers is also straightforward, involving a comparison of the number and order of transitions, retention time, and intensity with reference material from a given species. Another analytical technique for food authenticity testing is nuclear magnetic resonance (NMR) spectroscopy [49]. A key advantage of NMR is its ability to measure all components of a sample in a single experiment [49], whereas this study focuses solely on proteins and digested peptides. The simultaneous analysis of various ingredients can reduce the cost per sample. However, NMR data can result in highly complex spectra, which are prone to signal overlapping [49]. Furthermore, the NMR technology is non-destructive, meaning that the sample material can be reused for other experiments [50]. In contrast, while LC-MS samples can be frozen and stored after measurement for potential reanalysis, they cannot be repurposed for other applications (e.g., immunological experiments), as the native protein structure is irreversibly altered during sample preparation. This is especially relevant when considering the limited quantity of available sample material. Typically, crustaceans provide several grams of muscle tissue, which is sufficient for routine protein extraction using approximately 250 mg of tissue. However, for smaller shrimp species (e.g., *C. almannii*), it may only be possible to obtain a single biological replicate per specimen. In such cases, verified reference materials may be difficult to acquire, making them particularly valuable. NMR applications have also demonstrated high reproducibility and robustness, even without a prior chromatographic separation [49, 50]. In the previous section, it was mentioned that the m/z-values of the developed targeted LC-LRMS MRM method could be transferred to a different LRMS device. However, the second dimension of retention time depends on chromatographic conditions (e.g., columns, mobile phases) and is therefore not easily transferable. In addition, the inherent ability of quantitative nuclear magnetic resonance (qNMR) to enable quantification without the need for calibrations or reference standards highlights its potential [49, 50]. In this study, *de novo* sequencing, peptide synthesis, and subsequent verification using targeted LC-MS MRM were time-consuming and costly, and were performed to preserve selected peptide standards for the identified biomarkers. Nevertheless, the main drawback of NMR technology compared to MS-based methods remains its lower sensitivity [49]. For the biomarker B-62 belonging to the crayfish species *P. leptodactylus*, a very low detection limit of 1 ng/ml (signal-to-noise ratio around 3) was exemplarily determined for the isolated and structurally confirmed peptide standard QFVTEVC(+57.02)QEVEK. Overall, crustacean peptide markers that were not yet described in the current literature were discovered due to an extensive data set, which included also further crustaceans apart from shrimp species. Those insights and the evaluated limits of LRMS for species identification

in general could be also useful for different protein-based matrices apart from crustaceans.

Concerning the chemometrics model, Chatterjee et al. [19], Hu et al. [18] or Lu et al. [21] used principle component analysis and orthogonal partial least square discriminant analysis to predict shrimp species identity or for biomarker discovery. Different to those approaches, this article presents so far as we know the first random forest classification model based on untargeted LC-LRMS training data to identify unknown crustacean species. Dalal et al. [51] combined FT-NIR (Fourier transform near-infrared) data with machine learning, including random forest, for fish species identification. Similarly, Tata et al. [52] used random forest, among other approaches, for the species identification of edible insect powders. In their study, the authors employed HRMS data obtained using a DART (direct analysis in real-time) method. However, to the best of our knowledge, there are currently no random forest models available in the field of proteomics for food authentication, nor specifically for crustacean species identification. Nevertheless, various chemometric techniques have been developed for the authentication and adulteration control of foods of animal origin [53]. It is important, to emphasize the novelty of the analyte-to-feature relationship approach presented here: both aforementioned studies based their random forest predictions on the sample level. This concept could also be practical for other food types and different analytes domains (e.g., metabolomics, lipidomics), especially in cases involving a high number of features combined with a relatively low number of samples.

Conclusion and outlook

The results obtained by the developed untargeted and targeted LC-LRMS methods plus an additional statistical model showed the possibility to determine seven shrimp species and seven lobster, crab and crayfish species without using HRMS devices putting stronger emphasis on the bioinformatic evaluation. Beyond the proven feasibility of solving authenticity questions with LRMS data e.g., for a lower number of closely related species, a variety of peptide markers for further crustaceans were identified. Those biomarkers could be even more extensively validated concerning specificity by a higher number of trade relevant crustacean species and further evaluated towards technical applicability with processed sample material or as a component in a more complex matrix e.g., oil. Selected biomarkers were synthesized based on the results of *de novo* sequencing and some suggested amino acid sequences were confirmed experimentally. Even though, for the synthesis part, some limitations cannot be overcome by applying LRMS, which makes the structural elucidation not impossible but complex without HRMS data. Independent to a possible peptide structure confirmation, all 49 provided biomarkers were clear enough to declare them as species-specific inside the proven data and therefore those analytes can be used for the crustacean authentication of multiple species. Further on, the novel analyte-to-feature random forest approach could be also transferred to different LC-MS datasets from various analytes and matrices. An extended methodological investigation in the future should offer further limitations and potentials of this new strategy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00217-026-05062-3.

Acknowledgements The authors would like to thank Martin Clare, Björn Neumann and Pablo Zamora for their technical assistance in the LC-LRMS section and Liane Weber for peptide synthesis. Furthermore, they also like to thank the DNA group of Dr. Kristina Kappel for the species verification of trade samples as well as Dr. Paola Ferrario, Dr. Benedikt Merz and Fynn Brix for further statistical advice. Lastly, the authors would like to thank Dr. Katja Kaltenbach for providing NMR expertise.

Author contributions T. R.: Writing – original draft, Investigation, Methodology, Validation, Visualization. C. B.: Investigation – DNA analysis. M. D.: Writing – review and editing, Investigation – random forest, Software. W. J.: Writing – review and editing, Investigation – peptide synthesis C. T.: Writing – review and editing, Investigation – HRMS dataR. H.: Writing – review and editing, Conceptualization, Funding acquisition, Resources. M. B.: Resources. A. Y.: Resources. A. T.: Supervision. I. C.-R.: Conceptualization, Supervision. U. S.: Writing – review and editing, Conceptualization, Funding acquisition, Project administration.

Funding Open Access funding enabled and organized by Projekt DEAL. This project was supported by funds from the Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) based on a decision by the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the Innovation Support Programme (Förderkennzeichen: 281A104416).

Data availability Due to the specifications of the Nagoya-protocol implemented into European law (EU ABS VO Nr. 511/2014) and the internal handling with this law applied by the Max-Rubner Institute (institute of the lead author T. Roggensack), only LRMS data from crustacean reference sample material were given in the form of training datasets. LRMS data from commercial sample material used for development or validation as well as HRMS data used for method comparison were not provided.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval This study does not involve any human testing.

References

1. Farmery AK, Alexander K, Anderson K, Blanchard JL, Carter CG et al (2022) Food for all: designing sustainable and secure future seafood systems. Rev Fish Biol Fish 32(1):101–121. https://doi.org/10.1007/s11160-021-09663-x
2. Zhang YF, Ren YY, Bi YG, Wang Q, Cheng KW et al (2019) Review: seafood allergy and potential application of high hydrostatic pressure to reduce seafood allergenicity. Int J Food Eng 15:812. https://doi.org/10.1515/ijfe-2018-0392
3. Khora SS (2016) Seafood-associated shellfish allergy: a comprehensive review. Immunol Invest 45(6):504–530. https://doi.org/10.1080/08820139.2016.1180301
4. Devesa V, Martínez A, Súñer MA, Vélez D, Almela C et al (2001) Effect of cooking temperatures on chemical changes in species of organic arsenic in seafood. J Agric Food Chem 49(5):2272–2276. https://doi.org/10.1021/jf0013297
5. Ortea I, O'Connor G, Maquet A (2016) Review on proteomics for food authentication. J Proteom 147:212–225. https://doi.org/10.1016/j.jprot.2016.06.033
6. Ortea I, Cañas B, Calo-Mata P, Barros-Velázquez J, Gallardo JM (2009) Arginine kinase peptide mass fingerprinting as a proteomic approach for species identification and taxonomic analysis of commercially relevant shrimp species. J Agric Food Chem 57(13):5665–5672. https://doi.org/10.1021/jf900520h
7. Kappel K, Schröder U (2015) Species identification of fishery products in Germany. J Verbrauch Lebensm 10:S31–S34. https://doi.org/10.1007/s00003-015-0988-y
8. Hellberg RSR, Morrissey MT (2011) Advances in DNA-Based techniques for the detection of seafood species substitution on the commercial market. Jala 16 4:308–321. https://doi.org/10.1016/j.jala.2010.07.004
9. Kranz B, Fritsche J, Jira W, Langenkämper G, Roggensack T et al (2025) Mass spectrometry of Proteins, peptides and small molecules for food authentication - a systematic survey. Food Rev Int. https://doi.org/10.1080/87559129.2025.2553677
10. Varunjikar MS, Pineda-Pampliega J, Belghit I, Palmblad M, Grøsvik BE et al (2024) Fish species authentication in commercial fish products using mass spectrometry and spectral library matching approach. Food Res Int 192:11. https://doi.org/10.1016/j.foodres.2024.114785
11. Klapper R, Velasco A, Döring M, Schröder U, Sotelo CG et al (2023) A next-generation sequencing approach for the detection of mixed species in canned tuna. Food Chem X 17:12. https://doi.org/10.1016/j.fochx.2023.100560
12. Montowska M, Kasałka-Czarna N, Sumara A, Fornal E (2024) Comparative analysis of the *longissimus* muscle proteome of European wild Boar and domestic pig in response to thermal processing. Food Chem 456:12. https://doi.org/10.1016/j.foodchem.2024.139871
13. Brenn C, Schröder U, Hanel R, Arbizu PM (2021) A multiplex real-time PCR screening assay for routine species identification of four commercially relevant crustaceans. Food Control 125:8. https://doi.org/10.1016/j.foodcont.2021.107986
14. Giusti A, Armani A, Sotelo CG (2017) Advances in the analysis of complex food matrices: species identification in surimi-based products using next generation sequencing technologies. PLoS ONE 12(10):18. https://doi.org/10.1371/journal.pone.0185586
15. Giusti A, Malloggi C, Lonzi V, Forzano R, Meneghetti B et al (2023) Metabarcoding for the authentication of complex seafood products: the fish burger case. J Food Compos Anal 123:9. https://doi.org/10.1016/j.jfca.2023.105559
16. Lorusso L, Shum P, Piredda R, Mottola A, Maiello G et al (2024) Mismanagement and poor transparency in the European processed seafood supply revealed by DNA metabarcoding. Food Res Int 194:10. https://doi.org/10.1016/j.foodres.2024.114901
17. Ortea I, Cañas B, Gallardo JM (2011) Selected tandem mass spectrometry ion monitoring for the fast identification of seafood species. J Chromatogr A 1218(28):4445–4451. https://doi.org/10.1016/j.chroma.2011.05.032
18. Hu LP, Zhang HW, Zhang XM, Zhang TT, Chang YG et al (2018) Identification of peptide biomarkers for discrimination of shrimp species through SWATH-MS-based proteomics and chemometrics. J Agric Food Chem 66(40):10567–10574. https://doi.org/10.1021/acs.jafc.8b04375
19. Chatterjee NS, Chevallier OP, Wielogorska E, Black C, Elliott CT (2019) Simultaneous authentication of species identity and geographical origin of shrimps: untargeted metabolomics to recurrent biomarker ions. J Chromatogr A 1599:75–84. https://doi.org/10.1016/j.chroma.2019.04.001
20. Salla V, Murray KK (2013) Matrix-assisted laser desorption ionization mass spectrometry for identification of shrimp. Anal Chim Acta 794:55–59. https://doi.org/10.1016/j.aca.2013.07.014
21. Lu WB, Wang PY, Ge LJ, Chen X, Guo SY et al (2022) Real-time authentication of minced shrimp by rapid evaporative ionization mass spectrometry. Food Chem 383:8. https://doi.org/10.1016/j.foodchem.2022.132432
22. Gu SQ, Deng XJ, Shi YY, Cai YC, Huo YH et al (2020) Identification of peptide biomarkers for authentication of Atlantic salmon and rainbow trout with untargeted and targeted proteomics approaches and quantitative detection of adulteration. J Chromatogr B 1155:7. https://doi.org/10.1016/j.jchromb.2020.122194
23. Li JY, Wang Q, Wang YC, Jiang BX, Chang YG et al (2023) Identification and detection of protein-derived adulterants in oyster peptide powder through an untargeted and targeted proteomics workflow. Food Control 153:8. https://doi.org/10.1016/j.foodcont.2023.109896
24. Aydogan C (2020) Recent advances and applications in LC-HRMS for food and plant natural products: a critical review. Anal Bioanal Chem 412(9):1973–1991. https://doi.org/10.1007/s00216-019-02328-6
25. Korte R, Lepski S, Brockmeyer J (2016) Comprehensive peptide marker identification for the detection of multiple nut allergens using a non-targeted LC-HRMS multi-method. Anal Bioanal Chem 408(12):3059–3069. https://doi.org/10.1007/s00216-016-9384-4
26. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72(1–2):248–254. https://doi.org/10.1016/0003-2697(76)90527-3
27. von Oesen T, Treblin M, Clawin-Räedecker I, Martin D, Maul R et al (2023) Identification of marker peptides for the whey protein quantification in Edam-type cheese. Foods 12(10):21. https://doi.org/10.3390/foods12102002
28. Kuklenyik Z, Calafat AM, Barr JR, Pirkle JL (2011) Design of online solid phase extraction-liquid chromatography-tandem mass spectrometry (SPE-LC-MS/MS) hyphenated systems for

quantitative analysis of small organic compounds in biological matrices. J Sep Sci 34(24):3606–3618. https://doi.org/10.1002/jssc.201100562
29. Altmann K, Wutkowski A, Klempt M, Clawin-Rädecker I, Meisel H et al (2016) Generation and identification of anti-inflammatory peptides from bovine β-casein using enzyme preparations from cod and hog. J Sci Food Agric 96(3):868–877. https://doi.org/10.1002/jsfa.7159
30. Peñafiel R, Ruzafa C, Monserrat F, Cremades A (2004) Gender-related differences in carnosine, anserine and lysine content of murine skeletal muscle. Amino Acids 26(1):53–58. https://doi.org/10.1007/s00726-003-0034-8
31. Breiman L (2001) Random forests. Mach. Learn 45(1):5–32. https://doi.org/10.1023/a:1010933404324
32. R Core Team (2025) _R: A language and environment for statistical computing_. R Foundation for statistical computing, Vienna, Austria. https://www.R-project.org/
33. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York. https://ggplot2.tidyverse.org
34. Rinker TW, Kurkiewicz D (2017) pacman: package management for R. version 0.5.0., Buffalo, New York. http://github.com/trinker/pacman
35. Wickham H, Bryan J (2025) _readxl: Read Excel Files_. R package version 1.4.5, https://CRAN.R-project.org/package=readxl, https://doi.org/10.32614/CRAN.package.readxl
36. Liaw A, Wiener M (2002) Classification and regression by randomForest. 2(3):18–22, _R News_. https://CRAN.R-project.org/doc/Rnews/
37. Maasch J (2020) _sanzo: color palettes based on the works of Sanzo Wada_. R package version 0.1.0. https://CRAN.R-project.org/package=sanzo
38. Mazzeo MF, Siciliano RA (2016) Proteomics for the authentication of fish species. J Proteom 147:119–124. https://doi.org/10.1016/j.jprot.2016.03.007
39. Nagai H, Minatani T, Goto K (2015) Development of a method for crustacean allergens using liquid chromatography/tandem mass spectrometry. J AOAC Int 98(5):1355–1365. https://doi.org/10.5740/jaoacint.14-248
40. Häfner L, Kalkhof S, Jira W (2021) Authentication of nine poultry species using high-performance liquid chromatography-tandem mass spectrometry. Food Control 122:10. https://doi.org/10.1016/j.foodcont.2020.107803
41. Kruve A, Rebane R, Kipper K, Oldekop ML, Evard H et al (2015) Tutorial review on validation of liquid chromatography-mass spectrometry methods: part II. Anal Chim Acta 870:8–28. https://doi.org/10.1016/j.aca.2015.02.016
42. Gerssen A, McElhinney MA, Mulder PPJ, Bire R, Hess P et al (2009) Solid phase extraction for removal of matrix effects in lipophilic marine toxin analysis by liquid chromatography-tandem mass spectrometry. Anal Bioanal Chem 394(4):1213–1226. https://doi.org/10.1007/s00216-009-2790-0
43. Kromidas S (2007) Methodenvalidierung in der Analytik. http://www.kromidas.de/Uploads/Dokumente/ValidierunginderAnalytik.pdf
44. Grienke U, Silke J, Tasdemir D (2014) Bioactive compounds from marine mussels and their effects on human health. Food Chem 142:48–60. https://doi.org/10.1016/j.foodchem.2013.07.027
45. Santaclara FJ, Espiñeira M, Cabado G, Aldasoro A, Gonzalez-Lavín N et al (2006) Development of a method for the genetic identification of mussel species belonging to *Mytilus*, *Perna*, *Aulacomya*, and other genera. J Agric Food Chem 54(22):8461–8470. https://doi.org/10.1021/jf061400u
46. Wu Q, Wang JY, Han DQ, Yao ZP (2020) Recent advances in differentiation of isomers by ion mobility mass spectrometry. Trac Trends Anal Chem 124:7. https://doi.org/10.1016/j.trac.2019.115801
47. Golubović J, Heath E, Heath D (2019) Validation challenges in liquid chromatography-tandem mass spectrometry methods for the analysis of naturally occurring compounds in foodstuffs. Food Chem 294:46–55. https://doi.org/10.1016/j.foodchem.2019.04.069
48. Stahl A, Schroder 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. J Agric Food Chem 65(34):7519–7527. https://doi.org/10.1021/acs.jafc.7b02826
49. Minoja AP, Napoli C (2014) NMR screening in the quality control of food and nutraceuticals. Food Res Int 63:126–131. https://doi.org/10.1016/j.foodres.2014.04.056
50. Bharti SK, Roy R (2012) Quantitative ^{1}H NMR spectroscopy. Trac Trends Anal Chem 35:5–26. https://doi.org/10.1016/j.trac.2012.02.007
51. Dalal N, Sáiz MJ, Caporale AG, Baldini F, Babayan SA et al (2024) Fishy forensics: FT-NIR and machine learning based authentication of mediterranean anchovies (*Engraulis encrasicolus*). J Food Compos Anal. https://doi.org/10.1016/j.jfca.2024.106847
52. Tata A, Massaro A, Marzoli F, Miano B, Bragolusi M et al (2022) Authentication of edible insects' powders by the combination of DART-HRMS signatures: the first application of ambient mass spectrometry to screening of novel food. Foods 11:1511. https://doi.org/10.3390/foods11152264
53. Karabagias IK (2024) Food authentication and adulteration control based on metrics data of foods and chemometrics. Eur Food Res Technol 250(5):1269–1283. https://doi.org/10.1007/s00217-024-04477-0