molecules

Advances in Chemical Analysis Procedures (Part II)

Edited by

Marcello Locatelli, Angela Tartaglia, Dora Melucci, Abuzar Kabir,
Halil Ibrahim Ulusoy and Victoria Samanidou

Printed Edition of the Special Issue Published in *Molecules*

www.mdpi.com/si/molecules

MDPI

Advances in Chemical Analysis Procedures (Part II)

Advances in Chemical Analysis Procedures (Part II)

Statistical and Chemometric Approaches

Editors

Marcello Locatelli
Angela Tartaglia
Dora Melucci
Abuzar Kabir
Halil Ibrahim Ulusoy
Victoria Samanidou

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Marcello Locatelli
University "G. d'Annunzio" of
Chieti-Pescara
Italy

Angela Tartaglia
University "G. d'Annunzio" of
Chieti-Pescara
Italy

Dora Melucci
University of Bologna
Italy

Abuzar Kabir
Florida International University
USA

Halil Ibrahim Ulusoy
Cumhuriyet University
Turkey

Victoria Samanidou
Aristotle University of
Thessaloniki
Greece

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Molecules* (ISSN 1420-3049) (available at: https://www.mdpi.com/journal/molecules/special_issues/chemical_analysis_statistical_chemomitric_approaches).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03936-786-3 (Hbk)
ISBN 978-3-03936-787-0 (PDF)

Contents

About the Editors

Marcello Locatelli (Associate Professor, Analytical Chemistry) earned his degree in Chemistry from the University of Bologna, Department of Chemistry “G. Ciamician” with his thesis on “Development of an Analytical Methodology for the Analysis and Identification of Protein Adducts by Mass Spectrometry”. He earned his PhD from the University of Bologna, Department of Chemistry “G. Ciamician” with his thesis “Combined Analytical Methods of Mass Spectrometry for the Study of Impurities in Drugs and Metabolites of Biomolecules”. Currently, he is Associate Professor of Analytical Chemistry at the University “G. d’Annunzio” of Chieti-Pescara, Department of Pharmacy.

His research activity is devoted to the **development and validation of chromatographic methods** for the qualitative and quantitative determination of biologically active molecules in complex matrices from human and animal (whole blood, serum, plasma, bile, tissues, feces, and urine), cosmetics, foods, and the environment. This includes the study of all processes related to pre-analytical stages, such as sampling, extraction and purification, separation, enrichment, and even the application of conventional and coupled analytical methods for the accurate, sensitive, and selective determination of biologically active molecules. Recently, particular attention has been focused on **innovative (micro)extraction procedures** like MEPS, FPSE, MIP, DLLME, and SULLE. These procedures have been applied to different compounds, from synthetic and natural origin (glucosamine, 5-amino-salicylic acid, natural or synthetic bile acids, anti-inflammatory, drugs association and fluoroquinolones, secondary metabolites from natural sources, heavy metals). In the development of the method are tested also **predictive models and chemometric** both for the optimization of extraction protocols and for final data processing. Particular attention is given to the **new instrument configurations** for the quantitative analysis in complex matrices.

He has published more than 157 manuscripts, 116 congress communications, 1 patent subject to approval, 13 book chapters, and 3 books and served as Guest Editor of 13 Special Issues with attested scientific activity (h-index = 33, 148 papers, and 2778 citations, based on Scopus (8th of June 2020)). In addition, he is a reviewer of the following international journals: *Analytica Chimica Acta*, *Current Bioactive Compounds*, *Journal of Chromatography A*, *Talanta*, and *Trends in Analytical Chemistry* (a selection of the full list covering more than 100 international peer-reviewed journals). He is a referee for MIUR Institution for National Projects (SIR) and included in the register REPRISE (Register of Expert Peer Reviewers for Italian Scientific Evaluation) in the “Basic Research” section and former referee of VQR (2011–2014). He serves other universities as a referee of proposals through competitive tenders for the allocation of University funds for the activation of research grants. He is Editorial Board Member of the journals *Molecules* section “Analytical Chemistry”, *Current Analytical Chemistry*, *Separations*, *Current Bioactive Compounds*, *American Journal of Modern Chromatography*, *Journal of Selcuk University Science Faculty*, *Reviews in Separation Sciences*, and *Cumhuriyet Science Journal*. He is Associate Editor of *Frontiers in Pharmacology* section “Ethnopharmacology”, Review Editor of *Frontiers in Oncology* section “Pharmacology for Anticancer Drugs”, and Review Editor of *Frontiers in Medical Technology* section “Nano-Based Drug Delivery”. He is a member of the Scientific Committee of *Scienze e Ricerche* published by the Italian Book Association.

ADDITIONAL TITLES

Member of the Italian Chemical Society (SCI, card number 13779)

Member of the American Chemical Society (ACS, card number 30617260)

Member of the Italian Society of Toxicology (Sitox)

Member of the Italian Society of Phytochemistry (SIF)

Angela Tartaglia, Dr., obtained her master's degree in Pharmacy in March 2018 from the University "G. d'Annunzio" of Chieti (Italy) with her thesis on "FPSE–HPLC–DAD Method for the Quantification of Anticancer Drugs in Human Whole Blood, Plasma, and Urine". From March 2018 to June 2018, she conducted research at the Aristotle University of Thessaloniki, School of Chemistry, Laboratory of Analytical Chemistry with additional work on microextraction procedures and method validation focused on food products. She is currently a PhD candidate at University "G. d'Annunzio", Chieti (Italy) since beginning in November 2018.

Her PhD research activity is focused on the optimization of new protocols for sample preparation (MEPS, FPSE, MIP) and validation of new analytical methods for the qualitative and quantitative determination of small drugs in complex matrices, mostly biological fluids (plasma, urine, whole blood, saliva). Another line of research regards natural products and the study of biologically active products from plants with beneficial properties for human health, mostly phenolic compounds, due to their positive impact on health through reducing the risk of cardiovascular disease, neurodegenerative disorders, and cancer, among others.

She is a member of Italian Chemical Society and Italian Society of Toxicology (Sitox).

Dora Melucci, Prof., obtained her Master of Science in Chemistry at the University of Bologna with her thesis entitled "Determination of Polypeptides by HPLC". She then obtained her master's in Chemical Methodologies for Control and Analysis at the University of Bologna with a thesis entitled "The Gravitational Field-Flow Fractionation Technique (GrFFF). Fractionation and Absolute Quantitative Analysis of Particulate in Dispersion". Finally, she obtained her PhD in Chemical Sciences at the University of Ferrara with a thesis entitled "Characterization of Polymers by Means of Thermal Field-Flow Fractionation (ThFFF) Using Decalin as a Solvent".

She has served as Researcher and Assistant Professor (branch Analytical Chemistry) at the Department of Chemistry "Giacomo Ciamician", School of Sciences, University of Bologna, since her appointment in 1999.

Her research interests includethe following. Separation Science: FFF of macromolecules in solution and dispersed micro-particles. Standardless and absolute analysis in ETA-AAS and HPLC. ThFFF of industrial polymers in collaboration with industry. Flow-FFF and GrFFF of real samples (starch, yeast, metal nanoparticles for biostatic materials), miniaturization of the separation tool, cell-sorting, hyphenation of FFF with chemiluminescence, development of multianalyte competitive immunoenzymatic methods using dispersed nano- particles and microparticles. In the framework of these subjects, she was the Coordinator of the Local Unity of Bologna in a national project (PRIN) entitled "Microfluidic Separation of Nanosystems".

Since 2005, she started an autonomous research line, encapsulated in the basic keyword chemometrics. More specifically, this includes the application of chemometrics to the development and validation of innovative analytical methodologies, with focus on direct and non-altering methods of analysis. The fields of application are food, environment, pharmaceutics, forensics, biotechnology, and cultural heritage. The analytical techniques employed are AAS, NIR, voltammetry, Raman, GC, LC–MS, FT-IR, and XRD. Up to May 2020, her work has produced 76 articles, 20 contributions in books, and over 100 communications in national and international meetings, with h-index = 17.

Her teaching experience in 2002–2020 includes teaching courses in Analytical Chemistry, Laboratory of Analytical Chemistry, Analytical Chemistry and Law, Principles of Quality and Safety (Bachelor of Chemistry); Chemometrics (Master of Science in Chemistry); Chemometrics for Forensic Analysis (Master in Forensic Chemical and Chemical–Toxicological Analysis).

Her additional academic roles include Local Coordinator for the University of Bologna of the National Ministerial Project for high-school student guidance (Scientific Degrees Plan) and Department Delegate for university student tutoring.

Abuzar Kabir, Dr., is Research Assistant Professor at the Department of Chemistry and Biochemistry, Florida International University, Miami, Florida, USA. His research focuses on the synthesis, characterization, and applications of novel sol–gel-derived advanced material systems in the form of chromatographic stationary phases, surface coatings of high-efficiency microextraction sorbents, nanoparticles, microporous and mesoporous functionalized sorbents, and molecularly imprinted polymers for analyzing trace and ultra-trace level concentrations of polar, medium polar, nonpolar, ionic analytes, heavy metals, and organometallic pollutants from complex sample matrices. His inventions fabric phase sorptive extraction (FPSE), dynamic fabric phase sorptive extraction (DFPSE), capsule phase microextraction (CPME), molecular imprinting technology, superpolar sorbents, in-vial microextraction (IVME), sol–gel-based reversed phase LC stationary phases and SPE sorbents, organic polymeric LC stationary phases and SPE sorbents, synthesis of mesoporous silica and its application in reversed phase LC stationary phases and SPE sorbents have drawn global attention. He has developed and formulated numerous high efficiency sol–gel hybrid inorganic–organic sorbents based on silicon, titanium, zirconium, tantalum, and germanium chemistries. Dr. Kabir has authored 18 patents, 10 book chapters, 70 journal articles and 125 conference papers. His recent inventions, Biofluid Sampler (BFS) and Universal Biofluid Sampler (UBFS) are capable of handling whole blood (5–1000 μL) without any sample pre-treatment for chromatographic separation and analysis. These technologies will likely change the current practices of blood analysis in the near future.

Google Scholar Link: https://scholar.google.com/citations?user=ovZ73-UAAAAJ&hl=en

ResearchGate Link: https://www.researchgate.net/profile/Abuzar_Kabir

Halil Ibrahim Ulusoy (Full Professor, Analytical Chemistry), Prof. Dr., received his master's degree in 2007 and his doctorate degree in 2012 on the field of Analytical Chemistry. He has served as Full Professor of Analytical Chemistry at the Cumhuriyet University, Faculty of Pharmacy (Sivas/TURKEY) since 2015. He is a member of the Turkish Chemical Society. His research interests are in the development of new analytical methodologies for trace organic and inorganic species in the food samples, pharmaceutical samples, and biological matrices.

His research activity is devoted to the development and validation of chromatographic and spectroscopic methods for trace determination of biologically active molecules and elements in complex matrices cosmetics, foods, and environmental samples. Recently, particular attention of his academic studies is focused on easy applicable and reliable determination drug active ingredients such as antibiotics, antidepressants, pesticides, and vitamins.

He has authored more than 65 manuscripts, 96 congress communications, and 5 chapters in scientific books in addition to serving as Guest Editor of 4 Special Issues with attested scientific activity (h-index = 18, 55 papers, and 653 citations, based on Scopus (9th of June 2020)). He is Editorial Board Member of the journals *Current Analytical Chemistry*, *Journal of Quality Assurance and*

Pharma Analysis (IJQAPA), *Pharmaceutica Analytica Acta*, and *Asian Journal of Medicinal and Analytical Chemistry*. He is one of the Editors-in-Chief of *Cumhuriyet Science Journal*.

Victoria Samanidou, Professor (Analytical Chemistry), Dr., was born on the 11th of January 1963, in Thessaloniki, Greece. She obtained her Bachelor of Science degree in Chemistry, in 1985, from the Chemistry Department of Aristotle University of Thessaloniki, Greece. From 20-7-86 to 25-8-86, she was at the Institute of Ecological Chemistry, in GSF, Attaching/Freising, Germany, to conduct additional work on her PhD thesis as well as research work on photochemistry and the study of photodecomposition products of chlorophenols by HPLC–diode array and GC–MS. From 15-7-87 to 4-9-87, she was at the Institute of Ecological Chemistry, in GSF, Neuherberg-Munich, Germany, to conduct additional work on her PhD thesis, as well as research on carbamate analysis by HPLC and GC–MS. From 1-7-88 to 30-9-88, she joined the Institute of Ecological Chemistry, in GSF, Neuherberg-Munich, Germany, to conduct additional work on her PhD thesis as well as research work on the controlled release of pesticides by HPLC and GC–MS.

In 1990, she obtained a doctorate (PhD) in Chemistry from the Department of Chemistry of the Aristotle University of Thessaloniki. The topic of her thesis was "Distribution and Mobilization of Heavy Metals in Waters and Sediments from Rivers in Northern Greece". In the same year, Dr. Samanidou joined the Laboratory of Analytical Chemistry at the Department of Chemistry, Aristotle University of Thessaloniki, as Technical Assistant. Nine years later, she was appointed as Lecturer in the Laboratory of Analytical Chemistry in the Department of Chemistry of the Aristotle University of Thessaloniki. In 2007, she joined the Institute of Analytical Chemistry and Radiochemistry in Graz Technical University for four months, developing methods by LC–MS/MS.

Since 2015, Dr. Samanidou has been Full Professor in the Department of Chemistry, Aristotle University of Thessaloniki, Greece, where she currently serves as Director of the Laboratory of Analytical Chemistry.

Dr. Samanidou has authored and co-authored more than 170 original research articles and 45 reviews in peer-reviewed journals as well as 50 chapters in scientific books, with an h-index = 36 (Scopus June 2020, http://orcid.org/0000-0002-8493-1106, Scopus Author ID 7003896015) and ca. 3500 citations. She has supervised four PhD theses, 24 Postgraduate Diploma theses, 2 postdoctoral researchers, and more than 15 undergraduate Diploma theses. She has served as Member of 10 advisory PhD committees, 21 examination PhD committees, and 32 examination committees of postgraduate Diploma theses. She is Editorial Board Member of over 10 scientific journals and has reviewed ca. 500 manuscripts for more than 100 scientific journals. She has also served as Guest Editor of more than 10 Special Issues of scientific journals. She has served as Academic Editor for Separations (MDPI), as Regional Editor of Current Analytical Chemistry, and as Editor-in-Chief of Pharmaceutica Analytica Acta.

Her research interests include: 1. Development and validation of analytical methods for the determination of inorganic and organic substances using chromatographic techniques. 2. Development and optimization of methodology for sample preparation of various samples, e.g., food, biological fluids, etc. 3. Study of new chromatographic materials used in separation and sample preparation (polymeric sorbents, monoliths, carbon nanotubes, fused core particles, etc.) compared to conventional materials.

She has also been a member of the organizing and scientific committee for 20 scientific conferences.

Dr. Samanidou has been serving as President of the Steering Committee of the Division of Central and Western Macedonia of the Greek Chemists' Association since being elected in December 2015. In November 2018, she was re-elected to serve in the same leading position for an additional term of 3 years.

A milestone in her career occurred in 2016, when she was included in top 50 Power List of women in Analytical Science, as proposed by Texere Publishers.

https://theanalyticalscientist.com/power-list/the-power-list-2016

Preface to ”Advances in Chemical Analysis Procedures (Part II)”

Analytical chemistry deals with both qualitative and quantitative measurements, although modern approaches are more inclined towards quantitative science. In analytical laboratories, the measurements are usually made on a small group of representative samples to determine the presence and concentration of target analytes. Following data collection, the results are tabulated to evaluate the quality of the data. An important area in evaluating analytical data is represented by statistical approaches, which should not be considered only for evaluating the results of experiments, but also in the planning and design of experiments. The design and optimization process should include the identification of those experimental factors and then combine them in an optimal way to obtain the best sensitivity and selectivity among other factors. The major quantitative chemical problems can also be performed with chemometric measurements. The starting point of multivariate measurements is usually represented by principal component analysis (PCA), which can reduce the dimensionality of the data, eliminate false information, search for outliers, and more. The modern tools for various measurements are completely devoid of manual controls and are controlled by personal computers that record and manage the obtained data. In recent years, appreciable progress has been made, and in the most modern analytical chemistry laboratories, instruments not only allow quick and precise data calculations but also include instrument performance control and reporting of any malfunctions.

Marcello Locatelli, Angela Tartaglia, Dora Melucci,
Abuzar Kabir, Halil Ibrahim Ulusoy, Victoria Samanidou
Editors

molecules

MDPI

Article

Identification of Metabolites of Eupatorin In Vivo and In Vitro Based on UHPLC-Q-TOF-MS/MS

Luya Li [1], Yuting Chen [1], Xue Feng [1], Jintuo Yin [1], Shenghao Li [2], Yupeng Sun [1] and Lantong Zhang [1,*]

[1] School of Pharmacy, Hebei Medical University, Shijiazhuang 050017, China
[2] School of Pharmacy, Hebei University of Chinese Medicine, Shijiazhuang 050000, China
* Correspondence: zhanglantong@263.net or zhanglantong@hebmu.edu.cn; Tel./Fax: +86-311-8626-6419

Academic Editor: Marcello Locatelli
Received: 23 June 2019; Accepted: 19 July 2019; Published: 23 July 2019

Abstract: Eupatorin is the major bioactive component of Java tea (*Orthosiphon stamineus*), exhibiting strong anticancer and anti-inflammatory activities. However, no research on the metabolism of eupatorin has been reported to date. In the present study, ultra-high-performance liquid chromatography coupled with hybrid triple quadrupole time-of-flight mass spectrometry (UHPLC-Q-TOF-MS) combined with an efficient online data acquisition and a multiple data processing method were developed for metabolite identification in vivo (rat plasma, bile, urine and feces) and in vitro (rat liver microsomes and intestinal flora). A total of 51 metabolites in vivo, 60 metabolites in vitro were structurally characterized. The loss of CH_2, CH_2O, O, CO, oxidation, methylation, glucuronidation, sulfate conjugation, N-acetylation, hydrogenation, ketone formation, glycine conjugation, glutamine conjugation and glucose conjugation were the main metabolic pathways of eupatorin. This was the first identification of metabolites of eupatorin in vivo and in vitro and it will provide reference and valuable evidence for further development of new pharmaceuticals and pharmacological mechanisms.

Keywords: eupatorin; UHPLC-Q-TOF-MS/MS; metabolism; in vivo *and in vitro*; rat liver microsomes; rat intestinal flora

1. Introduction

Eupatorin (5,3′-di-hydroxy-6,7,4′-tri-methoxy-flavone, Figure 1), belonging to the natural methoxyflavone compound, is widely found in Java tea (*Orthosiphon stamineus*, OS) which is a popular medicinal herb used in traditional Chinese medicine as a diuretic agent and for renal system disorders in Southeast Asia and European countries [1–3]. OS has gained a great interest nowadays due to its wide range of pharmacological effects such as antibacterial, antioxidant, hepatoprotection, antidiabetic, anti-hypertension, anti-inflammatory and antiproliferative activities [4–9]. Eupatorin, as a major bioactive flavonoid constituent in OS possesses numerous strong biological activities, including anticancer, anti-inflammatory and vasorelaxation activities [10–17]. Its anticancer activities have attracted more and more attention and it was expected to be developed as a cancer chemopreventive and as an adjuvant chemotherapeutic agent. Although there is literature on the qualitative and quantification profile of eupatorin in OS [6], the metabolism study of eupatorin has not been studied to date, which was necessary for the exploration of the biological activity and the clinical therapeutic effect of eupatorin. Thus, an investigation is essential to explore the identification of metabolites of eupatorin for further understanding of its biological activities.

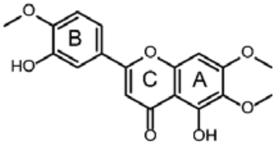


Figure 1. Chemical structure of eupatorin.

To the best of our knowledge, a series of biotransformations will occur when drugs are orally taken into the body, there are four aspects of pharmacological consequences in these biotransformation processes: (1) Transforming into inactive substances; (2) transforming the drug with no pharmacological activity into active metabolites; (3) changing the types of pharmacological actions of drugs; (4) and producing toxic substances [18]. Therefore, it is extremely crucial to study the metabolism of drugs in vivo to make sure of safety of use. In addition, as the main metabolic organ of the human body, the liver is rich in enzymes, especially cytochrome P450 enzymes, which are closely related to the biological transformation of drugs [19]. Furthermore, the gastrointestinal tract is also a vital place for drug metabolism, and its intestinal flora have a significant impact on drug absorption, metabolism and toxicology [20,21]. Hence, in this paper, mass spectrometry was employed to investigate the metabolism of eupatorin in rats, liver microsomes and intestinal flora, in order to characterize the metabolites and structural information of the products, which will lay a foundation for further studies on the safety and efficacy of metabolites and will provide greater possibilities for the development of new drugs.

With the development of technology, a quadrupole time-of-flight mass spectrometry has been widely used as a reliable analytical technique to detect metabolites due to its advantages of high resolution, high sensitivity, high-efficiency separation and accurate quality measurement [22,23]. In this study, high-sensitivity ultra-high-performance liquid chromatography coupled with hybrid triple quadrupole time-of-flight mass spectrometry (UPLC-Q-TOF-MS) full scan mode, electrospray ionization (ESI) source negative ion mode monitoring combined with multiple mass loss (MMDF) and dynamic background subtraction (DBS) were employed to collect data online. Correspondingly, multiple data processing methods were applied by using PeakView 2.0 and MetabolitePilot 2.0.4 software developed by AB SCIEX company, including a variety of data handing functions such as the extraction of ion chromatograms (XIC), mass defect filter (MDF), product ion filter (PIF) and neutral loss filtering (NLF), which provided accurate secondary mass spectral information [24]. Based on the above methods, the metabolic pathways of eupatorin were explored and summarized for the first time and 51 metabolites in vivo and 60 metabolites in vitro were finally identified. These metabolic studies are important parts of drug discovery and development and can also provide a basis for further pharmacological research.

2. Results and Discussion

2.1. Analytical Strategy

In this study, UHPLC-Q-TOF-MS/MS combined with an online data acquisition and multifarious processing methods was adopted to systematically identify the metabolites of eupatorin in vivo and in vitro.

The workflow of the analytic procedure was segmented into three steps. First, an online full-scan data acquisition was performed based on the MMDF and DBS to collect data online and to capture all potential metabolites. Next, a multiple data processing method was employed by using PeakView 2.0 and MetabolitePilot 2.0.4 software, which contained many data-processing tools such as XIC, MDF, PIF and NIF, these provided accurate MS/MS information to determine the metabolites of eupatorin. Finally, plenty of metabolites were identified according to accurate mass datasets, specific secondary mass spectrometry information and so on. With regard to the isomers of metabolites, Clog P values calculated by ChemDraw 14.0 were used to further distinguish them. Generally speaking, the larger the Clog P value, the longer the retention time will be in the reversed-phase chromatography system [25–27].

2.2. Mass Fragmentation Behavior of Eupatorin

In order to identify the metabolites of eupatorin, it is of significance to understand the pyrolysis of parent drug (M0). The chromatographic and mass spectrometric behaviors of eupatorin were explored in the negative ESI scan mode by UHPLC-Q-TOF-MS. Eupatorin ($C_{18}H_{16}O_7$) was eluted at 12.22 min and yielded at 343.0821 $[M-H]^-$. The characteristic fragment ions of M0 at *m/z* 328.0585, 313.0348, 298.0111, 285.0398, 270.0160, 267.0285, 254.0217, 241.0503, 221.0434, 147.0461, 132.0214 were detected according to the MS/MS spectrum. Fragment ions at *m/z* 328.0585, 313.0348, 298.0111, 270.0160 and 254.0217 were generated by M0 through losing CH_3, CH_3, CH_3, CO and O continuously. The ion at *m/z* 343.0821 yielded other representative fragment ions at *m/z* 267.0285, 241.0503 and 221.0434 by loss of CO_2 and 2O, $C_4H_6O_3$, $C_7H_6O_2$, respectively. The product ion at *m/z* 285.0398 was created by dropping CO from the ion at *m/z* 313.0348. Last but not the least, the conspicuous product ion at *m/z* 147.0461 was formed because of the Retro-Diels-Alder (RDA) reaction in ring C of the flavonoid, which gained the ion at *m/z* 132.0214 by loss of CH_3 [28]. The MS/MS spectrum and the fragmentation pathways of eupatorin are shown in Figure 2.

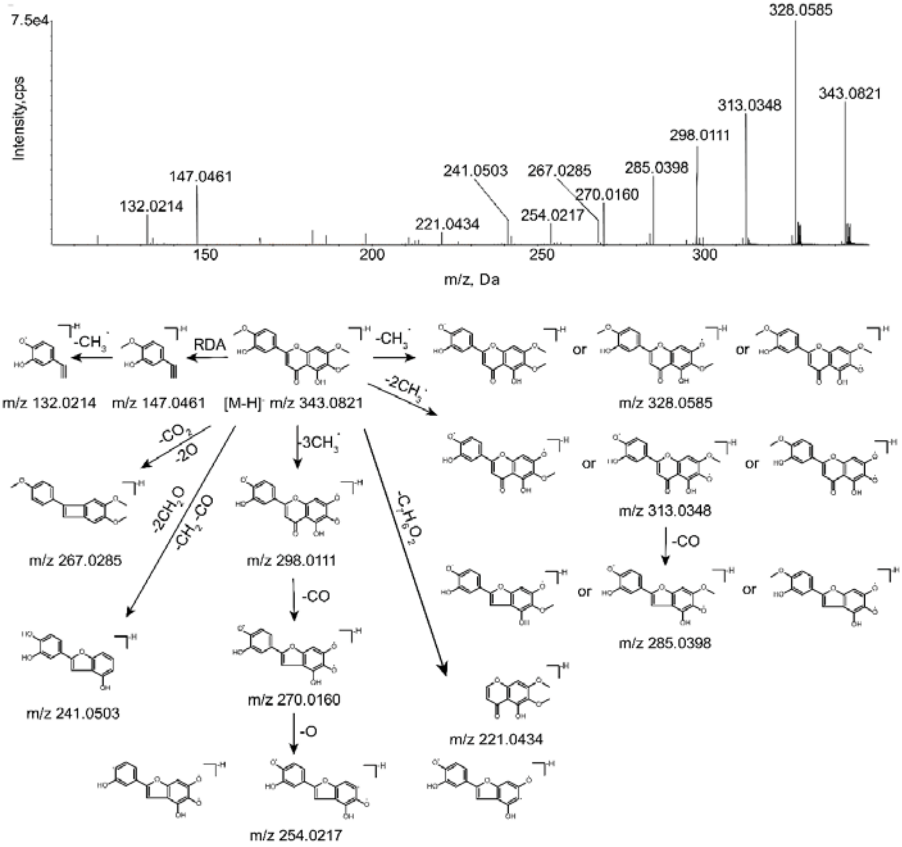


Figure 2. MS/MS spectrum of eupatorin and its predominant fragmentation pathways.

2.3. Identification of Metabolites in Vivo and in Vitro

Metabolites M1, M2 and M3 ($C_{17}H_{14}O_7$) were isomers with the deprotonated molecular ions [M-H]$^-$ at *m*/*z* 329.0660, 329.0668 and 329.0662, which were 14 Da (CH_2) lower than that of M0. They were eluted at 9.93 min, 10.27 min and 10.79 min, respectively. In the MS/MS spectrum, product ions at *m*/*z* 314.0427, 313.0384, 299.0188 and 285.0371 were formed after losing CH_3, O, $2CH_3$ and CO_2, respectively. The prominent fragment ion at *m*/*z* 133.0287 created after the RDA reaction was 14 Da lower than the ion *m*/*z* 147.0461 of the parent drug, suggested that CH_2 was lost at the methoxy group at 4′position. At the same time, the fragment ions at *m*/*z* 207.7129 and 207.7166 were 14 Da lower than that of M0, which showed that the loss of CH_2 occurred at the methoxy group at 6 or 7 position of A ring. Additionally, the Clog P values of M1, M2 and M3 were 2.26422, 2.26434, 2.51422, respectively. Therefore, M1–M3 were illustrated according to the above information.

Metabolites M4 and M5 ($C_{16}H_{12}O_7$) were eluted at 7.26 and 8.50 min, with the deprotonated molecular ions [M-H]$^-$ at *m*/*z* 315.0500 and 315.0504, 28 Da (C_2H_4) lower than that of the parent drug, which indicated that it lost $2CH_2$. Fragment ions at *m*/*z* 300.0279 and 297.1740 were generated by loss of CH_3 and H_2O, respectively. The product ion at *m*/*z* 269.1760 was obtained through dropping CO from the ion at *m*/*z* 297.1740. According to the dominant fragment ion at *m*/*z* 133.0270 gained by the RDA reaction, loss of CH_2 and CH_2 occurred at the position of 4′, 6 or 4′, 7. In addition, the distinctive ion at *m*/*z* 147.0821 was similar with that of the parent drug, which implied that the reaction occurred at the position of 6 and 7.

Metabolite M6 ($C_{17}H_{14}O_6$) was obtained with a peak at *m*/*z* 313.0713 in the UPLC system, which was eluted at 13.86 min, 30 Da (CH_2O) lower than that of eupatorin. Prominent fragment ions at *m*/*z* 298.0483 and 283.0250 were created by dropping CH_3 and CH_3 successively. In addition, the characteristic fragment ions at *m*/*z* 117.0364 was produced by RDA reaction, which was 30 Da lower than that of M0, showing that loss of CH_2O occurred at the position of 4′. Similarly, the product ion at *m*/*z* 147.0078 was consistent with M0, indicating that loss of CH_2O occurred at the position of 6 or 7. Thus, it was speculated that it may have three missing CH_2O sites.

Metabolite M7 ($C_{16}H_{12}O_6$) was detected at 10.10 min and exhibited the molecular ion [M-H]$^-$ at *m*/*z* 299.0562, which was 44 Da lower than that of M0. Based on the information of chemical elements and software provided, it indicated that M7 lost CH_2O and CH_2. Crucial fragment ions at *m*/*z* 284.0326 and 251.1281 were obtained by loss of CH_3 and 3O from M7, respectively. Furthermore, M7 had common fragment ion at *m*/*z* 146.9687 with that of the parent drug, it is equally important that the noteworthy fragment ion at *m*/*z* 281.1787 was generated by loss of H_2O from M7, which implied that loss of CH_2O and CH_2 occurred at the position of 7 or 6, respectively. Hence, it was identified.

Metabolite M8 ($C_{16}H_{12}O_5$) was eluted at 13.60 min, which displayed deprotonated molecular ion [M-H]$^-$ at *m*/*z* 283.0614, 60 Da ($C_2H_4O_2$) lower than that of the parent drug. Fragment ions at *m*/*z* 268.0379 and 240.0428 were produced by dropping CH_3 and CO continuously from *m*/*z* 283.0614. In addition, the dominant fragment ion at *m*/*z* 146.9655 was consistent with that of the parent drug, while the diagnostic fragment ion at *m*/*z* 161.0025 was 60 Da lower than 221.0434 of M0, these suggested that loss of CH_2O and CH_2O reaction happened at C-6 and C-7 of A ring. So, the structure of M8 could be inferred.

Metabolites M9 and M10 ($C_{18}H_{16}O_6$) appeared as deprotonated molecular ions [M-H]$^-$ at *m*/*z* 327.0882 and 327.0872, together with the retention time of 4.98 min and 7.47 min, respectively, which were 16 Da lower than M0, suggesting they lacked one oxygen atom compared with the parent. The MS/MS spectra showed the fragment ions at *m*/*z* 309.0800, 299.0957 and 281.2489, which were created by loss of O, CO and C_2H_6O, respectively. In addition, M9 had common fragment ion at *m*/*z* 146.9380 with that of the parent drug, and meanwhile the characteristic fragment ion at *m*/*z* 205.0025 was 16 Da lower than 221.0434 of M0, which implied that loss of O occurred at C-5 of A ring. Nevertheless, the ion at *m*/*z* 130.9716 gained after the RDA cleavage was 16 Da lower than that of the parent drug, showing that loss of O occurred at C-4′ of B ring. Therefore, the structures of metabolites

M9 and M10 were determined. Moreover, they were also validated with the Clog P values of M9 and M10 which were 2.45814 and 3.44497, respectively.

Metabolite M11 ($C_{18}H_{16}O_5$) was turned up in the chromatogram at 9.55 min with the deprotonated molecular ion at *m/z* 311.0930 $[M-H]^-$ and was 32 Da less than that of M0, suggesting that the loss of two oxygen atoms reaction took place. A series of diagnostic product ions at *m/z* 250.9816, 204.9868 and 130.9658 were yielded by loss of $C_2H_4O_2$, C_7H_6O and RDA reaction. In addition, the product ion at *m/z* 174.9556 was obtained through dropping CH_2O from the ion at *m/z* 204.9868. According to the above characteristic fragment ions and analysis, loss of O and O occurred at C-5 and C-3′.

Metabolite M12 ($C_{17}H_{14}O_5$), the deprotonated molecular ion of *m/z* 297.0768 was observed at the retention time of 7.33 min and was 46 Da lower than that of eupatorin. According to its secondary mass spectrum and the information software provided, implying that M12 lost O and CH_2O. Fragment ions at *m/z* 267.1016, 253.0865, 175.0394 and 147.0452 were produced by loss of CH_2O, CO_2, $C_7H_6O_2$ and RDA reaction. It was important that the typical ion at *m/z* 147.0452 was similar with the fragment ion at *m/z* 147.0461 of the parent drug, together with the dominant fragment ion at *m/z* 175.0394, 46 Da lower than that of M0, all of which indicated that the reaction was likely to occur in the A ring. Above all, loss of O happened at the hydroxyl group at the 5 position, while loss of CH_2O occurred at the methoxy group at 6 or 7 position.

Metabolite M13 ($C_{17}H_{16}O_6$) exhibited a sharp peak at an elution time of 12.74 min in the XIC with a deprotonated ion at *m/z* 315.0862 and it was 28 Da (CO) less than eupatorin. Product ions at *m/z* 300.0633, 285.0401 and 270.0144 were formed after dropping CH_3 continuously. In addition, the MS^2 spectrum of M13 presented other vital fragment ions at *m/z* 193.0503 and 147.0445 by losing $C_7H_6O_2$ and undergoing RDA reaction.

Metabolites M14, M15, M16 and M17 ($C_{18}H_{16}O_8$): Four chromatographic peaks were eluted at 10.01 min, 10.50 min, 11.47 min and 12.23 min with deprotonated molecular ions $[M-H]^-$ at *m/z* 359.0772, 359.0768, 359.0767 and 359.0767, which were 16 Da (O) higher than that of eupatorin. Characteristic ions at *m/z* 344.0542, 329.0304, and 314.0064 were obtained by loss of CH_3 successively. Furthermore, noteworthy fragment ions at *m/z* 221.0098 and 163.0368 were produced by loss of $C_7H_6O_3$ and RDA reaction. The ion at *m/z* 163.0368 was 16 Da (O) larger than *m/z* 147.0461, showing that oxidation occurred at C-2′, C-5′ or C-6′ of B ring. However, the prominent ion at *m/z* 147.0130 was similar with the fragment ion at *m/z* 147.0461 of the parent drug, indicating that the reaction happened at the position of 8 in the A ring. The Clog P values of M14-M17 were 1.79518, 1.84518, 1.86518 and 1.87123, respectively. Thus, M14-M17 were characterized by comparing the different values of Clog P.

Metabolite M18 ($C_{18}H_{16}O_9$), the deprotonated molecular ion of *m/z* 375.0709 was observed at the retention time of 9.90 min, which was 32 Da (2O) higher than that of eupatorin. A series of product ions at *m/z* 329.0669, 221.1216 and 178.9947 were detected by loss of CH_2O_2, $C_7H_6O_4$ and RDA reaction in its secondary mass spectrum. Product ions at *m/z* 314.0434 and 299.0191 were produced by losing CH_3 and CH_3 continuously from the ion at *m/z* 329.0669. What's more, the key fragment ions at *m/z* 178.9947 was 32 Da higher than 147.0461 of eupatorin, implying that di-oxidation reaction occurred in the B ring, then M18 was identified.

Metabolite M19 ($C_{18}H_{16}O_{10}$) was detected at 12.26 min and showed the deprotonated molecular ion $[M-H]^-$ at *m/z* 391.0673, 48 Da (3O) higher than that of the parent drug, which contained the fragment ions at *m/z* 345.0869, 330.0636 and 315.0393 by loss of CH_2O_2, CH_3 and CH_3 continuously. More importantly, distinctive fragment ions at *m/z* 221.0399 and 195.0289 were created by loss of $C_7H_6O_5$ and RDA reaction. The pivotal fragment ions at *m/z* 195.0289 was 48 Da higher than 147.0461 of the parent drug, suggesting that tri-oxidation happened at C-2′, C-5′ and C-6′ of B ring. Hence, M19 was recognized.

Metabolites M20 and M21 ($C_{17}H_{14}O_8$) were eluted at 9.43 min and 10.29 min, with the deprotonated molecular ions $[M-H]^-$ at *m/z* 345.0605 and 345.0606 and were increased by 2 Da compared with M0, indicating that it carried out demethylation and oxidation reaction. The representative secondary fragment ions at *m/z* 330.0384, 301.0719, 221.0028, 125.0311 and 149.0234 generated by the loss of CH_3,

CO_2, $C_6H_4O_3$, $C_{11}H_8O_5$ and RDA reaction implied that demethylation and oxidation occurred in ring B. Furthermore, the Clog P values of M20 and M21 were 1.5017 and 1.59734, respectively, so their structures were identified.

Metabolites M22 and M23 ($C_{19}H_{18}O_7$) were obtained in the extracted chromatogram at *m*/*z* 357.0972 and 357.0969 with the retention time of 10.02 min and 12.86 min, which were 14 Da (CH_2) higher than that of eupatorin. The diagnostic fragment ions at *m*/*z* 342.0740, 327.0503, 312.0266 and 297.0033 were attributed to the loss of CH_3 successively. In addition, because of the prominent fragment ions at *m*/*z* 235.0434 and 147.0433 obtained after RDA reaction, it was proposed that methylation happened at hydroxyl group at 5 position. Nevertheless, the fragment ion at *m*/*z* 161.0269 was 14 Da higher than 147.0461 of eupatorin, indicating that it occurred at C-3′ of B ring. Furthermore, the Clog P values of M22 and M23 were 2.06632 and 3.18323, respectively, so they were verified.

Metabolites M24 and M25 ($C_{19}H_{18}O_6$) were eluted at 7.15 min and 8.79 min, respectively. They had the deprotonated molecular ions $[M-H]^-$ at *m*/*z* 341.1025 and 341.1027, which were 2 Da lower than that of eupatorin, presumably they occurred a loss of O and a methylation reaction. The distinctive fragment ion at *m*/*z* 130.9906 was 16 Da (O) lower than 147.0461 of M0, along with fragment ions at *m*/*z* 235.0607 and 107.0440 produced by loss of C_7H_6O and $C_{12}H_{10}O_5$, implying that the loss of O occurred at the hydroxyl group at C-3′, while methylation happened at the hydroxyl group at C-5. Similarly, according to the representative fragment ion at *m*/*z* 161.0595, 14 Da (CH_2) lower than 147.0461 of M0 and the diagnostic fragment ions at *m*/*z* 204.9196 and 137.0553 obtained by loss of $C_8H_8O_2$ and $C_{11}H_8O_4$, the loss of O occurred at the hydroxyl group at 5 position, while methylation took place at the hydroxyl group at 3′ position. In addition, they were also validated with the Clog P values of M24 and M25 which were 2.80306 and 2.9313, respectively.

Metabolite M26 ($C_{15}H_{10}O_5$), displayed a peak at 9.89 min, as well as a deprotonated molecular ion $[M-H]^-$ at *m*/*z* 269.0459, 14 Da (CH_2) lower than that of M8, suggesting that demethylation occurred on the basis of M8. The fragment ions at *m*/*z* 253.0124, 241.0500, 225.0555 and 133.0298 were attributed to the loss of O, CO, CO_2 and RDA reaction, which was 14 Da (CH_2) lower than that of M0 and M8, implying that demethylation occurred at the methoxy group at 4′ position. Like the M8, the loss of CH_2O and CH_2O took place at C-6 and C-7 of A ring.

Metabolite M27 ($C_{15}H_{10}O_6$) was detected at a retention time of 8.45 min with the deprotonated molecular ion $[M-H]^-$ at *m*/*z* 285.0402, 14 Da (CH_2) lower than that of M7, indicating that M27 was demethylated on the basis of M7. Product ions at *m*/*z* 267.0130, 241.0462, 221.0063, 177.0189 and 133.0307 were produced by loss of H_2O, CO_2, 4O, $C_6H_4O_2$ and RDA reaction which was 14 Da (CH_2) lower than that of M0 and M7, it means demethylation happened at the methoxy group at 4′ position. Similar to M7, loss of CH_2O and CH_2 occurred at C-7 and C-6 of A ring, respectively.

Metabolites M28 and M29 ($C_{24}H_{24}O_{13}$) arose as deprotonated molecular ions $[M-H]^-$ at *m*/*z* 519.1140 and 519.1151, together with the retention time of 7.10 min and 8.14 min, respectively, which were 176 Da higher than that of eupatorin, suggesting that glucuronidation was carried out. The key product ion at *m*/*z* 343.0822 was yielded by dropping a glucuronic acid. Moreover, the crucial ion at *m*/*z* 146.9662 was similar to the fragment ion at *m*/*z* 147.0461 and while the fragment ion at *m*/*z* 397.0442 was 176 Da higher than that of the parent drug, indicating that glucuronidation happened at the hydroxyl group at 5 position. Nevertheless, the prominent fragment ions at *m*/*z* 323.0173 was 176 Da larger than 147.0461 of M0, inferring that the reaction occurred at the hydroxyl group at 3′ position. Furthermore, M28 and M29 were also proved by the different Clog P values of −0.494983 and 0.621934, respectively.

Metabolite M30 ($C_{23}H_{22}O_{13}$) was detected at 7.88 min with the deprotonated molecular ion $[M-H]^-$ at *m*/*z* 505.0979, 14 Da lower than that of M28 and M29, which suggested that it occurred glucuronide conjugation and demethylation. The characteristic product ion at *m*/*z* 329.0669 was obtained by losing a glucuronic acid. The distinctive fragment ions at *m*/*z* 285.0735 and 309.0687 which was 162 Da larger than 147.0461 of M0 were attributed to the loss of $C_{11}H_8O_5$ and RDA reaction, implying glucuronide conjugation and demethylation occurred at the position of 3′ and 4′, respectively.

Metabolites M31 and M32 ($C_{18}H_{16}O_{10}S$) appeared as deprotonated molecular ions [M-H]$^-$ at *m*/*z* 423.0391 and 423.0387 with the retention time of 8.93 min and 9.20 min. S elemental was found, suggesting that it had been a sulfate bound. The characteristic product ion at *m*/*z* 343.0830 was created by the loss of SO_3. Remaining ions at *m*/*z* 328.0593, 313.0355, 285.0413 and 147.0037 were similar to the fragment ions of the parent drug, inferring that sulfate conjugation occurred at the hydroxyl group at 5 position. However, the pivotal fragment ion at *m*/*z* 227.0084 was 80 Da higher than 147.0461 of eupatorin, implying sulfate conjugation took place at the hydroxyl group at 3′ position of B ring. In addition, the Clog P values of M31 and M32 were 0.270316 and 1.38723, respectively. So, they were also validated.

Metabolite M33 ($C_{17}H_{14}O_{10}S$) was eluted at the retention time of 9.01 min on the UPLC system. Its deprotonated molecular ion [M-H]$^-$ at *m*/*z* 409.0233 lacked CH_2 compared with M31 and M32. The representative product ion at *m*/*z* 329.0670 was acquired by dropping SO_3. In addition, M33 created the dominant fragment ion at *m*/*z* 212.0456 through the RDA reaction, and the product ion at *m*/*z* 132.0208 was formed by the loss of SO_3 from it. Therefore, it might occur at the methoxy group at 4′ position.

Metabolite M34 ($C_{18}H_{16}O_9S$) was eluted at a retention time of 12.65 min. The MS/MS spectrum of M34 showed the deprotonated molecular ion [M-H]$^-$ at *m*/*z* 407.0434, lacked one oxygen atom compared with M31 and M32. The crucial fragment ions at *m*/*z* 327.0826, 301.0034 and 131.0573 were attributed to the loss of SO_3, C_7H_6O and RDA reaction. In addition, the product ion at *m*/*z* 220.9818 was acquired by the loss of SO_3 from the fragment ions at *m*/*z* 301.0034. Based on the information above, the loss of O and sulfate conjugation might occur at the hydroxyl group at 3′ and 5 position, respectively.

Metabolites M35 and M36 ($C_{20}H_{18}O_8$): Two isomers were simultaneously extracted in the XIC at 13.36 and 13.90 min and were detected at *m*/*z* 385.0917 and 385.0925, respectively. The noteworthy ion at *m*/*z* 343.0846 was yielded by the loss of acetyl. In M36, the diagnostic fragment ion at *m*/*z* 189.0551 generated by RDA reaction, which was 42 Da higher than 147.0461 of eupatorin and the distinctive fragment ion at *m*/*z* 221.0781 indicated that acetylation reaction happened at the hydroxyl group at 3′ position. Likewise, according to the prominent fragment ions at *m*/*z* 263.0551 and 147.0513, the acetylation reaction happened at the hydroxyl group at position 5 of M35. In addition, Clog P values of M35 and M36 were 1.49632 and 2.61323, respectively, which could also support the confirmation of the structures.

Metabolites M37 and M38 ($C_{20}H_{18}O_7$) were observed at 13.02 and 13.90 min in the XIC and were detected at *m*/*z* 369.0987 and 369.0975 in the mass spectra, respectively, which were decreased by 16 Da (O) compared with M35 and M36. The typical fragment ion at *m*/*z* 130.9934, 16 Da lower than 147.0461 of eupatorin, together with the representative fragment ion at *m*/*z* 263.1681, 42 Da higher than that of eupatorin, inferring that loss of O and acetylation reaction occurred at the hydroxyl group at 3′and 5 position, respectively. Similarly, based on the crucial product ions at *m*/*z* 174.9586 and 164.9289, loss of O and acetylation reaction occurred at the hydroxyl group at 5 and 3′ position, respectively. In addition, M37 and M38 were also verified based on their Clog P values of 2.23306 and 2.3613, respectively.

Metabolite M39 ($C_{19}H_{16}O_8$) detected at *m*/*z* 371.0761 and eluted at 11.45 min. In addition, it was 14 Da (CH_2) smaller than the size of M35 and M36. The characteristic ion at *m*/*z* 329.0680 was yielded by dropping of acetyl. The prominent fragment ion at *m*/*z* 175.0389, 28 Da larger than 147.0461 of eupatorin, implying that the loss of CH_2 and acetylation reaction took place at the methoxy group at 4′ position of B ring.

Metabolites M40 and M41 ($C_{19}H_{16}O_7$) were detected as deprotonated [M-H]$^-$ ion at *m*/*z* 355.0813 and 355.0814, which were eluted at 11.86 min and 13.15 min, 30 Da (CH_2O) lower than that of M35 and M36. In M40, the diagnostic fragment ion at *m*/*z* 117.0329 produced by RDA reaction was 30 less than 147.0461 of the parent drug, while the fragment ion at *m*/*z* 263.0361 was 42 higher than 221.0434 of eupatorin, indicating that loss of CH_2O and acetylation reaction occurred at the position of 4′ and 5, respectively. However, in M41, the distinctive fragment ion at *m*/*z* 159.10931 was 12 higher than 147.0461 of eupatorin, the crucial fragment ion at *m*/*z* 221.0027 was similar to fragment ion at *m*/*z*

221.0434, suggesting that the loss of CH_2O still occurred at the methoxy group at 4′ position while acetylation reaction took place at the hydroxyl group at 3′ position. The respective Clog P values were 1.64857 and 2.87497, so M40 and M41 were ensured.

Metabolite M42 ($C_{21}H_{18}O_8$) with the $[M-H]^-$ ion of *m/z* 397.0918, which was eluted at 15.21 min, 42 Da higher than that of M40 and M41, speculating that the loss of CH_2O and di-acetylation of amines took place. The characteristic fragment ion at *m/z* 159.0462 created by RDA reaction was 12 larger than 147.0461 of the parent drug, while the product ion at *m/z* 263.0559 increased by 42 Da compared with 221.0434 of eupatorin, inferring that loss of CH_2O occurred at the position of 4′ like M40 and M41, while di-acetylation reaction happened at the hydroxyl group at 5 and 3′ position.

Metabolite M43 ($C_{20}H_{16}O_9$), eluted at 14.14 min, which was detected with the deprotonated molecular ion $[M-H]^-$ at *m/z* 399.0704, 84 Da higher than that of M4 and M5, implying that di-acetylation reaction happened on the basis of the loss of CH_2 and CH_2. The diagnostic fragment ions at *m/z* 357.0641 and 315.0523 were attributed to the loss of C_2H_2O consecutively. Moreover, according to the product ions at *m/z* 175.0034 and 147.0316 acquired by RDA reaction, it may have three possible metabolites.

Metabolite M44 ($C_{20}H_{16}O_7$) exhibited a sharp peak at an elution time of 15.24 min in the XIC with a deprotonated ion at *m/z* 367.0806 and it was 32 Da (2O) lower than M43, suggesting that the loss of CH_2O and CH_2O and di-acetylation reaction occurred. The noteworthy fragment ion at *m/z* 283.0237 was yielded by dropping of $2C_2H_2O$. M44 generated the fragment ions at *m/z* 189.0633 and 159.0348 after RDA reaction, so the possible structures were inferred according to above MS/MS information.

Metabolites M45, M46, M47 and M48, eluted at 4.57 min, 5.14 min, 5.43 min, 9.69 min, respectively, all exhibited the deprotonated ion at *m/z* 329.1028, 329.1029, 329.1029, 329.1029, implying that they were isomers with the molecular formula $C_{18}H_{18}O_6$ and were 2 Da higher than that of M9 and M10, so hydrogenation happened on the basis of the loss of O. In M45, the representative product ion at *m/z* 130.9874 obtained by RDA reaction was 16 (O) less than 147.0461 of eupatorin, while the fragment ion at *m/z* 223.0900 was twice higher than 221.0434 of eupatorin, indicating that the loss of O and hydrogenation happened at the position of 3′ and 4, respectively. Like M45, according to the crucial product ions at *m/z* 147.0535 and 207.0777, 149.0538 and 207.0618, 133.0731 and 223.0742, the structures of M46, M47 and M48 were distinguished by the analysis above. Furthermore, M45, M46, M47 and M48 were also proved by the different Clog P values of 1.71027, 1.7953, 2.28982 and 2.89116, respectively.

Metabolites M49 and M50 ($C_{18}H_{18}O_5$) were observed in the extracted chromatogram at *m/z* 313.1080 and 313.1086 with the retention time of 7.61 min and 7.88 min, 2 Da higher than that of M11, while lacked one oxygen atom compared with M45, M46, M47 and M48. In M49, the distinctive product ion at *m/z* 130.9677 obtained by the RDA reaction was 16 (O) less than 147.0461 of eupatorin, inferring that hydrogenation happened at the position of 4′. Nevertheless, the characteristic fragment ion at *m/z* 133.0654 yielded in M49 by the RDA reaction was 14 less than 147.0461, so hydrogenation happened at the position of 2 and 3. Besides, M49 and M50 were also checked by the different Clog P values of 2.5321 and 3.02662, respectively.

Metabolite M51 ($C_{17}H_{16}O_7$) was obtained with a peak at *m/z* 331.0824 in the UPLC system, which was eluted at 10.05 min, 2 Da larger than that of M1, M2 and M3. According to MS/MS spectrum, diagnostic product ions at *m/z* 316.0595, 313.1396, 223.1657, 109.0288 and 135.0450 were formed by losing CH_3, H_2O, $C_6H_4O_2$, $C_{11}H_{10}O_5$ and RDA reaction. It's worth mentioning that the fragment ion at *m/z* 135.0450 was 12 less than 147.0461 of eupatorin, so demethylation happened at the methoxy group at 4′ position and hydrogenation occurred at the position of 2 and 3.

Metabolite M52 ($C_{18}H_{20}O_7$) was eluted at the retention time of 12.78 min. Its deprotonated molecular ion $[M-H]^-$ at *m/z* 347.1140 was increased 4 Da compared with eupatorin, so di-hydrogenation occurred. According to the dominant product ion at *m/z* 149.0591 obtained by RDA reaction, 2 Da higher than 147.0461 of eupatorin, and the fragment ion at *m/z* 225.0074, 4 Da higher than 221.0434, indicating that di-hydrogenation happened at the position of 2, 3 and 4.

Metabolites M53 and M54 ($C_{18}H_{20}O_6$) were observed in the chromatogram at *m/z* 331.1186 and 331.1185 with the retention time of 3.60 min and 4.07 min, respectively, 16 Da (O) lower than that of M52, inferring that the loss of O happened on the basis of di-hydrogenation. The crucial ion at *m*/*z* 133.0325 was generated after the RDA cleavage, which was less than *m*/*z* 147.0461 one oxygen, so the loss of O occurred at the hydroxyl group at 3′ position in M53. However, according to the characteristic product ions at *m*/*z* 149.0680 and 209.0813, the loss of O happened at the hydroxyl group at 5 position while the di-hydrogenation was at the same position as M52. The respective Clog P values were 1.55427 and 1.6393, so M53 and M54 were verified.

Metabolite M55 ($C_{18}H_{20}O_5$) was eluted at 6.36 min possessing the deprotonated molecular ion $[M-H]^-$ at *m*/*z* 315.1214, which was 16 Da (O) lower than that of M53, M54 and 4 Da higher than that of M11. Based on the previous analysis of M11, M53 and M54, the structure of M55 can be inferred.

Metabolites M56, M57 and M58 ($C_{17}H_{18}O_7$): Three chromatographic peaks were eluted at 10.18 min, 10.24 min and 10.80 min with deprotonated molecular ions $[M-H]^-$ at *m*/*z* 333.0972, 333.0982 and 333.0979, which were 4 Da larger than the size of M1-M3 and 14 Da (CH_2) higher than that of M52. According to the prominent product ions at *m*/*z* 149.0642 and 135.1164, together with the information of M1–M3 and M52, the structures of M56-M58 could be identified. In addition, M56, M57 and M58 were also ensured by the different Clog P values of 0.324751, 0.371274 and 0.644751, respectively.

Metabolite M59 ($C_{16}H_{16}O_7$) was observed with a peak at *m*/*z* 319.0815 in the chromatogram, which was eluted at 8.47 min, 4 Da larger than that of M3 and M4. According to the MS/MS information, the typical fragment ions at *m*/*z* 301.0701, 211.0353, 197.0452, 149.0269 and 135.0443 were created by loss of H_2O, $C_6H_4O_2$, $C_7H_6O_2$ and RDA reaction, so there were three possible metabolites of M59.

Metabolites M60 and M61 ($C_{16}H_{16}O_5$) appeared as deprotonated molecular ions $[M-H]^-$ at *m*/*z* 287.0923 and 287.0927, together with the retention time of 9.97 min and 11.07 min, respectively, which were 4 Da higher than M8, indicating that M60 and M61 might undergo the loss of CH_2O and CH_2O reaction followed by di-hydrogenation. In the secondary mass spectrum of M61, it obtained the fragment ions at *m*/*z* 272.0695, 241.2138, 165.0166, 123.0117 and 149.0683 yielded by dropping of CH_3, CH_2O_2, $C_7H_6O_2$, $C_9H_8O_3$ and RDA cleavage, so the loss of CH_2O and CH_2O reaction happened at the positions of 6 and 7 while di-hydrogenation happened at the positions of 2, 3 and 4. However, the characteristic fragment ions at *m*/*z* 195.0653, 93.0325 and 119.0500 produced by the loss of C_6H_4O, $C_{10}H_{10}O_4$ and RDA cleavage. So, there were two positions (4′, 7 or 4′, 6) to have lost CH_2O. Finally, the sizes of different Clog P values were combined to determine the structure of M60.

Metabolite M62 ($C_{17}H_{18}O_5$) was eluted at 7.37 min, which displayed deprotonated molecular ion $[M-H]^-$ at *m*/*z* 301.1078, 4 Da larger the size of M12. In M62, the characteristic fragment ion at *m*/*z* 149.0605 was twice higher than 147.0461 of eupatorin, while, the prominent fragment ion at *m*/*z* 179.0711 was 42 times lower than fragment ion at *m*/*z* 221.0434, so the loss of O, CH_2O and di-hydrogenation occurred at the same positions as M12 and M52, respectively.

Metabolite M63 ($C_{15}H_{10}O_4$), the deprotonated molecular ion of *m*/*z* 253.0512 was observed at the retention time of 7.81 min, which was 30 Da (CH_2O) lower than that of M8. M63 comprised the typical fragment ions at *m*/*z* 225.0558, 209.0606, 161.0249 and 117.0351 by dropping of CO, CO_2, C_6H_4O and RDA cleavage. And the fragment ion at *m*/*z* 101.0246 arose by loss of O from the ion at *m*/*z* 117.0351, so the structure was inferred.

Metabolites M64, M65 and M66 ($C_{18}H_{14}O_8$) were the isomeric metabolites with the deprotonated $[M-H]^-$ ions at *m*/*z* 357.0607, 357.0609 and 357.0610, 14 Da higher than that of eupatorin, which were eluted at 10.79 min, 10.81 min and 12.99 min, respectively, suggesting that ketone formation reaction occurred. Several conspicuous ions at *m*/*z* 342.0374, 327.0132, 313.0304, 221.0224, 235.0267, 161.0174 and 147.0012 all appeared in the secondary mass spectra after the loss of CH_3, CH_3, CO_2, $C_7H_4O_3$, $C_7H_6O_2$ and RDA reaction. Moreover, the Clog P values of M64, M65 and M66 were 2.17223, 2.27223 and 2.52223, respectively. In consequence, the structures of M64, M65 and M66 were distinguished according to the above information.

Metabolite M67 ($C_{18}H_{18}O_9$) was eluted at 12.29 min and showed the deprotonated molecular ion $[M-H]^-$ at *m*/*z* 377.0875, 18 Da higher than that of M14-M17, implying that M67 might undergo oxidation followed by internal hydrolysis. In the MS/MS spectrum of M67, the representative product ion at *m*/*z*

181.0136 tested after RDA reaction was 34 larger than 147.0461 of eupatorin, while the fragment ion at *m/z* 239.0435 was 18 times higher than that of eupatorin, so internal hydrolysis happened at C-2 and C-3 and oxidation is most likely to occur at C-5′ [29].

Metabolite M68 ($C_{19}H_{17}NO_8$) was observed with a peak at *m/z* 386.0865 in the UPLC system, which was eluted at 10.00 min, 57 Da higher than that of M1-M3. The fragment ion at *m/z* 329.0662 was acquired, corresponding to the loss of glycine. Additionally, the conspicuous fragment ions at *m/z* 264.1192 and 147.0973 were yielded through the loss of $C_7H_6O_2$ and RDA reaction, implying that the loss of CH_2 and glycine conjugation were connected to A ring. Hence, there were two possible metabolites of M68.

Metabolite M69 ($C_{19}H_{17}NO_6$) was detected at 6.09 min, which presented an accurate deprotonated ion [M-H]$^-$ at *m/z* 354.0992, 57 Da higher than that of M12, indicating that M69 might experience the loss of O and CH_2O reaction followed by glycine conjugation. A sequence of crucial fragment ions at *m/z* 324.2008, 250.9077, 174.9553 and 204.0387 were produced by the loss of $2CH_3$, $C_3H_5NO_3$, $C_9H_9NO_3$ and RDA reaction, while the characteristic fragment ion at *m/z* 204.0387 was 57 higher than 147.0461 of eupatorin, inferring that glycine conjugation was connected to B ring and the loss of O and CH_2O reaction was at the same position as M12.

Metabolite M70 ($C_{22}H_{22}N_2O_8$) exhibited a sharp peak at an elution time of 5.07 min in the XIC with a deprotonated ion at *m/z* 441.1302. The characteristic fragment ion at *m/z* 312.8496 was observed, corresponding to the loss of glutamine [24]. Furthermore, A strong ion at *m/z* 245.1021 appeared in the secondary mass spectrum of M70 after the RDA reaction, and was 98 higher than 147.0461 of eupatorin, which created the prominent fragment ion at *m/z* 117.2304 by dropping of glutamine, suggesting that the loss of CH_2O occurred at the methoxy group at 4′ position while glutamine conjugation took place at the hydroxyl group at 3′ position.

Metabolite M71 ($C_{22}H_{22}O_{12}$), displayed a peak at 5.49 min, as well as a deprotonated molecular ion [M-H]$^-$ at *m/z* 477.1036. The predominated fragment ion at *m/z* 315.6277 was attributed to the loss of glucose. In addition, the distinctive fragment ion at *m/z* 146.9654 resulted from RDA reaction was consistent with *m/z* 147.0461 of eupatorin, further noteworthy MS/MS fragment ions at *m/z* 355.0661 and 192.9548 were yielded corresponding to the consecutive loss of $C_7H_6O_2$ and glucose. Thus, the loss of CH_2 and CH_2 took place at the methoxy group at 6 and 7 position, while glucose conjugation happened at the hydroxyl group at 5 position.

The detected metabolites are listed in Table 1. Moreover, their XICs are exhibited in Figure 3.

Table 1. Summary of metabolites of eupatorin in vivo and in vitro.

Metabolites ID	Composition	Formula	*m/z*	Error (ppm)	R.T. (min)	MS/MS Fragments	Clog P	Score (%)	Plasma	Bile	Urine	Feces	RLMs	RIF
M1	Loss of CH_2 [M-H]$^-$	$C_{17}H_{14}O_7$	329.0660	−2.0	9.93	314.0427, 313.0384, 299.0188, 285.0371, 207.7129	2.26422	83.3	-	-	+^{a,b}	+^{a,b}	+^{I,II}	+^{a,b}
M2	Loss of CH_2 [M-H]$^-$	$C_{17}H_{14}O_7$	329.0668	0.5	10.27	314.0423, 313.0344, 299.0189, 285.0393, 133.0287	2.26434	91.8	-	-	+^{a,b}	+^{a,b}	+^{I,II}	+^{a,b}
M3	Loss of CH_2 [M-H]$^-$	$C_{17}H_{14}O_7$	329.0662	−1.4	10.79	314.0436,313.0357, 299.0204, 285.0396, 207.7166	2.51422	96.3	-	-	+^{a,b}	+^{a,b}	+^{I,II}	+^{a,b}
M4a M4b	Loss of CH_2 and CH_2 [M-H]$^-$	$C_{16}H_{12}O_7$	315.0500	−3.3	7.26	300.0279, 297.1740, 269.1760, 251.1269, 133.0270	1.82034 2.07034	85.4	-	-	+^{a,b}	+^{a,b}	+^{I}	-
M5	Loss of CH_2 and CH_2 [M-H]$^-$	$C_{16}H_{12}O_7$	315.0504	−2.0	8.50	300.0275, 297.0411, 269.1755, 251.1658, 147.0821	2.18513	93.3	-	-	+^{a,b}	+^{a,b}	+^{I}	-
M6a M6b M6c	Loss of CH_2O [M-H]$^-$	$C_{17}H_{14}O_6$	313.0713	−1.4	13.86	298.0483, 283.0250, 221.0632, 147.0078, 117.0364	2.86048 3.08571 3.33571	90.9	-	-	-	+^{a,b}	-	+^{a,b}
M7	Loss of CH_2O and CH_2 [M-H]$^-$	$C_{16}H_{12}O_6$	299.0562	0.3	10.10	284.0326, 281.1787, 251.1281, 146.9687	2.74964	83.8	-	-	+^{a}	-	+^{I}	+^{b}
M8	Loss of CH_2O and CH_2O [M-H]$^-$	$C_{16}H_{12}O_5$	283.0614	0.8	13.60	268.0379, 240.0428, 267.0306, 161.0025, 146.9655	3.30833	93.5	-	-	+^{a,b}	+^{b}	+^{I}	+^{b}
M9	Loss of O [M-H]$^-$	$C_{18}H_{16}O_6$	327.0882	2.4	4.98	308.9931, 299.1274, 281.2489, 205.0025, 146.9380	2.45814	75.3	-	-	-	-	-	+^{b}
M10	Loss of O [M-H]$^-$	$C_{18}H_{16}O_6$	327.0872	−0.8	7.47	309.0800, 299.0957, 281.2493, 221.0452, 130.9716	3.44497	83.5	-	-	-	-	-	+^{b}
M11	Loss of O and O [M-H]$^-$	$C_{18}H_{16}O_5$	311.0930	1.5	9.55	250.9816, 204.9868, 174.9556, 130.9658	3.19475	75.7	-	-	-	-	-	+^{b}
M12a M12b	Loss of O and CH_2O [M-H]$^-$	$C_{17}H_{14}O_5$	297.0768	−0.2	7.33	267.1016, 253.0865, 175.0394, 147.0452, 145.0305	2.78433	82.1	-	-	-	-	-	+^{b}
M13	Loss of CO [M-H]$^-$	$C_{17}H_{16}O_6$	315.0862	−3.8	12.74	300.0633, 285.0401, 270.0144, 193.0503, 147.0445	2.84747	87.9	-	-	-	+^{a,b}	-	+^{a,b}
M14	Oxidation [M-H]$^-$	$C_{18}H_{16}O_8$	359.0772	−0.2	10.01	344.0542, 329.0304, 314.0064, 220.9817, 163.0384	1.79518	82.9	-	-	+^{a,b}	+^{a,b}	+^{I}	+^{a,b}
M15	Oxidation [M-H]$^-$	$C_{18}H_{16}O_8$	359.0768	−1.3	10.50	344.0529, 329.0306, 314.0061, 221.0098, 163.0368	1.84518	82.8	-	-	+^{a,b}	+^{a,b}	+^{I}	+^{a,b}
M16	Oxidation [M-H]$^-$	$C_{18}H_{16}O_8$	359.0767	−1.6	11.47	344.0536, 329.0296, 314.0066, 221.0762, 163.0019	1.86518	81.9	-	-	+^{a,b}	+^{a,b}	+^{I}	+^{a,b}
M17	Oxidation [M-H]$^-$	$C_{18}H_{16}O_8$	359.0767	−1.4	12.23	344.0542, 329.0315, 314.0085, 237.0375, 147.0130	1.87123	85.5	-	-	+^{a,b}	+^{a,b}	+^{I}	+^{a,b}
M18	Di-Oxidation [M-H]$^-$	$C_{18}H_{16}O_9$	375.0709	−3.3	9.90	329.0669, 314.0434, 299.0191, 221.1216, 178.9947	0.9644	91.2	-	-	-	+^{a,b}	-	+^{a}
M19	Tri-Oxidation [M-H]$^-$	$C_{18}H_{16}O_{10}$	391.0673	0.5	12.26	345.0869, 330.0636, 315.0393, 221.0399, 195.0289	0.25226	77.1	-	-	+^{b}	-	-	+^{a,b}
M20a M20b	Demethylation and Oxidation [M-H]$^-$	$C_{17}H_{14}O_8$	345.0605	−3.0	9.43	330.0379, 301.1825, 221.1270, 149.0245, 125.0237	1.29734	84.9	-	-	+^{a,b}	+^{b}	+^{I}	+^{b}
M21	Demethylation and Oxidation [M-H]$^-$	$C_{17}H_{14}O_8$	345.0606	−2.8	10.29	330.0384, 301.0719, 221.0028, 149.0234, 125.0311	1.59734	87.1	-	-	+^{a,b}	+^{b}	+^{I}	+^{b}

Table 1. *Cont.*

Metabolites ID	Composition	Formula	*m/z*	Error (ppm)	R.T. (min)	MS/MS Fragments	Clog P	Score (%)	Plasma	Bile	Urine	Feces	RLMs	RIF
M22	Methylation $[M-H]^-$	$C_{19}H_{18}O_7$	357.0972	−2.1	10.02	342.0740, 327.0503, 312.0266, 235.0434, 147.0433	2.06632	80.6	-	-	+ [a,b]	+ [a,b]	-	+ [a,b]
M23	Methylation $[M-H]^-$	$C_{19}H_{18}O_7$	357.0969	−3.1	12.86	342.0737, 327.0508, 312.0266, 221.0766, 161.0269	3.18323	78.6	-	-	+ [a,b]	+ [a,b]	-	+ [a,b]
M24	Loss of O+Methylation $[M-H]^-$	$C_{19}H_{18}O_6$	341.1025	−1.5	7.15	326.1073, 311.0918, 235.0607, 130.9906, 107.0440	2.80306	82.8	-	-	-	-	-	+ [b]
M25	Loss of O+Methylation $[M-H]^-$	$C_{19}H_{18}O_6$	341.1027	−1.2	8.79	326.0798, 311.0451, 204.9196, 161.0595, 137.0553	2.9313	82.1	-	-	-	-	-	+ [b]
M26	Loss of CH_2O and CH_2O+Demethylation $[M-H]^-$	$C_{15}H_{10}O_5$	269.0459	1.3	9.89	253.0124, 241.0500, 225.0555, 133.0298, 117.0349	2.88784	95.6	+ [a,b]	-	+ [a,b]	-	-	+ [a,b]
M27	Loss of CH_2O and CH_2+Demethylation $[M-H]^-$	$C_{15}H_{10}O_6$	285.0402	−0.9	8.45	267.0130, 241.0462, 221.0063, 177.0189, 133.0307	2.31115	90.2	-	-	-	-	-	+ [b]
M28	Glucuronidation $[M-H]^-$	$C_{24}H_{24}O_{13}$	519.1140	−0.8	7.10	397.0442, 343.0822, 328.0587, 313.0346, 146.9662	−0.495	81.6	+ [a]	+ [b]	+ [a,b]	-	+ [II]	-
M29	Glucuronidation $[M-H]^-$	$C_{24}H_{24}O_{13}$	519.1151	1.3	8.14	343.0824, 328.0588, 323.0173, 313.0354, 221.0262	0.62193	83.1	+ [a]	+ [b]	+ [a,b]	-	+ [II]	-
M30	Demethylation and Glucuronide Conjugation $[M-H]^-$	$C_{23}H_{22}O_{13}$	505.0979	−1.8	7.88	329.0669, 309.0687, 299.0165, 285.0735	0.14693	81.3	-	-	+ [a,b]	-	-	-
M31	Sulfate Conjugation $[M-H]^-$	$C_{18}H_{16}O_{10}S$	423.0391	0.1	8.93	343.0830, 328.0593, 313.0355, 285.0413, 147.0037	0.27032	86.1	-	+ [a,b]	+ [a,b]	-	-	+ [a,b]
M32	Sulfate Conjugation $[M-H]^-$	$C_{18}H_{16}O_{10}S$	423.0387	−0.9	9.20	343.0836, 328.0606, 313.0371, 285.0457, 227.0084	1.38723	89.6	-	+ [a,b]	+ [a,b]	-	-	+ [a,b]
M33	Loss of CH_2+Sulfate Conjugation $[M-H]^-$	$C_{17}H_{14}O_{10}S$	409.0233	−0.4	9.01	329.0670, 314.0432, 299.0198, 212.0456, 132.0208	0.81223	91.5	+ [a]	-	+ [a,b]	-	-	-
M34	Loss of O+Sulfate Conjugation $[M-H]^-$	$C_{18}H_{16}O_9S$	407.0434	−2.1	12.65	327.0826, 301.0034, 220.9818, 131.0573	1.00706	63.3	-	-	-	-	-	+ [b]
M35	N-Acetylation $[M-H]^-$	$C_{20}H_{18}O_8$	385.0917	−3.1	13.36	370.0729, 355.0427, 343.0846, 263.0551, 147.0513	1.49632	74.4	-	-	-	-	-	+ [b]
M36	N-Acetylation $[M-H]^-$	$C_{20}H_{18}O_8$	385.0925	−0.9	13.90	370.0735, 355.0492, 343.0874, 221.0781, 189.0551	2.61323	77.9	-	-	-	-	-	+ [b]
M37	Loss of O+N-Acetylation $[M-H]^-$	$C_{20}H_{18}O_7$	369.0987	2.1	13.02	327.2227, 279.0680, 263.1681, 237.1098, 130.9934	2.23306	75.3	+ [a]	-	+ [a,b]	+ [a,b]	+ [I]	+ [a,b]
M38	Loss of O+N-Acetylation $[M-H]^-$	$C_{20}H_{18}O_7$	369.0975	−1.3	13.90	354.0755, 339.0542, 311.0594, 174.9586, 164.9289	2.3613	76.7	+ [a]	-	+ [a,b]	+ [a,b]	+ [I]	+ [a,b]
M39	Loss of CH_2+N-Acetylation $[M-H]^-$	$C_{19}H_{16}O_8$	371.0761	−3.2	11.45	329.0680, 314.0439, 299.0196, 220.9869, 175.0389	2.13823	76.8	-	-	-	+ [a,b]	-	+ [b]
M40	Loss of CH_2O+N-Acetylation $[M-H]^-$	$C_{19}H_{16}O_7$	355.0813	−2.9	11.86	340.0587, 325.0335, 313.1107, 263.0361, 117.0329	1.64857	78.2	-	-	+ [a,b]	+ [a,b]	+ [I]	+ [a,b]

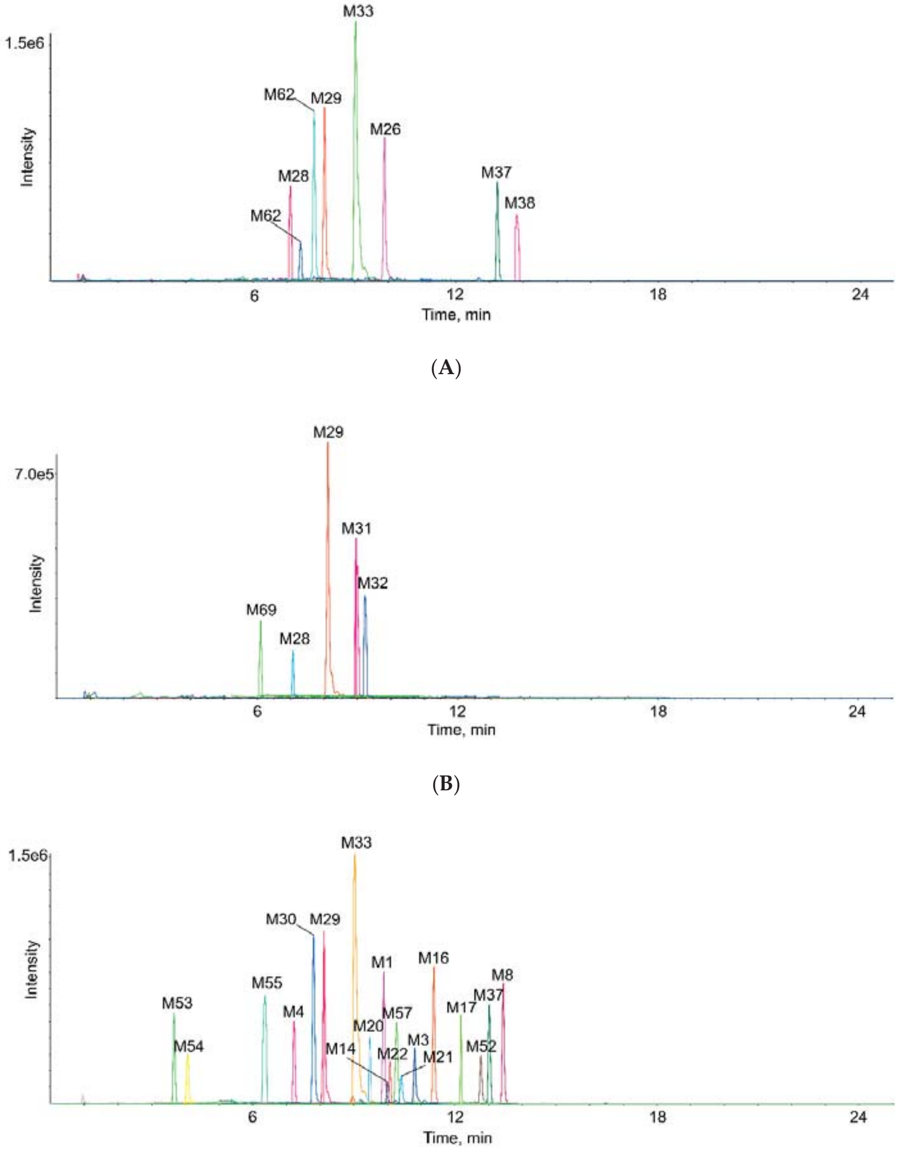
Table 1. *Cont.*

Metabolites ID	Composition	Formula	*m/z*	Error (ppm)	R.T. (min)	MS/MS Fragments	Clog P	Score (%)	Plasma	Bile	Urine	Feces	RLMs	RIF
M41	Loss of CH_2O+N-Acetylation [M-H]$^-$	$C_{19}H_{16}O_7$	355.0814	−2.6	13.15	340.0589, 325.0356, 313.0337, 221.0027, 159.1093	2.87497	78.1	-	-	+ [a,b]	+ [a,b]	+ [I]	+ [a,b]
M42	Loss of CH_2O+Di-Acetylation of Amines [M-H]$^-$	$C_{21}H_{18}O_8$	397.0918	−2.6	15.21	382.0691, 367.0426, 313.2553, 263.0559, 159.0462	1.66306	74.8	-	-	-	+ [b]	-	+ [b]
M43a M43b M43c	Loss of CH_2 and CH_2+Di-Acetylation of Amines [M-H]$^-$	$C_{20}H_{16}O_9$	399.0704	−4.3	14.14	384.0497, 357.0641, 315.0523, 175.0034, 147.0316	1.56823	71.8	-	-	-	+ [b]	-	-
M44a M44b M44c	Loss of CH_2O and CH_2O+Di-Acetylation of Amines [M-H]$^-$	$C_{20}H_{16}O_7$	367.0806	−4.8	15.24	283.0237, 233.1253, 202.9904, 189.0633, 159.0348	2.05475 2.11133 2.40475	71.9	-	-	-	+ [a,b]	-	-
M45	Loss of O+Hydrogenation [M-H]$^-$	$C_{18}H_{18}O_6$	329.1028	−0.8	4.57	314.0861, 299.1136, 283.2624, 223.0900, 130.9874	1.71027	75.9	-	-	-	-	-	+ [a,b]
M46	Loss of O+Hydrogenation [M-H]$^-$	$C_{18}H_{18}O_6$	329.1029	−0.5	5.14	314.0905, 299.1189, 283.1288, 207.0777, 147.0535	1.7953	76.7	-	-	-	-	-	+ [a,b]
M47	Loss of O+Hydrogenation [M-H]$^-$	$C_{18}H_{18}O_6$	329.1029	−0.3	5.43	314.0384, 299.0986, 283.2612, 207.0618, 149.0538	2.28982	78.7	-	-	-	-	-	+ [a,b]
M48	Loss of O+Hydrogenation [M-H]$^-$	$C_{18}H_{18}O_6$	329.1029	−0.6	9.69	314.0226, 299.0298, 283.0982, 223.0742, 133.0731	2.89116	75.8	-	-	-	-	-	+ [a,b]
M49	Loss of O and O+Hydrogenation [M-H]$^-$	$C_{18}H_{18}O_5$	313.1080	−0.3	7.61	298.0759, 269.1191, 239.1066, 206.9936, 130.9677	2.5321	88.5	-	-	-	-	-	+ [a,b]
M50	Loss of O and O+Hydrogenation [M-H]$^-$	$C_{18}H_{18}O_5$	313.1086	1.6	7.88	298.0767, 269.1189, 239.1061, 207.0818, 133.0654	3.02662	90.7	-	-	-	-	-	+ [a,b]
M51	Demethylation and Hydrogenation [M-H]$^-$	$C_{17}H_{16}O_7$	331.0824	0.3	10.05	316.0595, 313.1396, 301.0343, 223.1657, 135.0450	1.70816	90.5	-	-	+ [a,b]	-	-	+ [a]
M52	Di-Hydrogenation [M-H]$^-$	$C_{18}H_{20}O_7$	347.1140	1.0	12.78	332.0913, 317.0676, 225.0074, 149.0591, 123.0079	0.81747	60.2	-	-	+ [a,b]	-	-	+ [b]
M53	Loss of O+Di-Hydrogenation [M-H]$^-$	$C_{18}H_{20}O_6$	331.1186	−0.2	3.60	301.1097, 299.1257, 285.1728, 225.0501, 133.0325	1.55427	53.1	-	-	+ [b]	-	-	+ [b]
M54	Loss of O+Di-Hydrogenation [M-H]$^-$	$C_{18}H_{20}O_6$	331.1185	−0.7	4.07	301.1088, 299.1304, 285.0945, 209.0813, 149.0680	1.6393	52.8	-	-	+ [b]	-	-	+ [b]
M55	Loss of O and O+Di-Hydrogenation [M-H]$^-$	$C_{18}H_{20}O_5$	315.1214	−1.6	6.36	285.1155, 271.1557, 269.1318, 241.1035, 133.0693	2.3761	83.6	-	-	+ [a]	-	-	+ [a,b]

Table 1. *Cont.*

Metabolites ID	Composition	Formula	*m/z*	Error (ppm)	R.T. (min)	MS/MS Fragments	Clog P	Score (%)	Plasma	Bile	Urine	Feces	RLMs	RIF
M56	Loss of CH_2+Di-Hydrogenation [M-H]$^-$	$C_{17}H_{18}O_7$	333.0972	−2.3	10.18	318.0813, 317.1125, 303.0579, 211.0624, 149.0642	0.32475	94.4	-	-	+ [a,b]	-	-	-
M57	Loss of CH_2+Di-Hydrogenation [M-H]$^-$	$C_{17}H_{18}O_7$	333.0982	0.6	10.24	315.0034, 225.2199, 179.1064, 135.1164, 109.0679	0.37127	63.2	-	-	+ [a,b]	-	-	-
M58	Loss of CH_2+Di-Hydrogenation [M-H]$^-$	$C_{17}H_{18}O_7$	333.0979	−0.3	10.80	315.0884, 300.0638, 285.0374, 211.0614, 149.0693	0.64475	90.6	-	-	+ [a,b]	-	-	-
M59a M59b M59c	Loss of CH_2 and CH_2+Di-Hydrogenation [M-H]$^-$	$C_{16}H_{16}O_7$	319.0815	−2.6	8.47	301.0701, 211.0353, 197.0452, 149.0269, 135.0443	0.19855 −0.1214 0.22768	89.0	-	-	-	+ [a,b]	-	-
M60	Loss of CH_2O and CH_2O+Di-Hydrogenation [M-H]$^-$	$C_{16}H_{16}O_5$	287.0923	−0.8	9.97	272.0689, 241.0875, 195.0653, 119.0500, 93.0325	1.35527	80.7	-	-	-	-	-	+ [b]
M61	Loss of CH_2O and CH_2O+Di-Hydrogenation [M-H]$^-$	$C_{16}H_{16}O_5$	287.0927	0.5	11.07	272.0695, 241.2138, 165.0166, 149.0683, 123.0117	1.4214	85.8	-	-	-	-	-	+ [b]
M62a M62b	Loss of O and CH_2O+Di-Hydrogenation [M-H]$^-$	$C_{17}H_{18}O_5$	301.1078	−1.2	7.37	271.1359, 255.0541, 241.0814, 179.0711, 149.0605	1.9972	87.9	+ [a]	-	-	+ [a,b]	-	-
M63	Loss of CH_2O and CH_2O+Loss of Hydroxymethylene [M-H]$^-$	$C_{15}H_{10}O_4$	253.0512	2.1	7.81	225.0558, 209.0606, 161.0249, 117.0351, 101.0246	3.4753	70.9	+ [b]	-	+ [a,b]	-	-	+ [b]
M64	Ketone Formation [M-H]$^-$	$C_{18}H_{14}O_8$	357.0607	−2.5	10.79	342.0374, 327.0132, 313.0304, 221.0224, 161.0174	2.17223	79.1	-	-	-	+ [b]	+ [I]	-
M65	Ketone Formation [M-H]$^-$	$C_{18}H_{14}O_8$	357.0609	−1.8	10.81	342.0379, 327.0146, 313.0349, 235.0241, 147.0012	2.27223	76.1	-	-	-	+ [b]	+ [I]	-
M66	Ketone Formation [M-H]$^-$	$C_{18}H_{14}O_8$	357.0610	−1.6	12.99	342.0385, 327.0198, 313.0421, 235.0267, 147.0503	2.52223	77.4	-	-	-	+ [b]	+ [I]	-
M67	Oxidation and Internal Hydrolysis [M-H]$^-$	$C_{18}H_{18}O_9$	377.0875	−0.8	12.29	362.0508, 349.0873, 239.0435, 181.0136, 139.0054	0.21452	83.4	-	-	-	+ [a,b]	-	+ [a,b]
M68a M68b	Loss of CH_2+Glycine Conjugation [M-H]$^-$	$C_{19}H_{17}NO_8$	386.0865	−4.2	10.00	329.0662, 314.0421, 299.0189, 264.1192, 147.0973	2.15391 2.40391	80.8	-	-	-	+ [b]	-	-
M69a M69b	Loss of O and CH_2O+Glycine Conjugation [M-H]$^-$	$C_{19}H_{17}NO_6$	354.0992	2.5	6.09	324.2008, 250.9077, 221.0751, 204.0387, 174.9553	2.66894	73.9	-	+ [a]	-	-	-	-
M70	Loss of CH_2O+Glutamine Conjugation [M-H]$^-$	$C_{22}H_{22}N_2O_8$	441.1302	−0.2	5.07	312.8496, 245.1021, 221.5883, 117.2304	0.99254	50.9	-	-	-	-	-	+ [b]
M71	Loss of CH_2 and CH_2+Glucose Conjugation [M-H]$^-$	$C_{22}H_{22}O_{12}$	477.1036	−0.6	5.49	355.0661, 315.6277, 192.9548, 146.9654, 123.0795	−0.3982	52.9	-	-	-	-	-	+ [b]

+, Detected; -, Undetected. RLMs, rat liver microsomes; RIF, rat intestinal flora. (a) Metabolites obtained by methanol precipitation protein; (b) metabolites obtained by ethyl acetate extraction; a+b metabolites obtained by methanol precipitation protein and ethyl acetate extraction. (I) Phase I metabolites obtained from liver microsomes; (II) phase II metabolites obtained from liver microsomes.

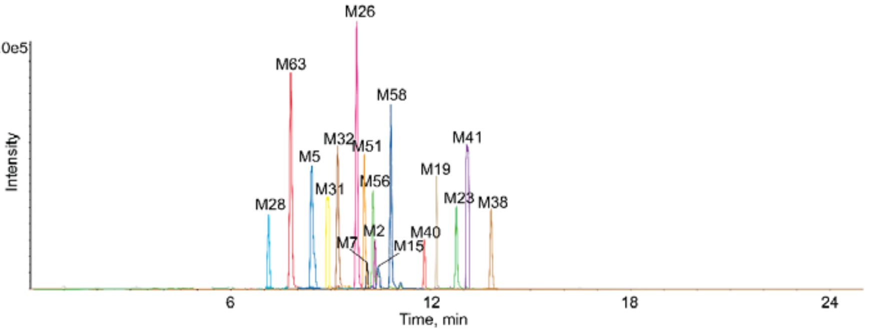


(**A**)

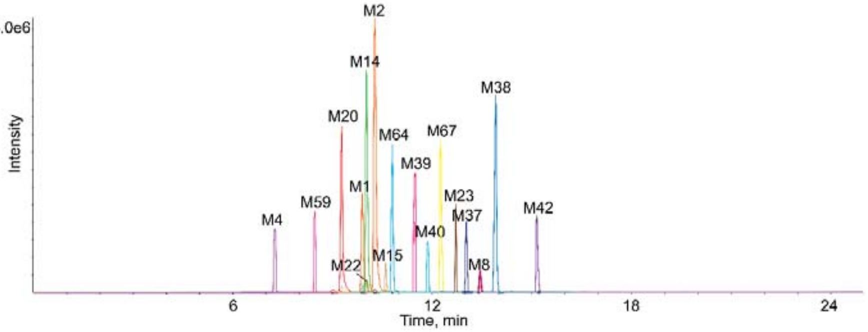
(**B**)

(**C1**)

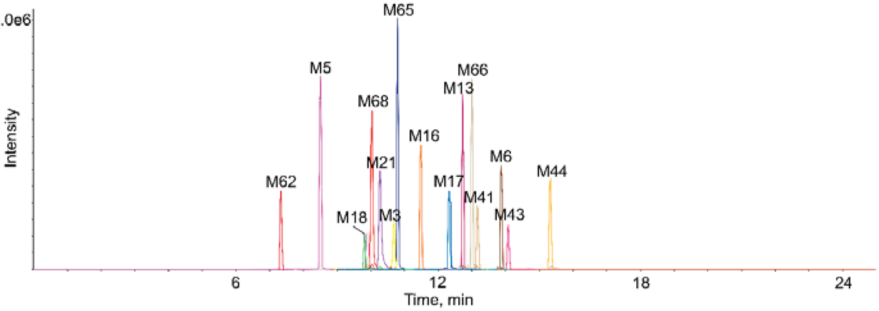
Figure 3. *Cont.*



(**C2**)

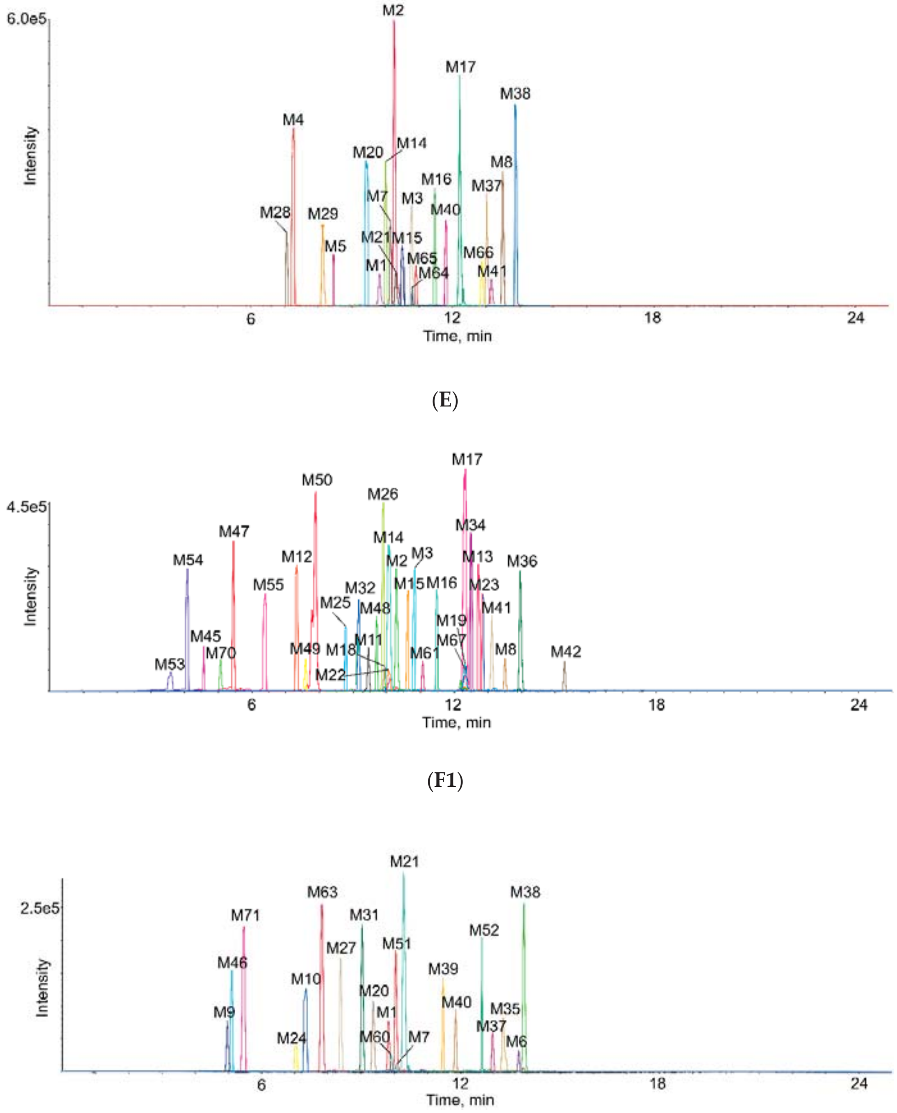


(**D1**)



(**D2**)

Figure 3. *Cont.*



(**E**)

(**F1**)

(**F2**)

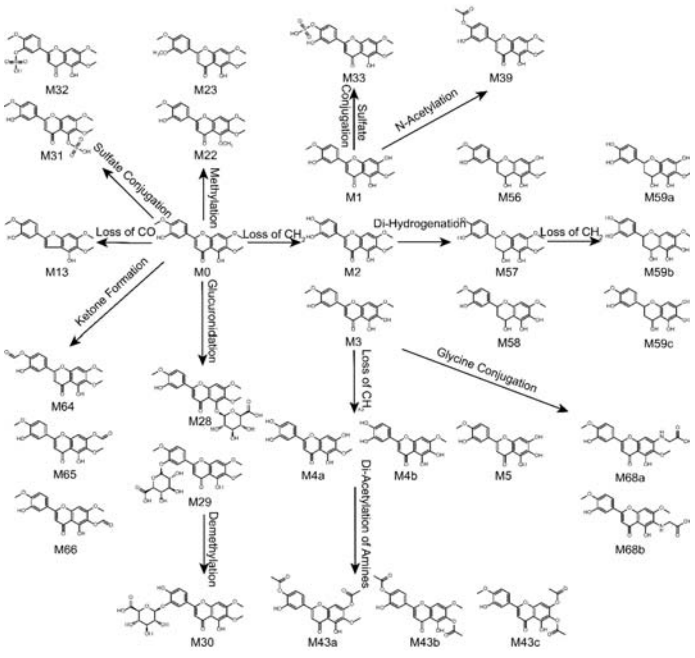
Figure 3. Extracted ion chromatograms of all metabolites of eupatorin in vivo and in vitro (**A**—in rat plasma sample, **B**—in rat bile sample, **C1**,**C2** in rat urine sample, **D1**,**D2** in rat feces sample, **E**—in rat liver microsomes, **F1**,**F2** in rat intestinal flora).

2.4. Metabolic Pathways of Eupatorin

The metabolites of eupatorin in rats after oral administration, in liver microsomes and intestinal flora through incubation was identified in this study. As a result, a total of 51 metabolites in vivo were detected, including 8 metabolites in plasma, 5 metabolites in bile, 36 metabolites in urine and 32 metabolites in feces. Meanwhile, 60 metabolites in vitro were observed, including 22 metabolites in liver microsomes and 53 metabolites in intestinal flora. The proposed metabolic pathways of eupatorin in vivo, in rat liver microsomes and in rat intestinal flora were shown in Figure 4. It is worth mentioning that the loss of CH_2, CH_2O, O, oxidation, glucuronidation and ketone formation was the primary metabolic step that produced further reactions such as sulfate conjugation, hydrogenation, N-acetylation, methylation, demethylation, internal hydrolysis, glycine conjugation, glutamine conjugation and glucose conjugation. Moreover, all metabolic changes above had taken place in vivo and in vitro. However, glycine conjugation was just present in vivo, while glutamine conjugation and glucose conjugation merely existed in vitro.

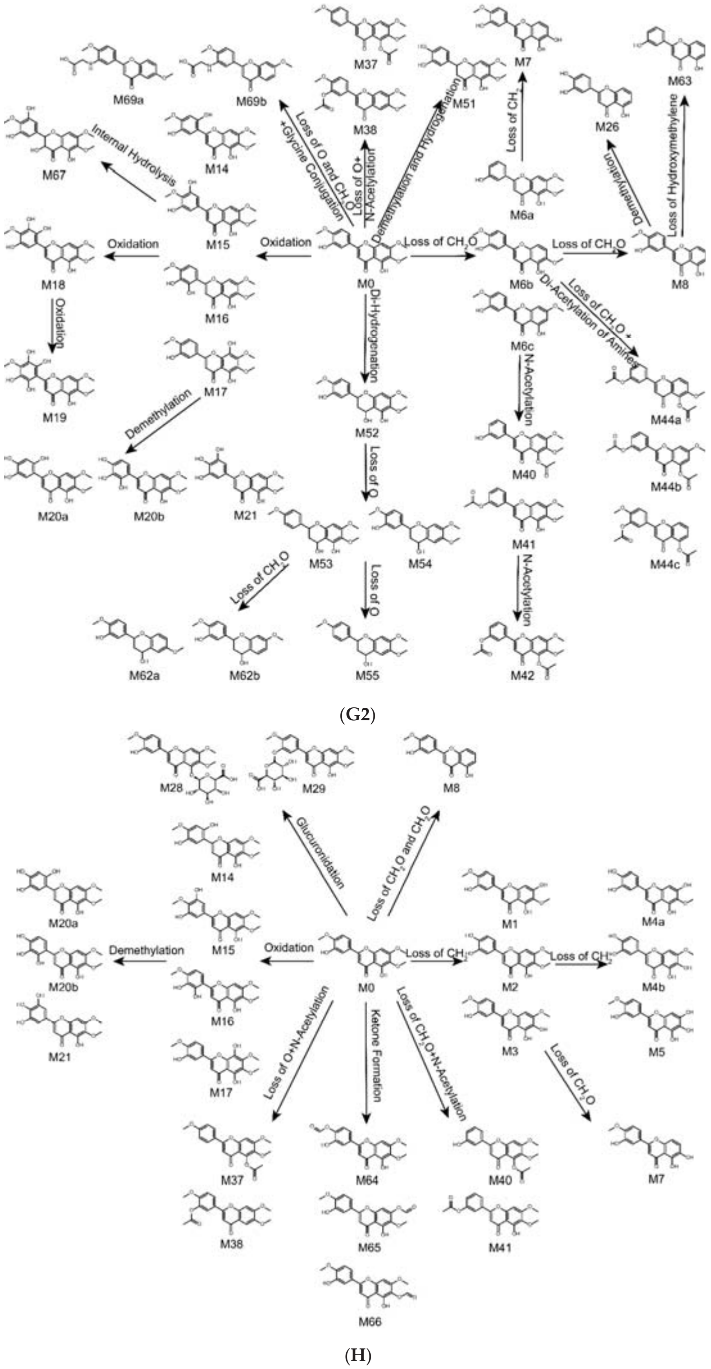
2.5. Comparison of Metabolites in Vivo and in Vitro

Drug metabolism plays a significant impact on various fields of pharmaceutical mechanisms as well as drug development and clinical use. In this work, the metabolism of eupatorin in vivo (plasma, bile, urine and feces) and in vitro (rat liver microsomes and intestinal flora) was investigated. In vivo; rat urine and feces possessed high activity for eupatorin metabolism, which were identified as having 36 and 32 metabolites, respectively. Nevertheless, only 8 metabolites were observed in rat plasma and 5 metabolites were detected in rat bile, suggesting that the rat plasma and bile might hold low biotransformation activity [30]. In vitro, 53 metabolites were obtained in rat intestinal flora while 22 metabolites were identified in rat liver microsomes, which implied that most metabolites could be excreted in intestinal flora samples and intestinal tract was more suitable for rapid identification of metabolites of eupatorin in vitro, with enormous catalytic and metabolic capacity which exceeds that of the liver microsomes [24]. Thus, the intestinal tract is considered as an extremely vital organ in the biotransformation of eupatorin.



(**G1**)

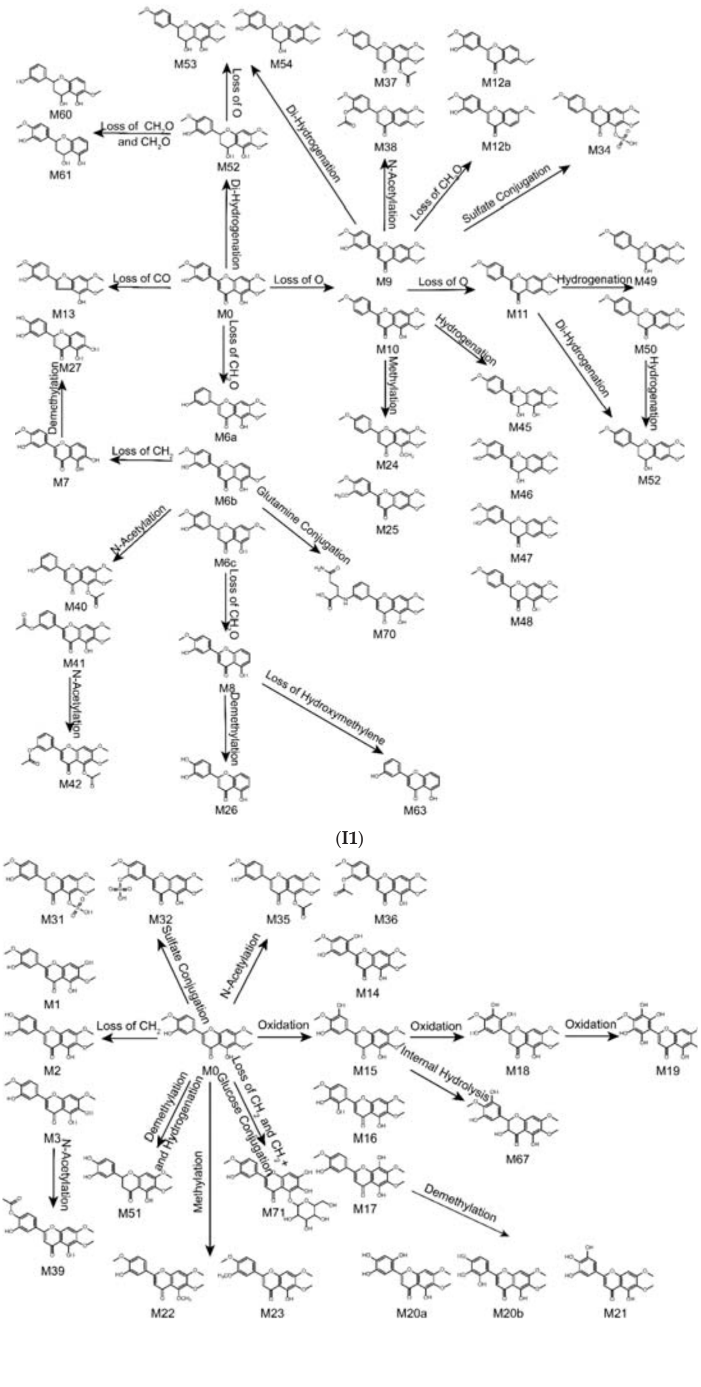
Figure 4. *Cont.*



(**G2**)

(**H**)

Figure 4. *Cont.*



(**I1**)

(**I2**)

Figure 4. Metabolic profile and proposed metabolic pathways of eupatorin in vivo and in vitro (**G1**,**G2** in vivo, **H** in rat liver microsomes, **I1**,**I2** in rat intestinal flora).

2.6. Metabolite Activity of Eupatorin

It has been reported in the literature that OS was taken as a beverage to improve health and for treatment of kidney disease, bladder inflammation and urethritis [1,2]. As its major active ingredient, eupatorin has also been reported to have meaningful anti-inflammatory activity [15,16]. In this study, the metabolites of eupatorin in urine samples were the largest, which may be related to the therapeutic effects of cystitis, nephritis and urethritis. In addition, many of the metabolites of eupatorin have been studied. For example, M4a namely nepetin, is a natural flavonoid present in different plants. In recent years, accumulating evidence has shown that nepetin exhibits various pharmacological activities, especially potent anti-inflammatory properties, which might be related to the strong anti-inflammatory activity of eupatorin [30–32]. Overall, the identification of metabolites of eupatorin provides a basis for new pharmacological studies and these metabolites will be further explored in the future.

3. Material and Methods

3.1. Chemicals and Materials

Eupatorin (855-96-9, purity > 98.94%) was purchased from Chengdu Desite Co., Ltd. (Chengdu, China). Beta-nicotinamide adenine dinucleotide phosphate (β-NADPH) was purchased from Sigma Chemical (St. Louis, MO, USA). Alamethicin and uridine 5′-diphosphoglucuronic acid trisodium salt (UDPGA) were purchased from BD Biosciences (Woburn, MA, USA). Phosphate buffer saline (PBS) was purchased from Sangon Biotech Co., Ltd. (Shanghai, China). Acetonitrile and methanol were all HPLC grade and were purchased from J.T.-Baker Company (Phillipsburg, NJ, USA). Formic acid (HPLC grade) was provided by Diamond Technology (Dikma Technologies Inc., Lake Forest, CA, USA). Purified water was purchased from Wahaha (Hangzhou Wahaha Group Co., Ltd., Hangzhou, China). L-ascorbic acid, L-cysteine, eurythrol, tryptone and nutrient agar were purchased from Beijing AoBoXing Bio-tech Co., Ltd. (Beijing, China). Sodium carboxymethyl cellulose (CMC-Na), sodium carbonate (Na_2CO_3), magnesium chloride ($MgCl_2$), potassium dihydrogen phosphate (KH_2PO_4), dipotassium phosphate (K_2HPO_4), calcium chloride ($CaCl_2$), ammonium sulfate ($(NH_4)_2SO_4$), sodium chloride (NaCl) and magnesium sulfate ($MgSO_4$) were obtained from Tianjin Guangfu Technology Development Co., Ltd. (Tianjin, China).

3.2. Instruments and Conditions

UHPLC-Q-TOF-MS/MS analysis was performed on a Nexera-X2 UHPLC system (Shimadzu Corp., Kyoto, Japan), which was combined with a triple TOF^{TM} 5600$^+$ MS/MS system (AB SCIEX, Concord, Ontario, Canada). The chromatographic separation was achieved on Poroshell 120 EC-C_{18} column (2.1 × 100 mm, 2.7 μm) with a SecurityGuard® UHPLC C18 pre-column (Poroshell).

The mobile phase was composed of 0.1% aqueous formic acid (eluent A) and acetonitrile (eluent B). The gradient elution program was as follows: 10–55% B from 0 to 15 min, 55–95% B from 15 to 20 min, 95–95% B from 20 to 25 min. The column temperature remained at 40 °C. In addition, the injection flow rate and the volume were set at 0.3 mL/min and 3 μL, respectively. Before the next injection, equilibration was performed for 3 min.

Mass spectrometric detection was carried out by a Triple TOF^{TM} 5600 system equipped with Duo-Spray^{TM} ion sources in the negative electrospray ionization (ESI) mode. The following mass spectrometry parameter settings were applied: ion spray voltage (IS), −4.5 kV; the turbo spray temperature, 550 °C; the optimized delustering potential (DP), −60 V; collision energy (CE), −10 eV; and the collision energy spread (CES), 15 eV. Moreover, the nebulizer gas (gas 1), the heater gas (gas 2) and the curtain gas were set to 55, 55 and 35 L/min, respectively.

3.3. Metabolism in Vivo

3.3.1. Animals and Drug Administration

Eighteen male Sprague-Dawley (SD) rats (220–220 g, 12–14 weeks old) were purchased from the Experimental Animal Research Center of Hebei Medical University (Certificate No.1811164). The conditions of temperature (22–25 °C), humidity (55–60%) and light (12 h light/dark cycle) were standard for 7 days before being used. All rats were fasted for 12 h but allowed water before the experiments. These rats were divided into six groups randomly with three rats per group. Groups 1, 3 and 5 were the control groups for blank blood, blank bile, blank urine and feces, respectively. Groups 2, 4 and 6 were the drug groups for blood, bile, urine and feces, respectively. Rats in groups 2, 4, 6 were given eupatorin by gavage, which dissolved in a 0.5% CMC-Na solution at a dose of 50 mg/kg. Nevertheless, an equal 0.5% CMC-Na solution without eupatorin was orally given to groups 1, 3, 5. All rat experiments were conducted in accordance with the committee's guidelines on the Care and Use of Laboratory Animals.

3.3.2. Bio-Sample Collection

The plasma samples collection: About 300 μL–500 μL for each blood sample was gathered from the eye canthus of rat into 1.5 mL heparinized tubes at 0.083, 0.167, 0.25, 0.5, 1, 2, 3, 6, 9, 12 and 24 h after gavage. Every blood sample were centrifuged immediately at 1920 g for 10 min at 4 °C to collect the supernatant. After that all collected plasma samples were combined and stored at −80 °C.

The bile collection: Each rat was injected 20% urethane solution intraperitoneally with 1–2 mL to anesthetize the rats after gavage. Then the rats were performed with bile duct cannulation operation and the bile samples were gathered during 0–1 h, 1–3 h, 3–5 h, 5–8 h, 8–12 h, 12–20 h and 20–24 h with PE-10 tubes (ID = 0.07 cm) [33,34]. Lastly, all bile samples were consolidated and frozen at −80 °C.

The urine and feces collection: The rats were separately housed in metabolic cages with free access to deionized water to collect the urine and feces samples over a 0–72 h period after gavage [35,36]. Finally, all the urine and feces samples were separately mixed, and they were placed at −80 °C before pretreatment was conducted.

3.3.3. Bio-Sample Pretreatment

All biological samples were disposed with two methods: Protein precipitation and liquid-liquid extraction were performed on the combined plasma, bile and urine with three times of methanol and ethyl acetate, respectively. Next, the mixture was vortexed for 5 min and centrifuged at 21,380× *g* for 10 min at 4 °C to obtain the supernatant, which was then collected and dried under nitrogen flow.

Dried and powdered feces samples were severally added to 3-fold methanol and ethyl acetate and then were ultrasonically extracted for 45 min. After centrifugation for 10 min at 21,380× *g*, they were dried under nitrogen gas like the supernatant in plasma, bile and urine samples.

150 μL methanol was added to the residua above with an ultrasonic operation for 15 min, centrifugation at 21,380× *g* for 10 min to gain the supernatant which were ultimately passed through the 0.22 μm millipore filters before injecting into the chromatographic system for further analysis. The control group was handled the same as the drug group.

3.4. Metabolism in Vitro by Rat Liver Microsomes

3.4.1. Phase I Metabolism

The typical incubation mixture was carried out in a PBS buffer (pH 7.4) with a final volume of 200 μL, which consisted of liver microsomal protein (1.0 mg/mL), eupatorin (100 μmol/L), $MgCl_2$ (3.3 mmol/L), and β-NADPH (1.3 mmol/L) [37]. Preincubation was conducted at 37 °C for 5 min, subsequently NADPH was added to start the reaction. After incubation at 37 °C for 90 min, the reaction was terminated by adding 1 mL of ethyl acetate. Next, vortex and centrifugation for 5 and 10 min, respectively, and then the organic phase was gathered and evaporated under nitrogen gas. 100 μL of acetonitrile was put in the residua and they were eventually passed through the 0.22 μm millipore filters and placed at −20 °C before analysis. Groups contained blank groups incubated without the addition of eupatorin, the control groups incubated without the addition of NADPH and the sample groups, which were implemented in triplicate with the same treatment [38,39].

3.4.2. Phase II Metabolism

The representative incubation mixture was performed in a PBS buffer (pH 7.4) with a final volume of 200 μL, which including liver microsomal protein (1.0 mg/mL), eupatorin (100 μmol/L), $MgCl_2$ (3.3 mmol/L), and UDPGA (2 mmol/L). Preincubation was implemented at 37 °C for 20 min, subsequently UDPGA was added to begin the reaction. After incubation at 37 °C for 1 h, the reaction was ceased by adding 200 μL of ice-acetonitrile. Next, vortex and centrifugation for 5 and 10 min, respectively. In addition, the supernatant was passed through the 0.22 μm millipore filter before injecting into the UHPLC-Q-TOF-MS/MS system for analysis. Groups contained blank groups incubated without the addition of eupatorin, the control groups incubated without the addition of UDPGA and the sample groups, which were carried out in triplicate with the same treatment.

3.5. Metabolism in Vitro by Rat Intestinal Flora

3.5.1. Preparation of Anaerobic Culture Medium

Solution A: K_2HPO_4 (0.78%) 37.5 mL; Solution B: KH_2PO_4 (0.47%), NaCl (1.18%), $(NH_4)_2SO_4$ (1.2%), $CaCl_2$ (0.12%) and $MgSO_4$ (0.25%) 37.5 mL; Solution C: Na_2CO_3 (8%) 50 mL; Solution D: L-ascorbic acid (25%) 2 mL together with L-cysteine 0.5 g, eurythrol 1 g, tryptone 1 g and nutrient agar 1 g, which were all mixed up. Ultrapure water was added to 1 L and then HCl (1 mol/L) was put to adjust the pH of the solution to 7.5–8.0.

3.5.2. Preparation of Intestinal Flora Culture Solution

Fresh intestinal contents (3 g) taken from SD rats were combined with anaerobic culture medium (30 mL) instantly. After stirring with a glass rod, filtered with gauze to obtain the intestinal bacterial liquid.

3.5.3. Sample Preparation

Eupatorin (1 mg/ mL,100 μL) was added to intestinal flora culture medium (1 mL), which was then filled with nitrogen without oxygen. The reactions were terminated by adding 3 volumes of methanol after incubation for 12 h. Next, the mixtures were vortexed for 5 min and centrifuged for 10 min at 21,380 g. Subsequently, the organic phases were collected and evaporated under nitrogen gas, and 100 μL of methanol was added to the residua, vortexed and centrifuged again for 5 and 10 min, respectively. Before analysis, the supernatant was passed through the 0.22 μm millipore filter. Blank groups were incubated without eupatorin, meanwhile the control groups were incubated not in intestinal flora culture solution but in anaerobic culture medium, but others were the same.

4. Conclusions

In conclusion, the identification of metabolites of eupatorin in vivo and in vitro had achieved great success firstly by means of UHPLC-Q-TOF-MS/MS combined with a powerful and efficient data acquisition and processing method. The results displayed that a total of 71 metabolites were characterized: 51 metabolites were identified in vivo (8 metabolites in the plasma, 5 metabolites in the bile, 36 metabolites in the urine and 32 metabolites in the feces), while 60 metabolites were detected in vitro (22 metabolites in the rat liver microsomes and 53 metabolites in rat intestinal flora). This study was expected to benefit future efficacy and safety studies on eupatorin and provide guidelines for intake of OS. There is no doubt that further studies are needed to confirm the impact of these metabolites on human health and safety, thus providing reasonable recommendations for the consumption of foods and drugs containing eupatorin.

Author Contributions: Conceptualization, L.-T.Z. and L.-Y.L.; investigation, L.-Y.L., X.F., J.-T.Y., S.-H.L. and Y.-P.S.; formal analysis, L.-Y.L., Y.-T.C. and X.F.; writing—original draft preparation, review, and editing, L.-Y.L. and Y.-T.C. All authors read and approved the final manuscript.

Funding: This work was funded by the National Natural Science Foundation of China, grant number 81473180.

Conflicts of Interest: All the authors have declared no conflicts of interest.

References

1. Pariyani, R.; Ismail, I.S.; Azam, A.; Khatib, A.; Abas, F.; Shaari, K.; Hamza, H. Urinary metabolic profiling of cisplatin nephrotoxicity and nephroprotective effects of *Orthosiphon stamineus* leaves elucidated by ^{1}H NMR spectroscopy. *J. Pharm. Biomed. Anal.* **2017**, *135*, 20–30. [CrossRef] [PubMed]
2. Hossain, M.A.; Rahman, S.M. Isolation and characterisation of flavonoids from the leaves of medicinal plant *Orthosiphon stamineus*. *Arab. J. Chem.* **2015**, *8*, 218–221. [CrossRef]
3. Yuliana, N.D.; Khatib, A.; Link-Struensee, A.M.; Ijzerman, A.P.; Rungkat-Zakaria, F.; Choi, Y.H.; Verpoorte, R. Adenosine A1 receptor binding activity of methoxy flavonoids from *Orthosiphon stamineus*. *Planta Med.* **2009**, *75*, 132–136. [CrossRef] [PubMed]
4. Ho, C.-H.; Noryati, I.; Sulaiman, S.-F.; Rosma, A. In vitro antibacterial and antioxidant activities of *Orthosiphon stamineus* Benth. extracts against food-borne bacteria. *Food Chem.* **2010**, *122*, 1168–1172. [CrossRef]
5. Zhang, J.; Wen, Q.; Qian, K.; Feng, Y.; Luo, Y.; Tan, T. Metabolic profile of rosmarinic acid from Java tea (*Orthosiphon stamineus*) by ultra-high-performance liquid chromatography coupled to quadrupole-time-of-flight tandem mass spectrometry with a three-step data mining strategy. *Biomed. Chromatogr.* **2019**. [CrossRef] [PubMed]
6. Guo, Z.; Liang, X.; Xie, Y. Qualitative and quantitative analysis on the chemical constituents in *Orthosiphon stamineus* Benth. using ultra high-performance liquid chromatography coupled with electrospray ionization tandem mass spectrometry. *J. Pharm. Biomed. Anal.* **2019**, *164*, 135–147. [CrossRef] [PubMed]
7. Saidan, N.H.; Aisha, A.F.A.; Hamil, M.S.R.; Majid, A.M.S.A.; Ismail, Z. A novel reverse phase high-performance liquid chromatography method for standardization of *Orthosiphon stamineus* leaf extracts. *Pharmacogn. Res.* **2015**, *7*, 23–31.
8. Yam, M.F.; Mohamed, E.A.H.; Ang, L.F.; Pei, L.; Darwis, Y.; Mahmud, R.; Asmawi, M.Z.; Basir, R.; Ahmad, M. A simple isocratic HPLC method for the simultaneous determination of sinensetin, eupatorin, and 3′-hydroxy-5,6,7,4′-tetramethoxyflavone in *Orthosiphon stamineus* extracts. *J. Acupunct. Meridian Stud.* **2012**, *5*, 176–182. [CrossRef] [PubMed]
9. Akowuah, G.; Zhari, I.; Norhayati, I.; Sadikun, A.; Khamsah, S. Sinensetin, eupatorin, 3′-hydroxy-5,6,7,4′-tetramethoxyflavone and rosmarinic acid contents and antioxidative effect of *Orthosiphon stamineus* from Malaysia. *Food Chem.* **2004**, *87*, 559–566. [CrossRef]
10. Razak, N.A.; Abu, N.; Ho, W.Y.; Zamberi, N.R.; Tan, S.W.; Alitheen, N.B.; Long, K.; Yeap, S.K. Cytotoxicity of eupatorin in MCF-7 and MDA-MB-231 human breast cancer cells via cell cycle arrest, anti-angiogenesis and induction of apoptosis. *Sci. Rep.* **2019**, *9*, 1514. [CrossRef]
11. Lee, K.; Hyun Lee, D.; Jung, Y.J.; Shin, S.Y.; Lee, Y.H. The natural flavone eupatorin induces cell cycle arrest at the G2/M phase and apoptosis in HeLa cells. *Appl. Biol. Chem.* **2016**, *59*, 193–199. [CrossRef]
12. Estevez, S.; Marrero, M.T.; Quintana, J.; Estevez, F. Eupatorin-induced cell death in human leukemia cells is dependent on caspases and activates the mitogen-activated protein kinase pathway. *PLoS ONE* **2014**, *9*, e112536. [CrossRef] [PubMed]
13. Androutsopoulos, V.; Arroo, R.R.J.; Hall, J.F.; Surichan, S.; Potter, G.A. Antiproliferative and cytostatic effects of the natural product eupatorin on MDA-MB-468 human breast cancer cells due to CYP1-mediated metabolism. *Breast Cancer Res.* **2008**, *10*. [CrossRef] [PubMed]
14. Doleckova, I.; Rarova, L.; Gruz, J.; Vondrusova, M.; Strnad, M.; Krystof, V. Antiproliferative and antiangiogenic effects of flavone eupatorin, an active constituent of chloroform extract of *Orthosiphon stamineus* leaves. *Fitoterapia* **2012**, *83*, 1000–1007. [CrossRef] [PubMed]

15. Laavola, M.; Nieminen, R.; Yam, M.F.; Sadikun, A.; Asmawi, M.Z.; Basir, R.; Welling, J.; Vapaatalo, H.; Korhonen, R.; Moilanen, E. Flavonoids eupatorin and sinensetin present in *Orthosiphon stamineus* leaves inhibit inflammatory gene expression and STAT1 activation. *Planta Med.* **2012**, *78*, 779–786. [CrossRef]
16. Yam, M.F.; Lim, V.; Salman, I.M.; Ameer, O.Z.; Ang, L.F.; Rosidah, N.; Abdulkarim, M.F.; Abdullah, G.Z.; Basir, R.; Sadikun, A.; et al. HPLC and anti-inflammatory studies of the flavonoid rich chloroform extract fraction of *Orthosiphon stamineus* leaves. *Molecules* **2010**, *15*, 4452–4466. [CrossRef]
17. Yam, M.F.; Tan, C.S.; Ahmad, M.; Shibao, R. Mechanism of vasorelaxation induced by eupatorin in the rats aortic ring. *Eur. J. Pharmacol.* **2016**, *789*, 27–36. [CrossRef]
18. Liao, M.; Cheng, X.; Diao, X.; Sun, Y.; Zhang, L. Metabolites identificaion of two bioactive constituents in Trollius ledebourii in rats using ultra-high-performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *J. Chromatogr. B* **2017**, *1068*, 297–312. [CrossRef]
19. Almazroo, O.A.; Miah, M.K.; Venkataramanan, R. Drug Metabolism in the Liver. *Clin. Liver Dis.* **2017**, *21*, 1–20. [CrossRef]
20. Li, H.; He, J.; Jia, W. The influence of gut microbiota on drug metabolism and toxicity. *Expert Opin. Drug Metab. Toxicol.* **2016**, *12*, 31–40. [CrossRef]
21. Noh, K.; Kang, Y.R.; Nepal, M.R.; Shakya, R.; Kang, M.J.; Kang, W.; Lee, S.; Jeong, H.G.; Jeong, T.C. Impact of gut microbiota on drug metabolism: An update for safe and effective use of drugs. *Arch. Pharm. Res.* **2017**, *40*, 1345–1355. [CrossRef] [PubMed]
22. Zhang, J.Y.; Wang, Z.J.; Li, Y.; Liu, Y.; Cai, W.; Li, C.; Lu, J.-Q.; Qiao, Y.-J. A strategy for comprehensive identification of sequential constituents using ultra-high-performance liquid chromatography coupled with linear ion trap-Orbitrap mass spectrometer, application study on chlorogenic acids in Flos Lonicerae Japonicae. *Talanta* **2016**, *147*, 16–27. [CrossRef] [PubMed]
23. Yao, D.; Wang, Y.; Huo, C.; Wu, Y.; Zhang, M.; Li, L.; Shi, Q.; Kiyota, H.; Shi, X. Study on the metabolites of isoalantolactone in vivo and in vitro by ultra performance liquid chromatography combined with Triple TOF mass spectrometry. *Food Chem.* **2017**, *214*, 328–338. [CrossRef] [PubMed]
24. Chen, Y.; Feng, X.; Li, L.; Zhang, X.; Song, K.; Diao, X.; Sun, Y.; Zhang, L. UHPLC-Q-TOF-MS/MS method based on four-step strategy for metabolites of hinokiflavone in vivo and in vitro. *J. Pharm. Biomed. Anal.* **2019**, *169*, 19–29. [CrossRef] [PubMed]
25. Yuan, L.; Liang, C.; Diao, X.; Cheng, X.; Liao, M.; Zhang, L. Metabolism studies on hydroxygenkwanin and genkwanin in human liver microsomes by UHPLC-Q-TOF-MS. *Xenobiotica* **2018**, *48*, 332–341. [CrossRef] [PubMed]
26. Zhang, X.; Yin, J.; Liang, C.; Sun, Y.; Zhang, L. A simple and sensitive UHPLC-Q-TOF-MS/MS method for sophoricoside metabolism study in vitro and *in vivo*. *J. Chromatogr. B* **2017**, *1061*, 193–208. [CrossRef]
27. Ma, Y.; Xie, W.; Tian, T.; Jin, Y.; Xu, H.; Zhang, K.; Du, Y. Identification and comparative oridonin metabolism in different species liver microsomes by using UPLC-Triple-TOF-MS/MS and PCA. *Anal. Biochem.* **2016**, *511*, 61–73. [CrossRef] [PubMed]
28. Liang, C.; Zhang, X.; Diao, X.; Liao, M.; Sun, Y.; Zhang, L. Metabolism profiling of nevadensin in vitro and in vivo by UHPLC-Q-TOF-MS/MS. *J. Chromatogr. B* **2018**, *1084*, 69–79. [CrossRef]
29. Zhang, X.; Yin, J.; Liang, C.; Sun, Y.; Zhang, L. UHPLC-Q-TOF-MS/MS Method Based on Four-Step Strategy for Metabolism Study of Fisetin in vitro and *in vivo*. *J. Agric. Food Chem.* **2017**, *65*, 10959–10972. [CrossRef]
30. Chen, X.; Han, R.; Hao, P.; Wang, L.; Liu, M.; Jin, M.; Kong, D.; Li, X. Nepetin inhibits IL-1beta induced inflammation via NF-kappaB and MAPKs signaling pathways in ARPE-19 cells. *Biomed. Pharmacother.* **2018**, *101*, 87–93. [CrossRef]
31. Clavin, M.; Gorzalczany, S.; Macho, A.; Muñoz, E.; Ferraro, G.; Acevedo, C.; Martino, V. Anti-inflammatory activity of flavonoids from Eupatorium arnottianum. *J. Ethnopharmacol.* **2007**, *112*, 585–589. [CrossRef] [PubMed]
32. Lee, Y.S.; Yang, W.K.; Yee, S.M.; Kim, S.M.; Park, Y.C.; Shin, H.J.; Han, C.K.; Lee, Y.C.; Kang, H.S.; Kim, S.H. KGC3P attenuates ovalbumin-induced airway inflammation through downregulation of p-PTEN in asthmatic mice. *Phytomedicine* **2019**, *62*, 152942. [CrossRef] [PubMed]
33. Liao, M.; Diao, X.; Cheng, X.; Sun, Y.; Zhang, L. Nontargeted SWATH acquisition mode for metabolites identification of osthole in rats using ultra-high-performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *RSC Adv.* **2018**, *8*, 14925–14935. [CrossRef]

34. Diao, X.; Liao, M.; Cheng, X.; Liang, C.; Sun, Y.; Zhang, X.; Zhang, L. Identification of metabolites of Helicid in vivo using ultra-high performance liquid chromatography-quadrupole time-of-flight mass spectrometry. *Biomed. Chromatogr.* **2018**, *32*, e4263. [CrossRef] [PubMed]
35. Xie, W.; Jin, Y.; Hou, L.; Ma, Y.; Xu, H.; Zhang, K.; Zhang, L.; Du, Y. A practical strategy for the characterization of ponicidin metabolites in vivo and in vitro by UHPLC-Q-TOF-MS based on nontargeted SWATH data acquisition. *J. Pharm. Biomed. Anal.* **2017**, *145*, 865–878. [CrossRef] [PubMed]
36. Tian, T.; Jin, Y.; Ma, Y.; Xie, W.; Xu, H.; Zhang, K.; Zhang, L.; Du, Y. Identification of metabolites of oridonin in rats with a single run on UPLC-Triple-TOF-MS/MS system based on multiple mass defect filter data acquisition and multiple data processing techniques. *J. Chromatogr. B* **2015**, *1006*, 80–92. [CrossRef] [PubMed]
37. Zhang, X.; Liang, C.; Yin, J.; Sun, Y.; Zhang, L. Identification of metabolites of liquiritin in rats by UHPLC-Q-TOF-MS/MS: Metabolic profiling and pathway comparison in vitro and in vivo. *RSC Adv.* **2018**, *8*, 11813–11827. [CrossRef]
38. Jia, P.; Zhang, Y.; Zhang, Q.; Sun, Y.; Yang, H.; Shi, H.; Zhang, X.; Zhang, L. Metabolism studies on prim-O-glucosylcimifugin and cimifugin in human liver microsomes by ultra-performance liquid chromatography quadrupole time-of-flight mass spectrometry. *Biomed. Chromatogr.* **2016**, *30*, 1498–1505. [CrossRef]
39. Yisimayili, Z.; Guo, X.; Liu, H.; Xu, Z.; Abdulla, R.; Aisa, H.A.; Huang, C. Metabolic profiling analysis of corilagin in vivo and in vitro using high-performance liquid chromatography quadrupole time-of-flight mass spectrometry. *J. Pharm. Biomed. Anal.* **2019**, *165*, 251–260. [CrossRef]

Sample Availability: Samples of the compounds are available from the authors.

molecules

MDPI

Article

Artificial Neural Network Prediction of Retention of Amino Acids in Reversed-Phase HPLC under Application of Linear Organic Modifier Gradients and/or pH Gradients

Angelo Antonio D'Archivio

Dipartimento di Scienze Fisiche e Chimiche, Università degli Studi dell'Aquila, Via Vetoio, 67100 Coppito, L'Aquila, Italy; angeloantonio.darchivio@univaq.it; Tel.: +39-0862-433777

Received: 21 January 2019; Accepted: 7 February 2019; Published: 11 February 2019

Abstract: A multi-layer artificial neural network (ANN) was used to model the retention behavior of 16 *o*-phthalaldehyde derivatives of amino acids in reversed-phase liquid chromatography under application of various gradient elution modes. The retention data, taken from literature, were collected in acetonitrile–water eluents under application of linear organic modifier gradients (φ gradients), pH gradients, or double pH/φ gradients. At first, retention data collected in φ gradients and pH gradients were modeled separately, while these were successively combined in one dataset and fitted simultaneously. Specific ANN-based models were generated by combining the descriptors of the gradient profiles with 16 inputs representing the amino acids and providing the retention time of these solutes as the response. Categorical "bit-string" descriptors were adopted to identify the solutes, which allowed simultaneously modeling the retention times of all 16 target amino acids. The ANN-based models tested on external gradients provided mean errors for the predicted retention times of 1.1% (φ gradients), 1.4% (pH gradients), 2.5% (combined φ and pH gradients), and 2.5% (double pH/φ gradients). The accuracy of ANN prediction was better than that previously obtained by fitting of the same data with retention models based on the solution of the fundamental equation of gradient elution.

Keywords: amino acids; reversed-phase liquid chromatography; gradient elution; retention prediction; artificial neural network

1. Introduction

Reversed-phase high-performance liquid chromatography (RP-HPLC) is an extensively applied technique in the separation and determination of a wide range of multi-class compounds, including biomolecules, pharmaceuticals, and industrial chemicals, in human, environmental, and food samples [1–4]. Separation of complex mixtures by RP-HPLC generally requires the application of mobile-phase gradients to overcome the typical disadvantages of isocratic elution, such as poor resolution of early peaks, broadening of late peaks, band tailing, and long separation times [5,6]. In organic modifier mobile-phase gradients (φ gradients), the concentration of organic solvent in the mobile phase is increased, determining a progressive increase of the elution power of the eluent during the gradient run and a consequent decrease in solute retention. A similar effect occurs in the pH gradient of the mobile phase [7], where an increase or decrease in pH in the case of weak bases or acids, respectively, produces a progressive increase of the ionized form of the analyte and a consequent decrease in its retention time.

In the last few decades, various predictive models [8–11] were proposed with the aim of supporting the empirical strategies commonly utilized in the development of the chromatographic methods, which can be particularly slow and inefficient when a large number of parameters have to be fixed, such as in the case of programmed elution analysis.

Many attempts to describe the retention of solutes in RP-HPLC under the application of mobile-phase gradients are based on the solution of the fundamental equation of gradient elution [12–16],

$$\int_0^{t_R - t_0} \frac{dt}{t_0 k} = 1 \tag{1}$$

where t_R is the retention time, t_0 is the column hold-up time, and k is the retention factor. Analytical or numerical solutions of Equation (1) require the dependence of k upon the mobile-phase composition. To this end, popular relationships relating k and φ, or empirical models arising from the experimental properties of the system are often used, where the adjustable eluent- and sometimes solute-dependent parameters associated with these relationships are determined by appropriate fitting algorithms applied to the retention data.

Artificial neural networks (ANNs), since their introduction in 1990s, are used as regression tools to address various complex issues in chromatography. The main advantage of ANN regression is that both multilinear and non-linear phenomena can be handled without the need of prior definition of a fitting function. The ANN-based applications in retention prediction include the development of quantitative structure–retention relationships (QSSRs) [17,18], modeling of the combined effects of solute structure and separation conditions (column, eluent, or both) [19,20], and transfer of retention data between different columns or eluent types [21–23]. ANN models based simultaneously on molecular descriptors and instrumental conditions associated with the elution mode were used to predict the retention times of diverse sets of organic compounds in gradient RP-HPLC [24–27]. We previously used ANN regression to model the retention times of 16 selected purines, pyrimidines, and nucleosides under the application of multilinear φ gradients [28]. With this aim, a network was trained to associate the retention times with both gradient profiles and solutes, the latter being represented by "bit-string" categorical descriptors, which, unlike the aforementioned QSSR-inspired approaches, did not require any assumption of the chemical structure of the analytes. The generalization ability of the so-obtained model was tested on external multilinear gradients, providing an accurate prediction of the solute retention times (within 2–3% on average). This approach was here extended to the RP-HPLC retention of ionizable solutes of biological relevance, such as amino acids, analyzed under the application of linear φ gradients, pH gradients, or combined φ/pH gradients, whereby the target compounds were previously derivatized with *o*-phthalaldehyde (OPA) to allow their fluorescence detection. The data investigated in the present study were taken from three works of Pappa-Louisi and co-workers [14–16], who collected the experimental data and developed retention models based on the solution of the fundamental equation of gradient elution to verify the accuracy of the predicted retention by different equations or fitting algorithms.

The present study is aimed at exploring the capability of ANN regression calibrated with the retention data collected in representative gradients to predict the chromatographic behavior of ionizable solutes in external separation conditions. Retention in gradient RP-HPLC is governed by several factors, such as the chemical structure of solutes, their acid–base properties, the polarity/acidity of the mobile phase, and how these properties change during the chromatographic run. While, on the one hand, ANN is potentially able to treat such complexity, on the other, the network does not provide a fitting equation that could be useful for getting information about the relative role of the different factors in the retention process. Nevertheless, finding the optimal condition for the chromatographic separation of a complex mixture, which is anyway a multivariate problem, can be handled by statistical retention models, but their predictive performance is more important than the knowledge of their physical meaning. In this view, a network was trained to associate the experimental parameters describing the gradient elution profile with the retention times of a mixture of target analytes to be separated. The ANN-based model, once calibrated on a sufficiently large set of representative separation conditions, was later applied to simultaneously predict the retention times of all the solutes in external elution conditions. In the end, the ANN response can be useful for optimization purposes, because it allows deducing the retention of the target solutes at any point of the experimental domain explored in calibration, and it may replace or support inefficient trial-and-error empirical approaches usually adopted to search the optimal separation conditions. At first, two separate ANN-based retention models were generated to predict the data collected under application of linear φ gradients or pH gradients. In addition, the retention behavior of the amino acids under the independent or simultaneous application of linear φ and pH gradients was modeled by ANN. The predictive performance of the various ANN-based models developed in this work was compared with the prediction ability of the retention models based on the solution of the fundamental equation of gradient elution.

2. Results

2.1. Identification of Model Variables and Data Subsets

In this paper, ANN regression was used to model the RP-HPLC retention of *o*-phthaladehyde (OPA) derivatives of 16 amino acids collected under the application of φ gradients, pH gradients, or combined pH/φ gradients. The retention datasets (A, B, and C, respectively), taken from the literature [14–16], are described in Section 3.1. The following variables were considered to describe the linear φ gradients of dataset A: the starting pH (pH_i), the starting organic solvent content (φ_i), and the φ-gradient slope ($\Delta\varphi/t_g = (\varphi_f - \varphi_i)/t_g$), where φ_f is the φ value at the end of gradient run and t_g is the gradient time. The pH gradients of dataset B were described by φ_i, pH_i and the gradient slope ($\Delta pH/t_g = (pH_f - pH_i)/t_g$, where pH_f is the final pH value). The respective values of the above quantities, determined from the experimental conditions reported in the original papers [14,16], are collected in Table 1. Among these parameters, the constant ones ($\Delta pH/t_g$ and φ_i in dataset A, and $\Delta\varphi/t_g$ in dataset B) were not considered as network inputs in ANN modeling of φ gradients or pH gradients. Datasets A and B were successively fused in one comprehensive dataset, hereafter indicated as A+B, to attempt ANN modeling of retention data collected under independent applications of φ gradients and pH gradients. In this case, all four gradient descriptors reported in Table 1 are informative and were considered as ANN inputs. To describe the 27 double pH/φ gradients of dataset C (referring to double pH/φ gradients), the three non-constant experimental quantities (pH_f, φ_f, and t_g) varying according to a three-level experimental design in Reference [15] were assumed as ANN inputs. The level values selected for these variables are given in Section 3.1.

As described in Section 3.2, ANN regression requires a training set, which is processed to update the network weights and biases; however, the network performance must also be monitored during learning using unknown data (validation set) to avoid overfitting. Moreover, the real generalization ability of the learned network must be finally evaluated on external data (test set) neither used in training nor in validation. To design these three datasets, the various φ gradients of dataset A, the pH gradients of dataset B, and the φ/pH gradients of dataset C were graphically represented in the space of the variables previously selected to describe the changes in the eluent composition (Figure 1). These plots helped us generate three well-balanced subsets in terms of representativeness; the data samples assigned to each subset were selected to cover the investigated experimental domain as much as homogeneously possible. Regardless of the dataset, six gradients were selected for the final test; three gradients (dataset A) or four gradients (datasets B and C) were selected for the internal validation and the remaining elution conditions were used to train the networks (Table 1 and Figure 1). The training, validation, and test sets for dataset A+B were designed by fusing the respective subsets of the A and B matrices. Considering that the retention data of 16 amino acids are associated with each experimental elution mode, the training, validation, and test data points were 160, 48, and 96, respectively, for dataset A; 192, 64, and 96, respectively, for dataset B; 352, 112, and 196, respectively, for dataset A+B; and 272, 64, and 96 for dataset C. Rather than representing the solutes by molecular descriptors, according to conventional QSRR approach, each of the 16 amino acids was identified by a

16-bit string, consisting of all “0” values except the *n*-th bit, which was set to “1”, where *n* corresponds to the position of that solute in an arbitrary and predefined sequence of the investigated analytes. In this condition, the network was trained to properly associate the retention times to both solutes and gradient modes, without any explicit reference to the solute molecular structure.

2.2. ANN Modeling of Retention

The distinct networks handling the retention datasets A, B, A + B, and C were optimized following a usual procedure aimed at founding the combination of the ANN adjustable parameters providing the lowest validation error. A range-scaling between 0 and 1 was always applied to both input and output variables. Retention time (t_R(min)) values and their logarithmic values were alternatively considered as the ANN responses. Both options provided good ANN models and a random distribution of absolute residuals; however, logarithmic transformation of retention times was preferred to the unscaled values because it gave lower relative errors for the less retained amino acids.

Table 1. Descriptors of the linear φ gradients (dataset A) and pH gradients (dataset B).

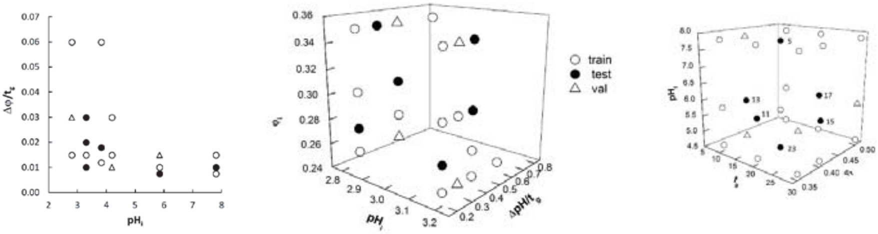
Dataset	Gradient Code	Subset [a]	pH_i	$\Delta pH/t_g$	φ_i	$\Delta\varphi/t_g$
A	1A	train	2.8	0	0.20	0.06
	2A	val	2.8	0	0.20	0.03
	3A	train	2.8	0	0.20	0.015
	4A	test	3.3	0	0.20	0.03
	5A	test	3.3	0	0.20	0.02
	6A	train	3.3	0	0.20	0.015
	7A	test	3.3	0	0.20	0.01
	8A	train	3.82	0	0.20	0.06
	9A	test	3.82	0	0.20	0.018
	10A	train	3.82	0	0.20	0.012
	11A	train	4.2	0	0.20	0.03
	12A	train	4.2	0	0.20	0.015
	13A	val	4.2	0	0.20	0.01
	14A	val	5.85	0	0.20	0.015
	15A	train	5.85	0	0.20	0.01
	16A	test	5.85	0	0.20	0.0075
	17A	train	7.8	0	0.20	0.015
	18A	test	7.8	0	0.20	0.01
	19A	train	7.8	0	0.20	0.0075
B	1B	train	2.8	0.79	0.35	0
	2B	val	2.8	0.527	0.35	0
	3B	test	2.8	0.395	0.35	0
	4B	train	2.8	0.263	0.35	0
	5B	val	2.8	0.527	0.25	0
	6B	train	2.8	0.527	0.27	0
	7B	test	2.8	0.527	0.30	0
	8B	train	2.8	0.263	0.25	0
	9B	test	2.8	0.263	0.27	0
	10B	train	2.8	0.263	0.30	0
	11B	train	3.2	0.580	0.25	0
	12B	train	3.2	0.387	0.25	0
	13B	val	3.2	0.290	0.25	0
	14B	train	3.2	0.193	0.25	0
	15B	train	3.2	0.387	0.27	0
	16B	test	3.2	0.387	0.30	0
	17B	test	3.2	0.387	0.35	0
	18B	train	3.2	0.290	0.30	0

Table 1. *Cont.*

Dataset	Gradient Code	Subset [a]	pH_i	$\Delta pH/t_g$	φ_i	$\Delta\varphi/t_g$
B	19B	val	3.2	0.290	0.35	0
	20B	test	3.2	0.193	0.27	0
	21B	train	3.2	0.193	0.30	0
	22B	train	3.2	0.193	0.35	0

[a] Training set (train), validation set (val), test set (test).

Based on the results of preliminary ANN runs, in which a sigmoid or a tangent hyperbolic activation function was applied to the hidden neurons, the latter was preferred, while application of a non-linear transformation in the output neuron was not required because it did not produce any improvement in the model performance. The number of hidden neurons was varied in the range between $N - 6$ and $N + 6$, where N is the number of inputs, and each tested network was trained until the validation error reached a minimum value.



(**a**) (**b**) (**c**)

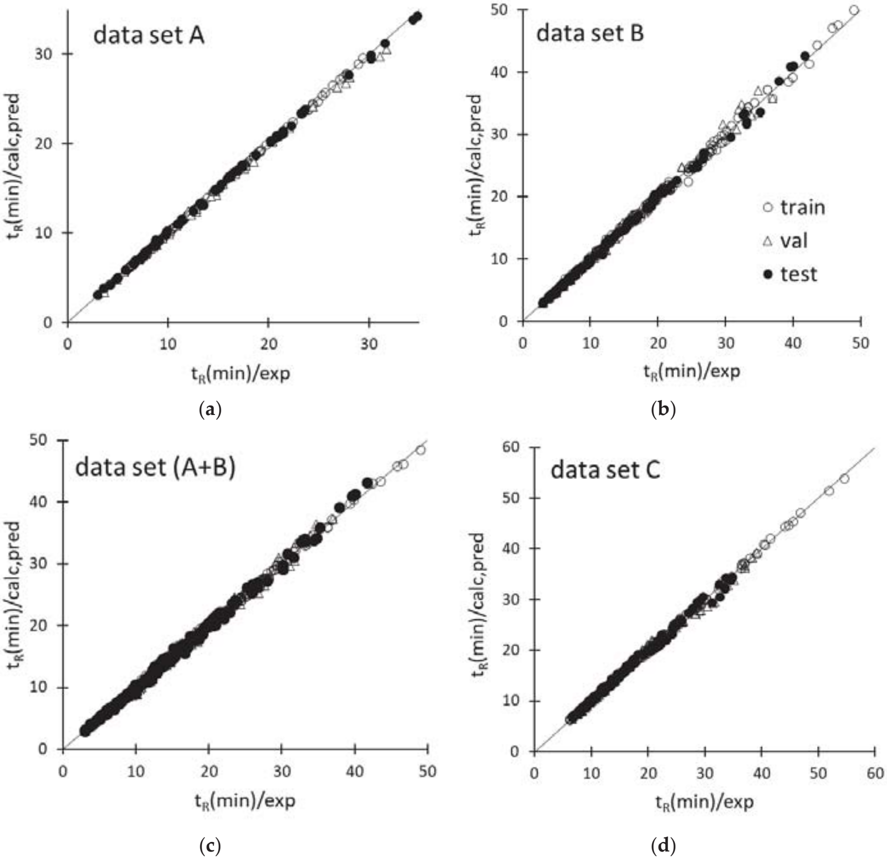
Figure 1. Gradients used in artificial neural network (ANN) training, validation, and test data projected in the space of the variables adopted as network inputs for datasets A (**a**), B (**b**), and C (**c**). Test samples of dataset C are labeled according to Reference [15].

The best ANN architectures and learning durations are presented in Table 2. Because of a randomization of the starting weights, here generated between −0.1 and 0.1, the optimal network produced slightly different responses upon being re-trained several times. To minimize the influence of the initial weights on the ANN-based model performance, the network was re-trained 100 times and the outputs were averaged. The agreement between computed or predicted ANN responses and the experimental t_R values for each retention dataset are graphically shown in Figure 2. Table 2 displays the determination coefficients in calibration and prediction (R^2 and Q^2) and the related standard errors (SEC and SEP, respectively) associated with the ANN-based models, where Q^2 was determined according to Todeschini et al. [29]. The average and maximum absolute percentage errors (mean(%) and max(%), respectively) in each subset are also reported. All the above statistical parameters refer to the unscaled t_R values.

Table 2. Description of the ANN models developed on the retention datasets A, B, A+B, and C: architecture and learning duration of the optimal network, coefficients of determination in training (R^2) and prediction (Q^2) and respective standard errors (SEC and SEP), and the mean and maximum percentage errors (mean(%) and max(%)).

Data Set	Network Topology	Learning Epochs	Training				Validation				Test			
			R^2	SEC	mean(%)	max(%)	Q^2	SEP	mean(%)	max(%)	Q^2	SEP	mean(%)	max(%)
A	18-14-1[a]	251	0.9999	0.06	0.3	1.6	0.9906	0.37	1.6	7.4	0.9984	0.22	1.4	6.4
B	19-21-1	63	0.9980	0.46	0.7	4.1	0.9778	0.78	1.4	4.1	0.9949	0.48	1.1	5.3
A+B	20-23-1	286	0.9993	0.23	1.2	6.3	0.9939	0.65	3.3	12.6	0.9799	0.48	2.5	10.4
C	19-18-1	125	0.9994	0.22	1.0	4.2	0.9938	0.72	2.6	6.9	0.9958	0.59	2.5	6.8

[a] Number of neurons in the input, hidden and output layer, respectively.



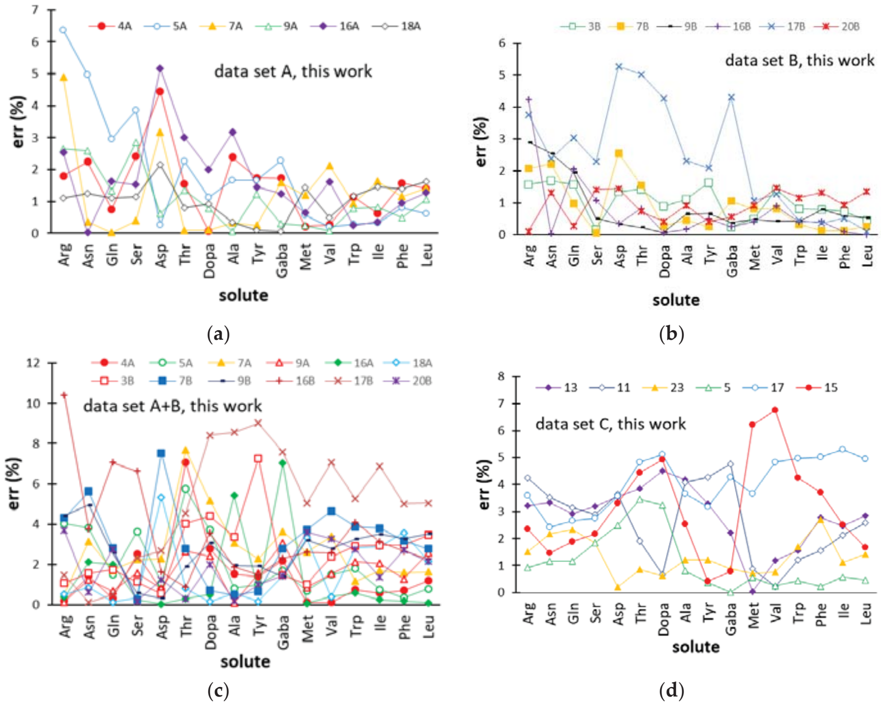
(**a**) (**b**) (**c**) (**d**)

Figure 2. Agreement between the experimental retention times (t_R(min)/exp) of solutes and calculated or predicted ANN responses (t_R(min)/calc,pred) of datasets A (**a**), B (**b**), A+B (**c**), and D (**d**).

2.3. Predictive Performance of the ANN-Based Models

Inspection of the agreement plots for the various retention datasets modeled by ANN (Figure 2) reveals that both computed and predicted responses were very close to the ideal line, ensuring an accurate prediction of the retention times of the amino acids within the respective experimental domains. As expected, the training data samples were better modeled than the validation and test data; nonetheless, worsening of the predictive performance as compared to the fitting ability was slight, as confirmed by the small differences among the statistical parameters of training, validation, and test sets, reported in Table 2. The data samples were also randomly distributed around the ideal line of the agreement plots, suggesting the absence of systematic errors, except for dataset C, for which a small group of validation cases in the t_R range between 30 and 40 min were all underestimated (Figure 2d). Most of these data samples were associated with the most retained amino acids analyzed under the application of a same gradient ($\varphi_f = 0.5$, $pH_f = 5.86$, $t_g = 30$ min), but the errors were anyway acceptable (within 4–7%). The retention data collected under the application of φ gradients and pH gradients were very well modeled according to the mean errors, which were smaller than 1% and 1.5% for the training and test data, respectively (Table 2). Only a slight worsening of the descriptive/predictive ANN performance was observed when the network was called to model the retention times of the amino acids under the independent application of φ gradients and pH gradients (dataset A+B) or double pH/φ gradients; the mean error in both cases was just above 1% for the training set and 2.5% for the test set (Table 2).

Figure 3 displays the trend of the relative (%) errors (err (%)) for the retention times of the 16 amino acids in the φ gradients and/or pH gradients of the test set. Therefore, these data quantify the ability of the ANN-based models to predict the retention of the amino acids in elution conditions external to those used in calibration, and give a measure of the applicability of this approach in optimization problems.



(**a**) (**b**)

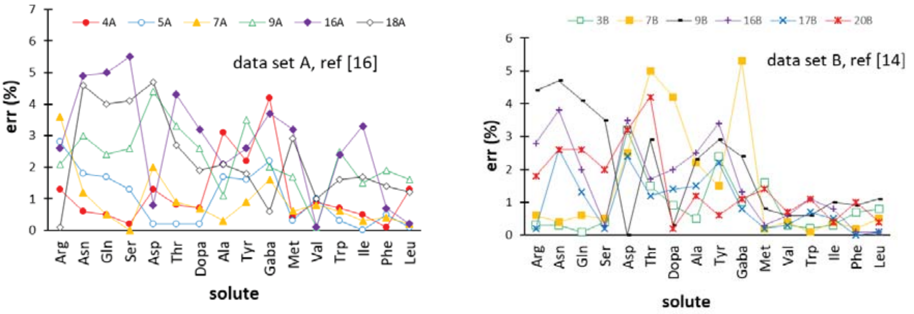
(**c**) (**d**)

Figure 3. Percentage errors (err (%)) for the retention times of the amino acids provided by the ANN-based models in the external gradients (test set) of datasets A (**a**), B (**b**), A+B (**c**), and C (**d**). Abbreviations used for the amino acids are reported in Section 3.1. Gradient codes are specified in Table 1 (datasets A, B, and A+B) and Figure 1 (dataset C).

For most gradients of dataset A (φ gradients) and B (pH gradients), err (%) almost regularly decreased, passing from the less retained (Arg) to the most retained amino acid (Leu), seen from the left to the right of the plots displayed in Figure 3a,b. This arose from the fact that the absolute errors were homogeneously distributed over the target amino acids and, therefore, the relative errors were inversely related to the t_R value. Most of the predicted errors associated with the 16 amino acids in the external gradients of datasets A and B were smaller than 3%, while ANN modeling of datasets A+B (Figure 3c) and C (Figure 3d) provided slightly greater residuals, although generally below 5%. It can be noted that the retention times of most amino acids were less accurately predicted in the pH gradient 17B, when the data referring to pH gradients were modeled separately (dataset B) and when pH gradients and φ gradients were combined (dataset A+B). The moderately worse performance of the ANN model in this experimental condition can be due to the fact that the values of the two eluent descriptors (φ_i and pH_i) of pH gradient 17B were the greatest within the respective variability ranges (Table 1) and, therefore, the network was called to extrapolate the response.

2.4. Comparison of the ANN-Based Models with Retention Models Based on the Solution of the Fundamental Equation of Gradient Elution

The error trends provided by the retention models based on the solution of the fundamental equation of gradient elution that Pappa-Louisi and co-workers applied to datasets A [16] and B [14] are displayed in Figure 4 for comparison purposes. With regards to dataset A, the ANN-based model gave a lower number of errors above 3% as compared with the retention model developed in Reference [16]. Concerning dataset B, it must be noted that the pH gradient retention data collected in the pH ranges of 2.8–10.7 and 3.2–9 (in Table 1, gradients 1B–10B and 11B–22B, respectively) were fitted by two separate models in Reference [14], while, in this work, all 22 elution conditions were modeled by the same network. Nevertheless, the comprehensive ANN-based model built here seems to give a better prediction of the retention times, whereby the number of errors above 2% was lower as compared with the results provided by the two separate retention models generated from the solution of the fundamental equation of gradient elution. Although t_R values of the most retained amino acids (Val, Trp, Ile, Phe, and Leu) were predicted by the two alternative approaches with a comparable accuracy (errors were close to 1% or lower), the behavior of the less retained solutes was better described by the ANN model.



(**a**) (**b**)

Figure 4. Percentage errors (err (%)) for the retention times of the amino acids provided by the retention models developed in References [14,16] for the external gradients (test set) of datasets A (**a**) and B (**b**). Abbreviations used for the amino acids are reported in Section 3.1. Gradient codes are specified in Table 1.

The comparison of Figures 3c and 4a,b reveals that the accuracy of prediction in the external φ gradients and pH gradients of dataset A+B was substantially equivalent to that provided by the solution of the fundamental equation of gradient elution. However, it should be remarked that a single ANN-based model was required to fit these data, while the data collected in pH gradients and φ gradients covering two different pH ranges were interpolated with three different retention models in References [14,16].

The ANN model describing retention under the application of double pH/φ gradients (dataset C) exhibited individual t_R errors in the external gradients that surpassed 5% only in a limited number of cases (Figure 3d). For this dataset, instead of the detailed trend of errors, not given in Reference [15], the mean errors provided by the retention model obtained from the solution of the fundamental equation of gradient elution could be considered for comparison. The mean and maximum errors reported for the model calibrated with all 27 gradients of dataset C were 2.9% and 18.9%, respectively. Moreover, the mean error associated with individual amino acids over the 27 gradients monotonically grew with the increase in retention time, from 1.5% (Arg) up to 6.5% (Leu) (Figure 5 of Reference [15]), revealing a poor modeling of the retention behavior of the most retained solutes. In the present work, the mean and maximum errors for the 17 gradients used to train the network were noticeably lower (1.0 and 4.2%, Table 2), and we observed a substantial independence of the training errors from the kind of amino acid. In Reference [15], the model initially developed using all the 27 gradients was recalibrated with 18 gradients and applied to the remaining nine gradients providing mean and maximum errors of 3.5 and 11.8%, respectively. In the present work, the network trained with 17 gradients gave lower mean and maximum errors both in internal validation (2.6 and 6.9%) and external prediction (2.5 and 6.8%). In summary, the ANN-based model, as compared with the retention models generated from the solution of the fundamental equation of gradient elution, provided a more accurate prediction of the retention times of the amino acids in double pH/φ gradients, as well as a more homogenous error distribution.

3. Methods

3.1. Retention Data

The data here analyzed were taken from three papers of Pappa-Louisi and co-workers [14–16] regarding the RP-HPLC retention of OPA derivatives of amino acids collected under the application of φ gradients, pH gradients, or combined pH/φ gradients. The mobile phases consisted of mixtures of aqueous phosphate buffer with a total ionic strength of 0.02 M and acetonitrile. In the first paper [16], 19 chromatographic runs were performed in different fixed eluent pHs (between 2.80 and 7.80), while the organic solvent volume fraction φ was linearly varied between 0.2 and 0.5 in different gradient durations (t_g, ranging between 5 and 40 min). In the second paper [14], φ was kept fixed (at 0.25, 0.27, 0.3, or 0.35), and 22 different linear pH gradients were applied in the pH ranges of 2.8–10.7 or 3.2–9, where t_g was varied between 10 and 30 min. A third retention dataset (dataset C) was collected by Zisi et al. [15] under the application of a double organic solvent and pH gradient, in which both φ and pH were linearly changed from initial values (φ_i and pH_i) to final values (φ_f and pH_f). This consisted of 27 different runs performed at fixed values of φ_i (0.25) and pH_i (3.21), while pH_f, φ_f, and t_g were varied according to a three-level experimental design. The selected levels were 4.68, 5.86, and 7.86 for pH_f; 0.35, 0.40, and 0.50 for φ_f; and 10, 20, and 30 min for t_g.

The amino acids analyzed in the above conditions were as follows: L-arginine (Arg), L-asparagine (Asn), L-glutamine (Gln), L-serine (Ser), L-aspartic acid (Asp), L-threonine (Thr), beta-(3,4-dihydroxyphenyl)-L-alanine (Dopa), L-alanine (Ala), L-tyrosine (Tyr), 4-aminobutyric acid (GABA), L-methionine (Met), L-valine (Val), L-tryptophan (Trp), L-isoleucine (Ile), L-phenylanine (Phe), and L-leucine (Leu). The amino acid L-glutamic acid (Glu), which was analyzed only in some experimental conditions, was not considered here. Apart from the different gradient profiles, all the retention data were collected with the same column, detector, and eluent flow rate. A 250 mm × 4.6 mm MZ-PerfectSil Target ODS-3HD analytical column with a 5-μm particle size kept at 25 °C was used, and the spectrofluorometric detector worked at 455 nm after excitation at 340 nm. Further experimental details can be found in the original papers [14–16].

3.2. Artificial Neural Network Modelling

A three-layer feed-forward ANN [30,31] was used in this work. The network consisted of one layer of input neurons, one output neuron, and an adjustable number of neurons in the hidden layer, fully connected to both the input and output neurons. Weights were associated to the connections, which modulated the information flowing from the input layer collecting the independent variables to the output neuron providing the network response. The weighted input variables entering each neuron of the hidden layer were summed and added to a bias value, and the result was transformed by a non-linear activation function, providing an output signal. The output neuron operated in a similar way on the weighted outputs of the hidden neurons producing the final response. A starting set of weights and biases, randomly generated, was sequentially updated in a learning (or training) procedure in which the network evaluated several input/output pairs (training set) to produce the best agreement between the target and computed responses. The optimized set of weights and biases, which represented a sort of memory of the learned network, could later be recalled, making predictions of the unknown response when the predictors were known. In this work, the network was trained by a quasi-Newton method [31], which incorporates second-order information about the error surface shape, ensuring faster convergence and a greater probability of avoiding local minima as compared to the classical error backpropagation learning algorithm. To avoid overfitting, the ANN performance during the learning step was monitored on unknown data samples (validation set), and the weight update was interrupted when the validation error started increasing. Minimization of the validation error was the criterion also adopted to select among alternative ANN models, differing in their network architecture, kind of activation function, kind of data scaling, and so on, the one with the best expected generalization ability. The real predictive performance of the final ANN-based model was finally evaluated on data samples (test set) external to both the training and validation sets. Software OpenNN [32] was used to perform ANN modeling.

4. Conclusions

In this paper, a three-layer artificial neural network was used to model the retention times of 16 amino acids under the separate or simultaneous application of linear organic modifier and pH gradients. We focused on the ANN's capability to predict the retention data of the target solutes in external gradients, which is a useful response for optimization purposes. Using a "bit-string" representation of solutes allowed simultaneously modeling the retention behavior of all 16 amino acids with no explicit reference to their chemical structure or properties. It follows that the approach presented in this work can be transferred to chemical classes or heterogeneous groups of solutes different from those investigated. Moreover, the model generation did not require any assumption concerning the dependence of the retention factors on the eluent pH and composition, which is, by contrast, a prerequisite to attempt the solution of the fundamental equation of gradient elution. The predictive ability of the ANN-based models tested on external gradients was very good, whereby the mean errors for the retention times were 1.1% for φ gradients, 1.4% for pH gradients, and 2.5% for pH/φ gradients, and better than that provided by retention models based on the solution of the fundamental equation of gradient elution. In summary, ANN modeling seems a powerful and flexible regression tool to describe the effect of the experimental conditions in linear gradient elution on the retention of ionizable solutes and, in combination with experimental design, can be applied to optimize HPLC methods.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fekete, S.; Veuthey, J.L.; Guillarme, D. New trends in reversed-phase liquid chromatographic separations of therapeutic peptides and proteins: Theory and applications. *J. Pharm. Biomed. Anal.* **2012**, *69*, 9–27. [CrossRef] [PubMed]
2. Domínguez-Álvarez, J.; Mateos-Vivas, M.; Rodríguez-Gonzalo, E.; García-Gómez, D.; Bustamante-Rangel, M.; Delgado Zamarreño, M.M.; Carabias-Martínez, R. Determination of nucleosides and nucleotides in food samples by using liquid chromatography and capillary electrophoresis. *TrAC Trends Anal. Chem.* **2017**, *92*, 12–31. [CrossRef]
3. Mazzeo, P.; Di Pasquale, D.; Ruggieri, F.; Fanelli, M.; D'Archivio, A.A.; Carlucci, G. HPLC with diode-array detection for the simultaneous determination of di(2-ethylhexyl)phthalate and mono(2-ethylhexyl)phthalate in seminal plasma. *Biomed. Chromatogr.* **2007**, *21*, 1166–1171. [CrossRef] [PubMed]
4. D'Archivio, A.A.; Maggi, M.A.; Ruggieri, F.; Carlucci, M.; Ferrone, V.; Carlucci, G. Optimisation by response surface methodology of microextraction by packed sorbent of non steroidal anti-inflammatory drugs and ultra-high performance liquid chromatography analysis of dialyzed samples. *J. Pharm. Biomed. Anal.* **2016**, *125*, 114–121. [CrossRef] [PubMed]

5. Fanali, S.; Haddad, P.R.; Poole, C.F.; Schoenmakers, P.; Lloyd, D. *Liquid Chromatography: Fundamentals and Instrumentation*; Elsevier: Amsterdam, The Netherlands, 2013; ISBN 9780124158078.
6. Jandera, P.; Churáček, J. Gradient elution in liquid chromatography. II. Retention characteristics (retention volume, band width, resolution, plate number) in solvent-programmed chromatography—Theoretical considerations. *J. Chromatogr. A* **1974**, *91*, 223–235. [CrossRef]
7. Kaliszan, R.; Wiczling, P.; Markuszewski, M.J. pH Gradient Reversed-Phase HPLC. *Anal. Chem.* **2004**, *76*, 749–760. [CrossRef]
8. Poole, C.F.; Lenca, N. Applications of the solvation parameter model in reversed-phase liquid chromatography. *J. Chromatogr. A* **2017**, *1486*, 2–19.
9. Vitha, M.; Carr, P.W. The chemical interpretation and practice of linear solvation energy relationships in chromatography. *J. Chromatogr. A* **2006**, *1126*, 143–194. [CrossRef]
10. Torres-Lapasió, J.R.; García-Alvarez-Coque, M.C.; Rosés, M.; Bosch, E.; Zissimos, A.M.; Abraham, M.H. Analysis of a solute polarity parameter in reversed-phase liquid chromatography on a linear solvation relationship basis. *Anal. Chim. Acta* **2004**, *515*, 209–227. [CrossRef]
11. Cela, R.; Ordoñez, E.Y.; Quintana, J.B.; Rodil, R. Chemometric-assisted method development in reversed-phase liquid chromatography. *J. Chromatogr. A* **2013**, *1287*, 2–22. [CrossRef]
12. Andrés, A.; Téllez, A.; Rosés, M.; Bosch, E. Chromatographic models to predict the elution of ionizable analytes by organic modifier gradient in reversed phase liquid chromatography. *J. Chromatogr. A* **2012**, *1247*, 71–80. [CrossRef] [PubMed]
13. Fasoula, S.; Zisi, C.; Gika, H.; Pappa-Louisi, A.; Nikitas, P. Retention prediction and separation optimization under multilinear gradient elution in liquid chromatography with Microsoft Excel macros. *J. Chromatogr. A* **2015**, *1395*, 109–115. [CrossRef] [PubMed]
14. Pappa-Louisi, A.; Zisi, C. A simple approach for retention prediction in the pH-gradient reversed-phase liquid chromatography. *Talanta* **2012**, *93*, 279–284. [CrossRef]
15. Zisi, C.; Fasoula, S.; Nikitas, P.; Pappa-Louisi, A. Retention modeling in combined pH/organic solvent gradient reversed-phase HPLC. *Analyst* **2013**, *138*, 3771–3777. [CrossRef] [PubMed]
16. Fasoula, S.; Zisi, C.; Nikitas, P.; Pappa-Louisi, A. Retention prediction and separation optimization of ionizable analytes in reversed-phase liquid chromatography under organic modifier gradients in different eluent pHs. *J. Chromatogr. A* **2013**, *1305*, 131–138. [CrossRef] [PubMed]
17. Héberger, K. Quantitative structure-(chromatographic) retention relationships. *J. Chromatogr. A* **2007**, *1158*, 273–305. [CrossRef] [PubMed]
18. D'Archivio, A.A.; Incani, A.; Ruggieri, F. Retention modelling of polychlorinated biphenyls in comprehensive two-dimensional gas chromatography. *Anal. Bioanal. Chem.* **2011**, *399*, 903–913. [CrossRef]
19. D'Archivio, A.A.; Maggi, M.A.; Mazzeo, P.; Ruggieri, F. Quantitative structure-retention relationships of pesticides in reversed-phase high-performance liquid chromatography based on WHIM and GETAWAY molecular descriptors. *Anal. Chim. Acta* **2008**, *628*, 162–172. [CrossRef]
20. D'Archivio, A.A.; Maggi, M.A.; Ruggieri, F. Multiple-column RP-HPLC retention modelling based on solvatochromic or theoretical solute descriptors. *J. Sep. Sci.* **2010**, *33*, 155–166. [CrossRef]
21. D'Archivio, A.A.; Giannitto, A.; Maggi, M.A. Cross-column prediction of gas-chromatographic retention of polybrominated diphenyl ethers. *J. Chromatogr. A* **2013**, *1298*, 118–131. [CrossRef]
22. D'Archivio, A.A.; Incani, A.; Ruggieri, F. Cross-column prediction of gas-chromatographic retention of polychlorinated biphenyls by artificial neural networks. *J. Chromatogr. A* **2011**, *1218*, 8679–8690. [CrossRef] [PubMed]
23. D'Archivio, A.A.; Giannitto, A.; Maggi, M.A.; Ruggieri, F. Cross-column retention prediction in reversed-phase high-performance liquid chromatography by artificial neural network modelling. *Anal. Chim. Acta* **2012**, *717*, 52–60. [CrossRef] [PubMed]
24. Fatemi, M.H.; Abraham, M.H.; Poole, C.F. Combination of artificial neural network technique and linear free energy relationship parameters in the prediction of gradient retention times in liquid chromatography. *J. Chromatogr. A* **2008**, *1190*, 241–252. [CrossRef] [PubMed]
25. Golubović, J.; Protić, A.; Otašević, B.; Zečević, M. Quantitative structure-retention relationships applied to development of liquid chromatography gradient-elution method for the separation of sartans. *Talanta* **2016**, *150*, 190–197. [CrossRef] [PubMed]

26. Barron, L.P.; McEneff, G.L. Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods. *Talanta* **2016**, *147*, 261–270. [CrossRef] [PubMed]
27. D'Archivio, A.A.; Maggi, M.A.; Ruggieri, F. Prediction of the retention of s-triazines in reversed-phase high-performance liquid chromatography under linear gradient-elution conditions. *J. Sep. Sci.* **2014**, *37*, 1930–1936. [CrossRef] [PubMed]
28. D'Archivio, A.A.; Maggi, M.A.; Ruggieri, F. Artificial neural network prediction of multilinear gradient retention in reversed-phase HPLC: Comprehensive QSRR-based models combining categorical or structural solute descriptors and gradient profile parameters. *Anal. Bioanal. Chem.* **2015**, *407*, 1181–1190. [CrossRef]
29. Todeschini, R.; Ballabio, D.; Grisoni, F. Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *J. Chem. Inf. Model.* **2016**, *56*, 1905–1913. [CrossRef]
30. Marini, F.; Bucci, R.; Magrì, A.L.; Magrì, A.D. Artificial neural networks in chemometrics: History, examples and perspectives. *Microchem. J.* **2008**, *88*, 178–185. [CrossRef]
31. Svozil, D.; Kvasnička, V.; Pospíchal, J. Introduction to multi-layer feed-forward neural networks. *Chemometr. Intell. Lab. Syst.* **1997**, *39*, 43–62. [CrossRef]
32. Lopez, R. Open NN: An Open Source Neural Networks C++ Library. 2014. Available online: http://opennn.cimne.com/ (accessed on 20 January 2019).

molecules

MDPI

Article

Quality Evaluation of *Gastrodia Elata* Tubers Based on HPLC Fingerprint Analyses and Quantitative Analysis of Multi-Components by Single Marker

Yehong Li [2,†], Yiming Zhang [1,†], Zejun Zhang [1], Yupiao Hu [1], Xiuming Cui [1,3,4] and Yin Xiong [1,3,4,*]

[1] Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming 650500, China; jr93586@163.com (Y.Z.); 18380802826@126.com (Z.Z.); hypflygo@163.com (Y.H.); cuisanqi37@163.com (X.C.)

[2] School of Pharmacy, China Pharmaceutical University, Nanjing 210009, China; yhlcpu@163.com

[3] Yunnan Key Laboratory of *Panax notoginseng*, Kunming University of Science and Technology, Kunming 650500, China

[4] Institute of Biology Leiden, Leiden University, 2333BE Leiden, The Netherlands

* Correspondence: yhsiung@163.com; Tel.: +86-0871-5915-818

† These authors contributed equally to this work.

Academic Editors: Marcello Locatelli, Angela Tartaglia, Dora Melucci, Abuzar Kabir, Halil Ibrahim Ulusoy and Victoria Samanidou
Received: 24 February 2019; Accepted: 15 April 2019; Published: 17 April 2019

Abstract: *Gastrodia elata* (*G. elata*) tuber is a valuable herbal medicine used to treat many diseases. The procedure of establishing a reasonable and feasible quality assessment method for *G. elata* tuber is important to ensure its clinical safety and efficacy. In this research, an effective and comprehensive evaluation method for assessing the quality of *G. elata* has been developed, based on the analysis of high performance liquid chromatography (HPLC) fingerprint, combined with the quantitative analysis of multi-components by single marker (QAMS) method. The contents of the seven components, including gastrodin, *p*-hydroxybenzyl alcohol, *p*-hydroxy benzaldehyde, parishin A, parishin B, parishin C, and parishin E were determined, simultaneously, using gastrodin as the reference standard. The results demonstrated that there was no significant difference between the QAMS method and the traditional external standard method (ESM) ($p > 0.05$, RSD < 4.79%), suggesting that QAMS was a reliable and convenient method for the content determination of multiple components, especially when there is a shortage of reference substances. In conclusion, this strategy could be beneficial for simplifying the processes in the quality control of *G. elata* tuber and giving references to promote the quality standards of herbal medicines.

Keywords: *Gastrodia elata* tuber; quality evaluation; HPLC; QAMS

1. Introduction

Gastrodia elata (G. elata) Blume is a traditional medicinal herb that has been used in oriental countries, for centuries, to treat general paralysis, headaches, dizziness, rheumatism, convulsion, and epilepsy [1,2]. Modern pharmacological studies have demonstrated that the extracts of *G. elata* tuber and some compounds that originate from it, possesses wide-reaching biological activities, including anti-tumor, anti-virus, memory-improving, anti-oxidation, and anti-aging actions [3–5]. Nowadays, it is also widely used as a sub-material in food and Chinese Patent Medicines (CPM) [6], and this herbal medicine is also listed as one of the functional foods approved by the Ministry of Health in China [7,8]. As the wild *G. elata* is not sufficient enough for commercial large-scale exploitation, its artificial cultivation in medicine has become essential, to meet the increasing requirement of markers [6].

Due to their high medicinal value, *G. elata* tubers have been cultivated and produced in many areas of Asia, like China and Korea, which could lead to great differences in quality and, possibly, could lead to differences in the following clinical efficacies. Many studies have indicated that the efficacy and quality of herbal medicines are somewhat different depending on the cultivation soil and climate, based on the geographic origin, even when coming from the same species [9,10]. Therefore, a reasonable and effective method for the quality evaluation of *G. elata* tuber, plays an important role in its medication safety.

Gastrodin and its aglycone (*p*-hydroxybenzyl alcohol) are major components of the *G. elata* tuber, which are also markers for the quality control of this herbal medicine [11]. However, over 81 compounds from *G. elata* tuber have been currently isolated and identified. Along with the above two marker components, others like *p*-hydroxy benzaldehyde, parishin A, parishin B, parishin C, parishin E, and so on have also been reported to be correlated with the bioeffects of the *G. elata* tuber [12,13]. Accordingly, a qualitative analysis and quantification of one or two compounds, could be insufficient for a complete profile of the chemical characterization of the *G. elata* tuber, due to its complex compositions. In recent years, the chromatographic fingerprint analysis has been accepted as a strategy for the quality assessment of herbal medicines and preparations by the US Food and Drug Administration [14], State Food and Drug Administration of China [15], and the European Medicines Agency [16]. Since the fingerprint is characterized by more chemical information, the method is often used for the origin identification, species authentication, and quality control for herbal medicines, by observing the presence or absence of a limited number of peaks in the chromatographic fingerprints [17,18]. Therefore, the fingerprint analysis of high performance liquid chromatography (HPLC) was developed for the qualitative analysis of *G. elata* tuber.

A single standard to determine multiple components, also known as the quantitative analysis of multi-components by single marker (QAMS) [19], is a novel method designed for the quality evaluation of herbal medicines and related products [20]. Researchers have used QAMS to determine three components in Fructus Evodiae, simultaneously, by using rutaecarpine as the internal reference compound to calculate the relative correction factor of evodin and evodiamine [21]. To make up for the limitations of the fingerprint which cannot be quantified accurately, a QAMS method using berberine as the standard, was developed and validated for a simultaneous quantitative analysis of fourteen components [22]. This strategy could not only reduce the cost of the experiment and time of detection but could also be independent of the availability of all target ingredients [19]. Thus, the QAMS method was applied for a quantitative analysis of *G. elata* tuber.

This study aimed to establish a reliable and practical method, realizing both qualitative and quantitative analyses for *G. elata* tuber, via HPLC fingerprinting, combined with QAMS. The differences and similarities of the HPLC fingerprints were visually compared, using a hierarchical cluster analysis (HCA) and similarity analysis. The contents of seven major active constituents were accurately determined by both the QAMS method and external standard method (ESM), through which we hoped to offer a suitable and efficient approach for assessing the quality of *G. elata* tuber.

2. Results and Discussion

2.1. Optimization of the Chromatographic Conditions

As the components of *G. elata* tuber are very intricate, it is critical to optimize the chromatographic conditions, including favorable mobile phase systems, gradient elution systems, and the detection wavelength, to obtain an efficient separation of the target components. Lei [23] indicated that the HPLC fingerprints of *G. elata* tubers were the most informative, while the UV wavelength was 220 nm from HPLC-DAD-3D spectrum of *G. elata* tuber. So in this case, we chose the UV wavelength of 220 nm, to determinate the selected components. We chose acetonitrile-water containing 0.1% phosphoric acid system. The samples were dissolved in 60% methanol and ultrasound, for 60 min. We optimized the gradient elution system as Section 3.5, and 35 °C was selected as the proper temperature for analysis, while the flow rate was set at 1.0 mL/min. The S1 sample of *G. elata* tuber and the mixed standards containing seven reference substances were analyzed to obtain the HPLC fingerprints (Figure 1) under the conditions of Section 3.5, producing sharp and symmetrical chromatographic peak shapes, good separation, and preventing the peak tailing.

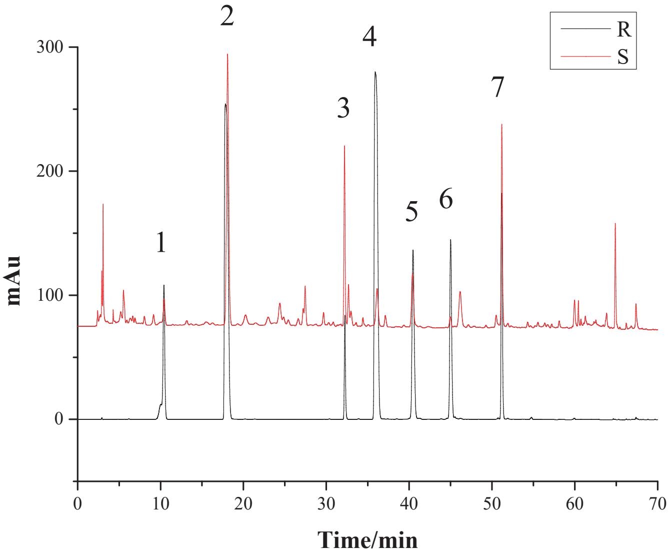


Figure 1. The HPLC fingerprints of the *Gastrodia elata* tuber sample and the mixed standards. R: The mixed standards; S: The *G. elata* tuber sample. 1—Gastrodin; 2—*p*-Hydroxy benzyl alcohol; 3—Parishin E; 4—*p*-Hydroxy benzaldehyde; 5—Parishin B; 6—Parishin C; 7—Parishin A.

According to the retention time of each peak in the chromatogram [24], the peaks of 1, 2, 3, 4, 5, 6, and 7 were identified to be gastrodin, *p*-hydroxybenzyl alcohol, parishin E, *p*-hydroxy benzaldehyde, parishin B, parishin C, and parishin A. The separation degree of each peak was greater than 1.5, in the present HPLC system, indicating the peaks were well-separated, under the chromatographic conditions.

2.2. Method Validation

2.2.1. Linearity

The mixed reference solution containing all the reference substances was diluted in series, with 60% methanol, to obtain six different concentrations for the seven reference curves. The linearity of each analyte was assessed by plotting its calibration curve with different concentrations and the corresponding peak areas. The results were shown in Table 1. The high correlation coefficient values indicated that there was a good correlation between the concentration and peak area of the seven compounds, at a relatively wide range of concentrations. The correlation coefficient of more than 0.9990, indicated a satisfactory linearity. The calibration curve could be utilized for the quantitative analysis in the given concentration range. The standard solution of the individual analyte was diluted gradually, to determine its Limit of Detection (LOD) and Limit of Quantity (LOQ) with signal-to-noise ratio of 3:1 and 10:1, respectively. LOD and LOQ values for the analytes are also listed in Table 1.

Table 1. The regression equations, Limit of Detection (LODs) and Limit of Quantity (LOQs) of seven components.

Analytes	Regression Equations	Linear Ranges (mg/mL)	R^2	LOD (mg/mL)	LOQ (mg/mL)
Gastrodin	$Y = 18634X - 264.07$	1.906~6.483	0.9997	0.042	0.139
p-Hydroxybenzyl alcohol	$Y = 39300X + 42.955$	0.075~1.773	0.9995	0.001	0.003
Parishin E	$Y = 14141X + 142.93$	2.273~7.052	0.9997	0.037	0.122
p-Hydroxy benzaldehyde	$Y = 52536X + 7.9174$	0.079~2.588	1.0000	0.001	0.005
Parishin B	$Y = 20791X + 6.7746$	1.450~5.190	1.0000	0.004	0.015
Parishin C	$Y = 31240X - 335.24$	0.286~0.356	0.9997	0.005	0.015
Parishin A	$Y = 11769X - 100.83$	0.181~19.301	0.9995	0.020	0.070

2.2.2. Precision, Stability, Repeatability, and Accuracy

The precision was evaluated according to the assay of S1, in which the solution was analyzed for six times in a day, to evaluate the intra-day precision, and was analyzed on three consecutive days, to evaluate the inter-day precision. Calculating the RSDs of each chromatographic peak, the results showed that the RSDs of gastrodin, *p*-hydroxybenzyl alcohol, parishin E, *p*-hydroxy benzaldehyde, parishin B, parishin C, and parishin A were 1.93%, 1.10%, 1.29%, 2.30%, 2.03%, 2.63%, and 0.89% (n = 6), respectively, indicating that the precision of the method was good.

The stability was tested with the S1 solution that was stored at room temperature (25 ± 5 °C) and analyzed at 0, 2, 4, 6, 8, 12, and 24 h, to calculate the RSDs. The results showed that the RSDs of gastrodin, *p*-hydroxybenzyl alcohol, parishin E, *p*-hydroxy benzaldehyde, parishin B, parishin C, and parishin A were 1.15%, 2.04%, 1.51%, 2.37%, 2.10%, 1.12%, and 2.25%, respectively, suggesting that the method was stable within 24 h.

In the repeatability test, six duplicates of S1 were extracted and analyzed, according to the sample preparation procedure, and the HPLC method. The RSDs of the peak areas were calculated. The results showed that the RSDs of gastrodin, *p*-hydroxybenzyl alcohol, parishin E, *p*-hydroxy benzaldehyde, parishin B, parishin C, and parishin A were 1.25%, 2.15%, 1.60%, 1.81%, 1.72%, 1.84%, and 1.60% (n = 6), respectively, indicating that the repeatability of the method was good.

In the accuracy test, certain amounts of the seven analytes' standards were added to the *G. elata* tuber samples (S1), with the six replicates. Then, these seven mixed samples were treated, as in the method described above. Recovery rate was used as the evaluation index and calculated as Recovery rate (%) = (Found amount − Known amount) × 100%/Added amount. The RSD of the accuracy values of the seven components are shown in Table 2, respectively.

Table 2. RSD of precision, stability, repeatability and accuracy for determination of seven components.

Analyte RSD (%)	Precision RSD (%)	Stability RSD (%)	Repeatability RSD (%)	Accuracy	
				Mean (%)	RSD (%)
Gastrodin	1.93	1.15	1.25	92.05%	2.02%
p-Hydroxybenzyl alcohol	1.10	2.04	2.15	95.78%	1.09%
Parishin E	1.29	1.51	1.60	98.05%	2.90%
p-Hydroxy benzaldehyde	2.30	2.37	1.81	92.44%	0.25%
Parishin B	2.03	2.10	1.72	93.33%	1.32%
Parishin C	2.63	1.12	1.84	92.91%	2.10%
Parishin A	0.89	2.25	1.60	91.80%	1.36%

The HPLC method was validated in terms of precision, repeatability, stability, and accuracy, as shown in Table 2. The RSD of the precision values of the seven components were less than 2.63%. RSD values for the stability and the repeatability were less than 2.37% and 2.15%, respectively. The recovery rates of the analytes ranged from 91.80% to 98.05%, with the RSD values being lower than 2.90%. All results indicated that the developed method was stable, accurate, and repeatable. This established

HPLC method could be applied for a simultaneous determination of gastrodin, *p*-hydroxybenzyl alcohol, parishin E, *p*-hydroxy benzaldehyde, parishin B, parishin C, and parishin A, in the *G. elata* tuber samples.

2.3. HPLC Fingerprints Analysis

The 21 batches of *G. elata* tuber samples from the different producing areas were prepared according to Section 3.3, and 10 μL of S1 sample solution was injected into the HPLC system according to the chromatographic conditions in Section 3.5, to obtain the fingerprints. The retention time was the horizontal axis and the peak area was the vertical axis; the 3D fingerprints of the 21 batches of *G. elata* tuber samples were established by the software Origin 9.0, as shown in Figure 2.

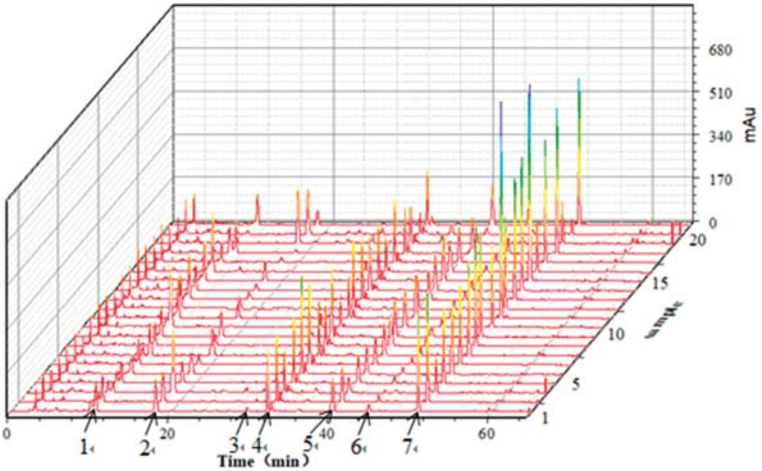


Figure 2. HPLC fingerprints of the 21 batches of *G. elata* tuber samples. 1—Gastrodin; 2—*p*-Hydroxy benzyl alcohol; 3—Parishin E; 4—*p*-Hydroxy benzaldehyde; 5—Parishin B; 6—Parishin C; 7—Parishin A.

According to Figure 2, the seven peaks with stable and better shape were determined to be the major ones for the HPLC fingerprints of *G. elata* tubers. The peak areas of the seven peaks are shown in Table 3. The variance coefficients of the peak area were greater than 32.2 percent, indicating that the content of each marker component varied greatly from place to place.

Table 3. The information and peak areas of the seven characteristic peaks in HPLC fingerprints of *G. elata* tubers.

No.	Peak Area of Seven Characteristic Peaks						
	Gastrodin	*p*-Hydroxy Benzyl Alcohol	Parishin E	*p*-Hydroxy Benzaldehyde	Parishin B	Parishin C	Parishin A
S1	1797.1	2249.5	2337.8	217.4	2263.6	420.7	4340.3
S2	1470.2	2144.3	2523.4	227.6	2526.6	462.2	4561.9
S3	623.8	4536.4	1528.4	301.7	1017	280.3	1487.8
S4	1325.1	1516.1	1412.1	116.7	1906.9	402.8	3150
S5	1659.3	2123.3	1991.3	111.9	2141.7	383.4	4006.4
S6	1161	1463.7	3734.3	108.1	1867.1	390.2	3167.6
S7	1492.8	663.8	2991.6	85.8	1818.7	392.5	3473.5
S8	1470.9	823.8	1573.2	127.1	2231.8	546.3	5104.3
S9	1898	1876.6	2572.7	82	2629.1	595.9	4430.8
S10	3816.6	136.2	1316.2	110.1	2663.3	441.9	3383.6
S11	2353.9	970.8	1563.7	131.9	3073.4	789.5	9224.6
S12	1794	830	2577.2	45.7	2141.8	101.1	3845.8
S13	2344.5	572.4	2363.1	57.6	2039.1	499.2	5019.1
S14	1369.4	427.8	1961.6	41.9	2408.9	622.1	5512.4
S15	2177.5	1270.2	2076.6	56.6	3133.4	791.2	8184.9
S16	3322.1	108.1	1240.9	73.1	1935.1	357.8	2127.8
S17	1081.8	322.8	2365	104.7	2363.6	500.7	5062.6
S18	1893.7	270.9	1719.4	78.3	2475.6	823	6072.7
S19	380.4	4012.7	1414.3	617.3	781.5	136.4	1789.1
S20	300.9	3287.7	878.9	564	479.5	102.3	687.1
S21	2175.1	1076.8	2057.2	143.7	2826.4	94.7	6278.5
C.V. (%) [1]	49.7	85.2	33.3	96.5	32.2	50.1	47.9

[1] C.V. (%) = $\delta/\mu \times 100$, δ—The standard deviation of peak area and μ—The average value of each peak area.

2.4. Similarity Analysis

According to the data of HPLC fingerprints in Figure 2, the similarity of HPLC fingerprints from the different producing regions were evaluated using the Similarity Evaluation System for chromatographic fingerprint of traditional Chinese medicines (TCM) (Version 2012), with correlation coefficient (median) on behalf of the similarity of HPLC fingerprints. We utilized the average correlation coefficient method of 21 batches of the samples for the multipoint correction, and the time window width was set to 0.5 [25], while the establishment of a common model was to generate a control fingerprints of the *G. elata* tuber. Compared with the reference fingerprint chromatogram (R), the similarities of the 21 batches of samples were higher than 0.96, indicating that the batch-to-batch consistency was good. The results suggested that those samples of *G. elata* tuber had a similar chemical composition, and the samples were collected from the same genus, even though they were from different producing countries or were produced under different processing conditions (Table 4). Therefore, the developed fingerprint by HPLC could be used as a practical tool for the qualitative identification of the *G. elata* tuber.

Table 4. Similarity of the *G. elata* tuber samples.

No.	Similarity	No.	Similarity	No.	Similarity	No.	Similarity
S1	0.983	S7	0.970	S13	0.988	S19	0.990
S2	0.987	S8	0.988	S14	0.982	S20	0.988
S3	0.983	S9	0.989	S15	0.979	S21	0.988
S4	0.975	S10	0.982	S16	0.989	*R*	1.000
S5	0.983	S11	0.987	S17	0.980		
S6	0.975	S12	0.990	S18	0.964		

2.5. Hierarchical Cluster Analysis (HCA)

Using the peak areas of the seven compounds from the 21 *G. elata* tuber samples as the clustering variable, the HCA of the standardized data was performed with the heat map software of Heml 1.0. The graph in Figure 3 illustrated that the samples could be categorized into three groups. Group 1 contained S1 and S2 from Zhaotong, Yunnan in China; Group 2 contained S19 and S20 tubers from South Korea; and Group 3 contained the rest of samples. From the result, the samples from the same producing area were not always classified into the same group. For example, Zhaotong has been considered as the Daodi production area (area which produces authentic and superior medicinal materials) of the *G. elata* tuber in China. However, samples 1 to 6 from Zhaotong, showed different levels and ratios of chemical components, which could be due to the variations in harvesting time, planting patterns, dying methods, and other factors. Additionally, the preliminary processing method also contributes to the differences in the chemical composition. For instance, *G. elata* tubers and slices from South Korea were classified into different categories. Therefore, it is insufficient to determine the quality of the *G. elata* tubers by only their producing areas or any other single factor. Although the HCA could be used to classify the *G. elata* tubers on the basis of the peak areas of the seven components, it was hard to tell which group had a better quality. Therefore, other methods for the quantitative analysis of *G. elata* tubers should be developed, to reflect the quality difference.

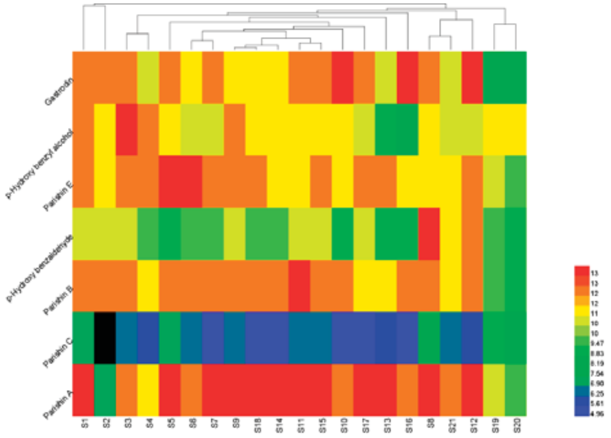


Figure 3. Clustering analysis graph of the 21 *G. elata* tuber samples.

2.6. Quantitative Analysis of Multiple Components by Single Marker

Theoretically, the quantity (mass or concentration) of an analyte is in direct proportion of the detector response. Then, in multi-component quantitation, a typical botanical compound (readily available) might be selected as an internal standard and the relative correction factor (RCF) of this marker, and the other components can be calculated.

2.6.1. Calculation of RCFs

It is of vital importance to select a proper internal referring standard for the accurate assay of multiple components in TCM. The component chosen as the internal referring substance should be stable, easily obtainable, and have relatively clear pharmacologic effects related to the clinical efficacy of the herbal medicine [26]. In this work, the gastrodin was used as an internal referring substance for its easy availability, lower cost, moderate retention value, and good stability.

In order to simultaneously determine the contents of the seven components in the *G. elata* tuber, by using the QAMS method, the relative correction factors (RCFs, f_x) were first determined, according to the ratio of the peak areas and the ratio of the concentration between the gastrodin and other compounds, as described in Section 3.6. We calculated the RCFs of six components (shown in Table 5).

Table 5. Relative correction factor (RCF) values of six components of the *G. elata* tuber.

Instrument	Chromatogram Column	RCF Values	
Agilent 1260	YMC-Tyiart C18 (250 × 4.6 mm, 5 µm)	$f_{\text{P-hydroxy benzyl alcohol/gastrodin}}$	2.1090
		$f_{\text{parishin E/gastrodin}}$	0.7589
		$f_{\text{P-hydroxy benzaldehyde/gastrodin}}$	2.8194
		$f_{\text{parishin B/gastrodin}}$	1.1156
		$f_{\text{parishin C/gastrodin}}$	1.6771
		$f_{\text{parishin A/gastrodin}}$	0.6316

2.6.2. Results from the QAMS Method

After preparing the sample solutions of *G. elata* tubers, they were injected into the HPLC system to obtain the peak areas. The contents of seven compounds were calculated, according to the calibration curves. Those scattered in the vicinity of the lowest concentration point on the standard curve were determined with a one point ESM. Meanwhile, the contents of the seven components of the *G. elata* tuber calculated according to QAMS method, are shown in Table 6.

The validated traditional ESM and QAMS method were employed to test the 21 batches of *G. elata* tuber samples from the different producing areas, which were based on the principle of the linear relationship between a detector response and the levels of components within certain concentration ranges. The validation of the QAMS method might be implemented, based on *t*-test, correlation coefficient [27], RSD [28], and relative error [29], through a comparison with an external standard. Correlation coefficient, as a statistical parameter, ranging from 0 (no correlation) to 1 (complete correlation), reflecting the closeness of two variables, is often used in similarity assessments of traditional Chinese medicine fingerprints [30]. As shown in Table 7, Correlation coefficients of the assay results obtained from the two methods were calculated here; all coefficients were found to be >0.998. The data showed that the results of the two methods were highly correlated. Then, a *t*-test was performed for the calculated results, by the QAMS method, and the on detected results, by an external standard method. *p*-values of gastrodin, *p*-hydroxy benzyl alcohol, parishin E, *p*-hydroxy benzaldehyde, parishin B, parishin C and parishin A, were all >0.05. The relative error and RSD values were all lower than 5%. Above all, the results indicated that there was no significant difference between the data from the QAMS and the ESM method, indicating that the present QAMS method was reliable for the simultaneous quantification of the seven components of the *G. elata* tuber.

Table 6. Contents of the seven components in *G. elata* tubes determined by the external standard method (ESM) and the quantitative analysis of multi-components by single marker (QAMS) methods ($mg·g^{-1}$) [1].

No.	Gastrodin	*p*-Hydroxy Benzyl Alcohol		Parishin E		*p*-Hydroxy Benzaldehyde		Parishin B		Parishin C		Parishin A		Total
		ESM	QAMS	ESM	QAMS	ESM	QAMS	ESM	QAMS	ESM	QAMS	ESM	QAMS	
S1	5.23 ± 0.16	1.77 ± 0.05	1.82 ± 0.05	5.14 ± 0.01	5.35 ± 0.03	0.24 ± 0.01	0.25 ± 0.06	3.54 ± 0.14	3.60 ± 0.05	0.18 ± 0.00	0.19 ± 0.01	11.58 ± 0.45	11.71 ± 0.49	27.68
S2	4.51 ± 0.38	1.61 ± 0.08	1.64 ± 0.06	2.27 ± 0.14	2.38 ± 0.08	0.24 ± 0.01	0.25 ± 0.01	3.43 ± 0.21	3.47 ± 0.01	0.13 ± 0.00	0.14 ± 0.00	0.18 ± 0.00	0.00 ± 0.00	12.38
S3	2.44 ± 0.28	1.07 ± 0.08	1.09 ± 0.10	4.82 ± 0.25	5.00 ± 0.04	0.26 ± 0.01	0.26 ± 0.02	2.84 ± 0.10	2.88 ± 0.04	0.16 ± 0.02	0.16 ± 0.02	8.33 ± 0.31	8.32 ± 0.11	19.91
S4	1.35 ± 0.02	3.36 ± 0.12	3.39 ± 0.08	2.62 ± 0.06	2.62 ± 0.23	0.23 ± 0.01	0.24 ± 0.04	1.45 ± 0.04	1.46 ± 0.09	0.16 ± 0.00	0.16 ± 0.10	4.23 ± 0.12	4.10 ± 0.31	13.41
S5	3.41 ± 0.38	1.55 ± 0.10	1.58 ± 0.07	2.31 ± 0.22	2.36 ± 0.10	0.12 ± 0.02	0.13 ± 0.01	2.49 ± 0.12	2.51 ± 0.11	0.20 ± 0.00	0.20 ± 0.01	8.94 ± 0.63	8.92 ± 0.29	19.03
S6	1.91 ± 0.12	1.03 ± 0.07	1.06 ± 0.10	7.05 ± 0.13	7.26 ± 0.23	0.23 ± 0.02	0.24 ± 0.05	2.69 ± 0.03	2.73 ± 0.17	0.17 ± 0.00	0.17 ± 0.02	7.64 ± 0.31	7.63 ± 0.61	20.72
S7	3.12 ± 0.01	0.51 ± 0.00	0.531 ± 0.00	6.02 ± 0.14	6.20 ± 0.09	0.21 ± 0.01	0.21 ± 0.00	2.71 ± 0.02	2.73 ± 0.01	0.15 ± 0.00	0.16 ± 0.00	8.87 ± 0.02	8.86 ± 0.08	21.58
S8	3.06 ± 0.10	0.62 ± 0.01	0.64 ± 0.01	3.00 ± 0.18	3.13 ± 0.17	2.59 ± 0.02	2.60 ± 0.05	3.25 ± 0.15	3.27 ± 0.01	0.15 ± 0.01	0.14 ± 0.04	12.77 ± 0.58	12.75 ± 0.54	25.44
S9	2.85 ± 0.37	1.22 ± 0.18	1.24 ± 0.20	4.54 ± 0.03	4.69 ± 0.06	0.28 ± 0.01	0.28 ± 0.01	3.61 ± 0.06	3.63 ± 0.15	0.17 ± 0.01	0.16 ± 0.05	10.78 ± 0.15	10.77 ± 0.08	23.44
S10	5.89 ± 0.22	0.10 ± 0.01	0.10 ± 0.01	3.37 ± 0.24	3.52 ± 0.23	0.11 ± 0.01	0.11 ± 0.06	3.84 ± 0.13	3.91 ± 0.05	0.15 ± 0.02	0.16 ± 0.00	7.90 ± 0.67	7.91 ± 0.63	21.36
S11	4.74 ± 0.37	0.69 ± 0.08	0.71 ± 0.08	3.40 ± 0.22	3.55 ± 0.22	0.26 ± 0.02	0.26 ± 0.04	5.191 ± 0.09	5.23 ± 0.02	0.18 ± 0.01	0.19 ± 0.00	26.70 ± 0.46	26.93 ± 0.54	41.15
S12	7.10 ± 0.27	0.65 ± 0.04	0.66 ± 0.01	4.88 ± 0.23	5.04 ± 0.11	0.08 ± 0.12	0.08 ± 0.02	3.03 ± 0.16	3.18 ± 0.01	0.15 ± 0.02	0.15 ± 0.00	9.43 ± 0.54	9.80 ± 0.10	25.32
S13	4.03 ± 0.03	0.37 ± 0.01	0.39 ± 0.01	3.69 ± 0.11	3.86 ± 0.11	0.16 ± 0.01	0.17 ± 0.02	2.37 ± 0.06	2.41 ± 0.06	0.15 ± 0.00	0.15 ± 0.00	10.27 ± 0.24	10.33 ± 0.24	21.04
S14	2.59 ± 0.03	0.19 ± 0.16	0.20 ± 0.00	2.97 ± 0.08	3.10 ± 0.07	0.20 ± 0.00	0.21 ± 0.02	2.88 ± 0.06	2.90 ± 0.05	0.15 ± 0.01	0.15 ± 0.00	11.61 ± 0.37	11.58 ± 0.33	20.59
S15	4.29 ± 0.15	0.92 ± 0.04	0.94 ± 0.04	3.76 ± 0.16	3.90 ± 0.16	0.31 ± 0.02	0.31 ± 0.07	4.27 ± 0.17	4.30 ± 0.10	0.13 ± 0.00	0.14 ± 0.00	19.30 ± 0.83	19.42 ± 0.84	32.98
S16	6.48 ± 0.21	0.08 ± 0.00	0.08 ± 0.00	2.35 ± 0.11	2.41 ± 0.13	0.11 ± 0.00	0.11 ± 0.05	2.75 ± 0.11	2.78 ± 0.04	0.15 ± 0.01	0.15 ± 0.01	5.24 ± 0.15	5.16 ± 0.20	17.15
S17	5.09 ± 0.39	0.13 ± 0.22	0.13 ± 0.26	4.20 ± 0.05	4.39 ± 0.05	0.24 ± 0.00	0.25 ± 0.07	1.54 ± 0.17	1.56 ± 0.08	0.15 ± 0.01	0.14 ± 0.01	11.65 ± 0.22	11.77 ± 0.23	23.01
S18	3.71 ± 0.05	0.20 ± 0.01	0.20 ± 0.01	4.22 ± 0.09	4.41 ± 0.13	0.20 ± 0.00	0.20 ± 0.04	3.47 ± 0.08	3.52 ± 0.03	0.15 ± 0.00	0.16 ± 0.01	14.60 ± 0.21	14.76 ± 0.24	26.54
S19	0.42 ± 0.01	1.23 ± 0.02	1.26 ± 0.02	1.10 ± 0.01	1.14 ± 0.01	0.18 ± 0.01	0.18 ± 0.00	0.46 ± 0.01	0.47 ± 0.01	0.38 ± 0.01	0.39 ± 0.04	1.99 ± 0.06	1.98 ± 0.05	5.76
S20	0.36 ± 0.02	1.01 ± 0.01	1.04 ± 0.01	0.64 ± 0.00	0.65 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.28 ± 0.00	0.29 ± 0.00	0.30 ± 0.01	0.30 ± 0.03	0.83 ± 0.09	0.82 ± 0.00	3.52
S21	1.50 ± 0.14	0.33 ± 0.00	0.34 ± 0.00	1.70 ± 0.40	1.71 ± 0.04	0.68 ± 0.02	0.69 ± 0.01	1.75 ± 0.02	1.78 ± 0.01	0.17 ± 0.00	0.16 ± 0.01	6.62 ± 0.03	6.61 ± 0.03	12.75
Mean	3.53		0.91		3.65		0.34		2.79		0.18		9.53	20.70

[1] ESM—external standard method, and its content was determined by the calibration equation method; QAMS—quantitative analysis multi-components by single marker, and its content was determined by RCFs; RSD—relative standard deviation; Total—the sum of the six alkaloid contents in each batch.

Table 7. The relative error, RSD, correlation coefficient, and p values of the contents from the ESM and the QAMS [1].

No.	*p*-Hydroxy Benzyl Alcohol		Parishin E		*p*-Hydroxy Benzaldehyde		Parishin B		Parishin C		Parishin A	
	Relative Error	RSD	Relative Error	RSD	Relative Error	RSD	Relative Error	RSD	Relative Error	RSD	Relative Error	RSD
S1	2.38%	1.70%	4.04%	2.92%	2.39%	1.71%	1.76%	1.25%	1.47%	1.05%	1.15%	0.82%
S2	1.78%	1.27%	4.56%	3.30%	1.72%	1.23%	1.10%	0.78%	1.55%	1.11%	0.00%	0.00%
S3	2.38%	1.70%	3.72%	2.68%	1.94%	1.39%	1.37%	0.97%	1.56%	1.11%	0.02%	0.10%
S4	0.72%	0.51%	0.10%	0.07%	1.11%	0.79%	0.60%	0.43%	0.94%	0.67%	0.18%	2.13%
S5	1.52%	1.08%	2.13%	1.52%	2.14%	1.53%	0.90%	0.64%	1.11%	0.79%	0.03%	0.18%
S6	2.50%	1.79%	2.95%	2.12%	2.01%	1.44%	1.38%	0.98%	0.66%	0.47%	0.01%	0.07%
S7	3.39%	2.43%	2.84%	2.03%	1.70%	1.21%	0.97%	0.69%	3.09%	2.22%	0.02%	0.12%
S8	2.38%	1.70%	4.27%	3.08%	0.32%	0.23%	0.37%	0.26%	3.81%	2.74%	0.02%	0.09%
S9	1.65%	1.17%	3.18%	2.29%	1.17%	0.83%	0.60%	0.43%	0.74%	0.52%	0.02%	0.10%
S10	1.29%	0.92%	4.24%	3.06%	2.81%	2.01%	1.59%	1.13%	3.05%	2.19%	0.21%	0.15%
S11	2.57%	1.84%	4.20%	3.03%	1.38%	0.98%	0.74%	0.53%	4.84%	3.51%	0.88%	0.62%
S12	0.77%	0.54%	3.22%	2.32%	3.59%	2.59%	4.72%	3.42%	0.31%	0.22%	3.81%	2.75%
S13	4.77%	3.45%	4.54%	3.29%	2.41%	1.73%	1.48%	1.05%	0.90%	0.64%	0.61%	0.43%
S14	4.79%	3.47%	4.31%	3.12%	1.15%	0.82%	0.39%	0.27%	0.66%	0.47%	0.27%	0.19%
S15	1.95%	1.39%	3.72%	2.68%	1.10%	0.78%	0.59%	0.42%	4.61%	3.34%	0.61%	0.44%
S16	4.71%	3.41%	2.10%	1.50%	2.69%	1.93%	1.15%	0.82%	0.39%	0.28%	1.44%	1.02%
S17	0.76%	0.54%	4.31%	3.12%	2.23%	1.60%	1.61%	1.15%	3.77%	2.72%	0.99%	0.70%
S18	0.35%	0.25%	4.31%	3.11%	2.23%	1.59%	1.43%	1.02%	4.95%	3.59%	1.14%	0.81%
S19	2.54%	1.82%	3.76%	2.71%	2.50%	1.79%	2.33%	1.67%	0.94%	0.67%	0.38%	0.27%
S20	3.25%	2.33%	1.43%	1.02%	3.59%	2.59%	3.34%	2.40%	0.07%	0.05%	1.13%	0.80%
S21	2.53%	1.81%	0.70%	0.50%	1.75%	1.24%	1.70%	1.21%	2.10%	1.50%	0.20%	0.14%
Correlation coefficient	0.999 **		0.999 **		0.999 **		1.000 **		0.998 **		0.999 **	
***p* values**	0.940		0.802		0.978		0.923		0.960		0.986	

[1] RSD—relative standard deviation; p values—the paired t-test results; ESM—external standard method, and its content was determined by the calibration equation method; QAMS—quantitative analysis multi-components by single marker, and its content was determined by RCFs; ** $p < 0.01$.

The results from the QAMS determination of the 21 batches of *G. elata* tuber samples showed the mean contents of 3.5275 mg·g^{-1}, 0.9060 mg·g^{-1}, and 0.3398 mg·g^{-1} for gastrodin, *p*-hydroxy benzyl alcohol, and *p*-hydroxy benzaldehyde; and 3.6511 mg·g^{-1}, 9.5303 mg·g^{-1}, 2.7901 mg·g^{-1}, and 0.1766 mg·g^{-1} for the parishin E, parishin A, parishin B, and parishin C, respectively (Table 4). It was obvious that parishin A is one of the most abundant components in *G. elata* tuber, thus, is well-deserved as a reference substance and index for quality assessment and control of the *G. elata* tuber. Obvious inter-batch content variations could be found for all these components with the mean ranging from 0.1766 mg·g^{-1} to 9.5303 mg·g^{-1}; these seven components in total averaged 20.7031 mg·g^{-1} in the *G. elata* tuber, for the 21 batches of samples. The data in Table 4 shows differences among various samples. To show the clear classification of the *G. elata* tuber samples, the QAMS method with chemometrics analysis was performed in the subsequent analyses.

Meanwhile, the results (Table 6) illustrated that there were remarkable differences in the contents of the seven components, in *G. elata* tubers from different regions, which could be attributed to the variations of genetics, plant origins, environmental factors, drying process, storage conditions, and so on. It was obvious that gastrodin is one of the most abundant components in *G. elata* tuber. Combined with its activities related to the efficacies of *G. elata* tuber [31], gastrodin is well-deserved as a reference substance and index for quality assessment and control of *G. elata* tuber.

In the Chinese Pharmacopoeia of 2015 edition, gastrodin and *p*-hydroxy benzyl alcohol are determined as the marker components for the quality control and evaluation of *G. elata* tuber. Despite their close correlation with the efficacies of *G. elata* tuber, gastrodin can transform to *p*-hydroxybenzyl alcohol, which is the aglycone and metabolite of gastrodin [32]. Fresh *G. elata* tubers have to be processed before being traded as materia medica in the market. During the steaming process, the change trend of the gastrodin content was often contrary to the one of *p*-hydroxybenzyl alcohol. When the content of gastrodin was increased, the content of *p*-hydroxybenzyl alcohol was generally decreased, and vice versa. Additionally, different processing methods will result in different variation of the contents of the two components. Choi et al. [33] applied drying methods of freeze drying, hot air, infrared ray, and steaming, to process *G. elata* tuber. The results showed that after steaming, the content of gastrodin in *G. elata* tuber processed by freeze drying was decreased, whereas, the content of *p*-hydroxybenzyl alcohol was increased. However, tubers processed by hot-air and infrared ray drying showed the opposite results. Such transformations between gastrodin and *p*-hydroxybenzyl alcohol might be due to the deglycosylation or glycosylation, during the processing. Since the herbal medicine in the global market is often processed or dried by different methods, which results in the fluctuation in the content of single component, it is relatively stable and more comprehensive to reflect on the quality of *G. elata* tuber by monitoring multiple components, instead of a single one.

3. Materials and Methods

3.1. Plant Material

Samples of *G. elata* tuber from different producing areas were collected, as shown in Table 8.

Table 8. The information of *G. elata* tubers from different producing areas.

No.	Sample	Producing Areas	No.	Sample	Producing Areas
S1	*G. elata* tubers	Zhaotong, Yunnan, China	**S12**	*G. elata* tubers	Enshi, Hubei, China
S2	*G. elata* tubers	Zhaotong, Yunnan, China	**S13**	*G. elata* tubers	Yichang, Hubei, China
S3	*G. elata* tubers	Zhaotong, Yunnan, China	**S14**	*G. elata* tubers	Hanzhong, Shanxi, China
S4	*G. elata* tubers	Zhaotong, Yunnan, China	**S15**	*G. elata* tubers	Qinling, Shanxi, China
S5	*G. elata* tubers	Zhaotong, Yunnan, China	**S16**	*G. elata* tubers	Qinchuan, Sichuang, China
S6	*G. elata* tubers	Zhaotong, Yunnan, China	**S17**	*G. elata* tubers	Longnan, Gansu, China
S7	*G. elata* tubers	Lijiang, Yunnan, China	**S18**	*G. elata* tubers	Anhui, China
S8	*G. elata* tubers	Bijie, Guizhou, China	**S19**	*G. elata* tubers	Moju, South Korea
S9	*G. elata* tubers	Zhengyuan, Guizhou, China	**S20**	*G. elata* tubers	Chun chuan, South Korea
S10	*G. elata* tubers	Qiandongnan, Guizhou, China	**S21**	*G. elata* tuber slices	Yingyang, South Korea
S11	*G. elata* tubers	Bijie, Guizhou, China			

3.2. Chemicals

The reference standards of gastrodin (no. B21243, purity HPLC ≥ 98%), *p*-hydroxybenzyl alcohol (no. B20326, purity HPLC ≥ 98%), *p*-hydroxy benzaldehyde (no. B20327, purity HPLC ≥ 99%), parishin A (no. BP1063, purity HPLC ≥ 98%), parishin B (no. BP1064, purity HPLC ≥ 98%), parishin C (no. B20913, purity HPLC ≥ 98%), parishin E (no. BP1648, purity HPLC ≥ 98%) were purchased from Sichuan Victory Biological Technology Co., Ltd. (Sichuan, China), and their structures are shown in Figures 4 and 5. Methyl alcohol was purchased from the Tianjin Fengchuan Chemical Reagent Technology Co. Ltd. Acetonitrile (HPLC grade) was purchased from Sigma-Aldrich, Inc. (St. Louis, MO, USA). Phosphoric acid was purchased from the Tianjin JinDongTianZheng Precision Chemical Reagent Factory. Ultrapure water was generated with an UPT-I-20T ultrapure water system (Yunnan Ultrapure Technology, Inc., Yunnan, China). All other chemicals used were of analytical grade.

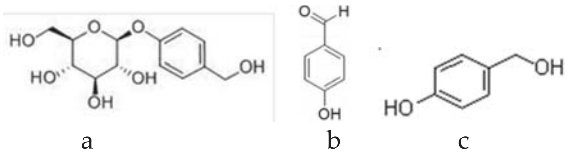


Figure 4. The structures of some compounds in the *G. elata* tuber. (**a**) Gastrodin [34], (**b**) *p*-hydroxy benzaldehyde [35], and (**c**) *p*-hydroxybenzyl alcohol [36].

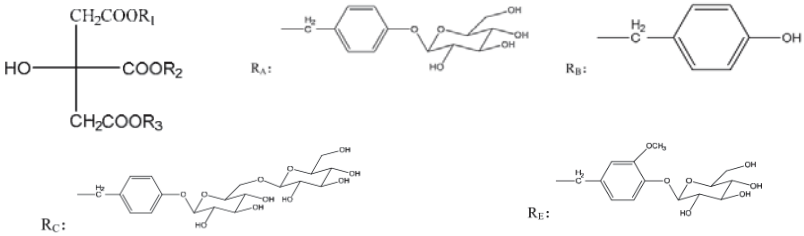


Figure 5. The structures of parishins in the *G. elata* tuber. The structure of parishins [13]: R_A. parishin A, R_B. parishin B, R_C. parishin C, R_E. parishin E.

3.3. Preparation of the Sample Solution

The 21 batches of dried *G. elata* tubers from different producing areas were crushed by a Wiggling high-speed Chinese medicine shredder, then powdered and sieved through a 40-mesh sieve. The sample solution of *G. elata* tuber was precisely absorbed (2.0 mg) and immersed in 25 mL volumetric flask, with 60% methanol. Additional 60% methanol was added to compensate for the weight loss after ultrasonic extraction for 60 min, and shaking it well. All solutions were filtered through 0.22 μm filter membranes, before being precisely injected into the HPLC system.

3.4. Reference Solution Preparation

The reference solution of *G. elata* tuber was prepared by accurately dissolving weighed samples of each compound in 60% methanol, making a mixture of 0.8 mg/mL of parishin A, 0.9 mg/mL of parishin B, 0.5 mg/mL of parishin E, 1.5 mg/mL of *p*-hydroxy benzaldehyde, 3.4 mg/mL of *p*-hydroxybenzyl alcohol, 0.9 mg/mL of gastrodin, 1.3 mg/mL of parishin C, mixed evenly. All the standard solutions were stored in a refrigerator at 4 °C, before use.

3.5. Chromatographic Procedures

The HPLC analysis of the *G. elata* tuber were done on an Agilent 1260 series system (Agilent Technologies, Santa Clara, CA, USA) consisting of a G1311B pump, a G4212B DAD detector, and a G1329B auto-sampler. The YMC-Tyiart C18 column (250 × 4.6 mm, 5 μm) was adopted for the analysis. The mobile phase consisted of A (0.1% phosphate solution) and B (acetonitrile). The gradient mode was as follows: 3–5% B for 0–11 min; 5% B for 11–18 min; 5–14% B for 18–31 min; 14% B for 31–38 min; 14–20% B for 38–48min; 20–24% B for 48–55 min; 24–80% B for 55–75 min; 80–100% B for 75–80 min; 100% B for 80–95 min; 100–70% B for 95–100 min; 70–50% B for 100–105 min; 50–30% B for 105–110 min; 30–3% B for 110–115 min; 3% B for 115–130 min. The flow rate was set at 1.0 mL/min. The detection wavelength was 220 nm. The column temperature was set at 35 °C and sample volume was 10 μL.

3.6. Theory of the QAMS Method

Methods for calculating the RCFs have been previously reported [24,37]. First, gastrodin was selected as the internal standard, and a multipoint method (Equation (1)) was used to calculate the relative correction factors (RCF) for *p*-hydroxy benzaldehyde, *p*-hydroxybenzyl alcohol, parishin A, parishin B, parishin E, and parishin C. Then the content of the measured component was calculated according to Equation (2) [38].

The RCFs were calculated using the calibration curves as follows:

$$f_{k/s} = \frac{\mathrm{a_k}}{\mathrm{a_s}} \tag{1}$$

The content of the measured component was calculated as follows:

$$\mathrm{C_k} = \frac{\mathrm{A_k}}{\left(\mathrm{A_s} \times f_{\mathrm{k/s}}\right)} \tag{2}$$

where, $\mathrm{a_s}$ is the ratio of the slope of internal standard reference calibration equations; $\mathrm{a_k}$ is the ratio of the slope of measured component calibration equations; $\mathrm{A_k}$ is the peak area of the measured component; and $\mathrm{A_s}$ is the peak area of the internal standard reference [37].

The content of the multi-marker components measured by QAMS was compared with results from ESM, to validate the methods of QAMS.

3.7. Data Analysis

We used the ESM and QAMS to calculate the seven components in 21 batches of *G. elata* tuber, to verify the feasibility of QAMS. At the same time, HCA was performed using the heat map software of Heml 1.0, to further investigate the difference among the *G. elata* tuber samples. The data were analyzed and evaluated by the Similarity Evaluation System for the chromatographic fingerprint of TCM (Version 2012), to evaluate similarities of the chromatographic profiles of the *G. elata* tuber.

4. Conclusions

In this study, the quality assessment method of *G. elata* tubers were established using QAMS methods, in combination with HPLC fingerprints analyses. The *G. elata* tubers from different areas were analyzed by HPLC fingerprints and the contents of the seven components in *G. elata* tuber samples was determined by the QAMS method. On the basis of these results, the quality of *G. elata* tubers could be quantified and better identified comprehensively by HCA of synthesis and similarity analysis. HPLC fingerprint analyses, combined with the QAMS methods, could be a powerful and reliable way to provide both qualitative insight and quantitative data for comprehensive quality assessment of the complex multi-component systems. QAMS combined with the HPLC fingerprint might offer a holistic phytochemical profile of botanicals, along with similarity analysis and HCA of synthesis, and the quality of *G. elata* tubers would be evaluated and better and more comprehensively identified. Moreover, in subsequent analyses, it is also necessary to combine the chemical analysis, biological evaluation, pharmacological activity, and other methods to evaluate the quality of *G. elata* tubers for better studying the clinical effect.

Author Contributions: Y.X. supervised the project and designed the experimental works; Y.L. performed the chemical analyses and wrote the paper; Y.Z., Z.Z., Y.H., and X.C. contributed to sample process and data analyses; Y.X. revised the paper. All authors read and approved the final manuscript.

Funding: This research was funded by Applied Basic Research Key Project of Yunnan (2017ZF005 and 2017ZF001).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations have been used in this manuscript.

HPLC	High performance liquid chromatography
QAMS	Qualitative and quantitative analysis of multi-component by single marker
ESM	External standard method
RSD	Relative standard deviation
HCA	Hierarchical cluster analysis
TCM	Traditional Chinese medicine
RCF	Relative correction factor

References

1. Zhan, H.D.; Zhou, H.Y.; Sui, Y.P.; Du, X.L.; Wang, W.H.; Dai, L.; Sui, F.; Huo, H.R.; Jiang, T.L. The rhizome of *Gastrodia elata* Blume–An ethnopharmacological review. *J. Ethnopharmacol.* **2016**, *189*, 361–385. [CrossRef] [PubMed]
2. Li, Z.F.; Wang, Y.W.; Ouyang, H.; Lu, Y.; Qiu, Y.; Feng, Y.L.; Jiang, H.L.; Zhou, X.; Yang, S.L. A novel dereplication strategy for the identification of two new trace compounds in the extract of *Gastrodia elata* using UHPLC/Q-TOF-MS/MS. *J. Chromatogr. B* **2015**, *988*, 45–52. [CrossRef] [PubMed]
3. Hu, M.; Yan, H.; Fu, Y.; Jiang, Y.; Yao, W.; Yu, S.; Zhang, L.; Wu, Q.; Ding, A.; Shan, M. Optimal extraction study of gastrodin-type components from *Gastrodia Elata* tubers by response surface design with integrated phytochemical and bioactivity evaluation. *Molecules* **2019**, *24*, 547. [CrossRef] [PubMed]
4. Heo, J.C.; Woo, S.U.; Son, M.; Park, J.Y.; Choi, W.S.; Chang, K.T.; Kim, S.U.; Yoon, E.K.; Shin, S.H.; Lee, S.H. Anti-tumor activity of Gastrodia elata Blume is closely associated with a GTP-Ras-dependent pathway. *Oncol. Rep.* **2007**, *18*, 849–853. [CrossRef]
5. Hu, Y.H.; Li, C.Y.; Shen, W. Gastrodin alleviates memory deficits and reduces neuropathology in a mouse model of Alzheimer's disease. *Neuropathology* **2014**, *34*, 370–377. [CrossRef] [PubMed]
6. Zuo, Y.; Deng, X.; Wu, Q. Discrimination of gastrodia elata from different geographical origin for quality evaluation using newly-build near infrared spectrum coupled with multivariate analysis. *Molecules* **2018**, *23*, 1088. [CrossRef]
7. Zhao, Y.; Kang, Z.J.; Zhou, X.; Yang, S.L. An edible medicinal plant-Gastrodia elata Bl. J. *Guizhou Norm. Univ.* **2013**, *31*, 9–12.

8. Kang, C.; Lai, C.J.; Zhao, D.; Zhou, T.; Liu, D.H.; Lv, C.; Wang, S.; Kang, L.; Yang, J.; Zhan, Z.L.; et al. A practical protocol for comprehensive evaluation of sulfur-fumigation of Gastrodia Rhizoma using metabolome and health risk assessment analysis. *J. Hazard. Mater.* **2017**, *340*, 221–230. [CrossRef] [PubMed]
9. Zervakis, G.I.; Koutrotsios, G.; Katsaris, P. Composted versus raw olive mill waste as substrates for the production of medicinal mushrooms: An assessment of selected cultivation and quality parameters. *Biomed. Res. Int.* **2013**, *2013*, 546830. [CrossRef]
10. Chen, W.C.; Lai, Y.S.; Lu, K.H.; Lin, S.H.; Liao, L.Y.; Ho, C.T.; Sheen, L.Y. Method development and validation for the high-performance liquid chromatography assay of gastrodin in water extracts from different sources of *Gastrodia elata* Blume. *J. Food Drug Anal.* **2015**, *23*, 803–810. [CrossRef] [PubMed]
11. Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People's Republic of China. Part I*; Chinese Medical Science and Technology Press: Beijing, China, 2015.
12. Ojemann, L.M.; Nelson, W.L.; Shin, D.S.; Rowe, A.O.; Buchanan, R. Tian ma, an ancient Chinese herb, offers new options for the treatment of epilepsy and other conditions. *Epilepsy Behav.* **2006**, *8*, 376–383. [CrossRef] [PubMed]
13. Xie, M.; Sha, M.; Zhai, Q.C.; Yu, H.; Guo, L.; Wei, Y.Q.; Li, Y. Research Progress on Parishins from *Gastrodia Elata*. *Guangdong Chem. Ind.* **2016**, *22*, 93–95.
14. FDA. *Guidance for Industry-Botanical Drug Products*; U. S. Food and Drug Administration: Silver Spring, MD, USA, 2004.
15. SFDA. *Technical Requirements for Studying Fingerprint of Traditional Chinese Medicine Injections (Draft)*; Drug Administration Bureau of China: Beijing, China, 2000.
16. EMEA. *Guidance on Quality of Herbal Medicinal Products/Traditional Herbal Medicinal Products*; European Medicines Agency: London, UK, 2006.
17. Cui, L.L.; Zhang, Y.; Shao, W.; Gao, D. Analysis of the HPLC fingerprint and QAMS from *Pyrrosia* species. *Ind. Crop. Prod.* **2016**, *85*, 29–37. [CrossRef]
18. Schaneberg, B.T.; Crockett, S.; Khan, I.A. The role of chemical fingerprinting: Application to Ephedra. *Phytochemistry* **2003**, *62*, 911–918. [CrossRef]
19. Bauer, R. Quality criteria and phytopharmaceuticals: Can acceptable drug standards be achieved? *Drug Inf. J.* **1998**, *31*, 101–110. [CrossRef]
20. Wang, X.; Tan, Y.; Wang, D.J.; Qing, D.S.; Yao, L.C.; Li, L.Y.; Luo, W.Z. Application and advance of QAMS in quality control of traditional Chinese medicines. *Chin. Tradit. Pat. Med.* **2016**, *38*, 395–402.
21. Li, D.W.; Zhu, M.; Shao, Y.D.; Shen, Z.; Weng, C.C.; Yan, W.D. Determination and quality evaluation of green tea extracts through qualitative and quantitative analysis of multi-components by single marker (QAMS). *Food Chem.* **2016**, *197*, 1112–1120. [CrossRef]
22. Song, Y.; Wang, Z.; Zhu, J.; Yan, L.; Zhang, Q.; Gong, M.; Wang, W. Assay of evodin, evodiamine and rutaecarpine in Fructus Evodiae by QAMS. *Zhongguo Zhong Yao Za Zhi* **2009**, *34*, 2781–2785. [PubMed]
23. Lei, Y.C. *Authenticity Identification and Quality Assessment of Gastrodia tuber (Tianma) Based on Chemical Characteristics*; Chengdu University of TCM: Chengdu, China, 2015; p. 5.
24. Gao, X.Y.; Jiang, Y.; Lu, J.Q.; Tu, P.F. One single standard substance for the determination of multiple anthraquinone derivatives in rhubarb using high-performance liquid chromatography-diode array detection. *J. Chromatogr. A* **2009**, *1216*, 2118–2123. [CrossRef]
25. Peng, Y.; Dong, M.H.; Zou, J.; Liu, Z.H. Analysis of the HPLC Fingerprint and QAMS for Sanhuang Gypsum Sou. *J. Anal. Methods Chem.* **2018**. [CrossRef]
26. Wang, N.; Li, Z.Y.; Zheng, X.L.; Li, Q.; Yang, X.; Xu, H. Quality Assessment of Kumu Injection, a Traditional Chinese Medicine Preparation, Using HPLC Combined with Chemometric Methods and Qualitative and Quantitative Analysis of Multiple Alkaloids by Single Marker. *Molecules* **2018**, *23*, 856. [CrossRef] [PubMed]
27. He, B.; Liu, Y.; Tian, J.; Li, C.H.; Yang, S.Y. Study on quality control of Houttuynia Cordata, a tradtional Chinese medicine by fingerprint combined with quantitative analysis of multi-components by single marker. *China J. Chin. Mater. Med.* **2013**, *38*, 2682–2689.
28. Ding, L.Y.; Zhou, L.; Wang, L.N.; Guo, Y.H.; Wang, H. Multi-components quantitation by one marker for simultaneous content determination of four components in Psoralea corylifolia. *Chin. J. Exp. Tradit. Med Formulae* **2013**, *19*, 152–154.
29. Dou, Z.H.; Qiao, J.; Bian, L.; Hou, J.Y.; Mao, C.F.; Chen, Z.X.; Shi, Z. Combinational quality control method of Rhei Radix et Rhizoma based on fingerprint and QAMS. *J. Chin. Pharm. Sci.* **2015**, *50*, 442–448.

30. Yan, D.M.; Chang, Y.X.; Kang, L.Y.; Gao, X.M. Quality evaluation and regional analysis of Psoraleae Fructus by HPLC-DAD-MS/MS plus chemometrics. *Chin. Herb. Med.* **2010**, *2*, 216–223.
31. Chen, W.H.; Luo, D. Research Development on Pharmacological Action in *Gastrodia elata blume* and *Gastrodia Elata Polysaccharide*. *China J. Drug Eval.* **2013**, *30*, 132–141.
32. Lu, G.W.; Zou, Y.J.; Mo, Q.Z. Kinetic aspects of absorption, distribution, metabolism and excretion of ~3H-gastrodin in rats. *Yao Xue Xue Bao* **1985**, *20*, 167–172. [PubMed]
33. Choi, S.R.; Jang, I.; Kim, C.S.; You, D.H.; Kim, J.Y.; Kim, Y.G.; Ahn, Y.S.; Kim, J.M.; Kim, Y.S.; Seo, K.W. Changes of components and quality in gastrodiae rhizoma by different dry methods. *Korean J. Med. Crop Sci.* **2011**, *19*, 354–361. [CrossRef]
34. KPC Pharmaceuticals, Inc. The Invention Relates to Gastrodin Compound, Preparation Method, Preparation and Application. CN107056853A, 2017.
35. Wang, W.Y. *Study on Solubility of p-Hydroxybenzaldehyde, m-Hydroxybenzaldehyde and Their Mixture in Supercritical Carbon Dioxide*; Beijing University of Chemical Technology: Beijing, China, 2014.
36. Wang, W.J. *Study on the Aerobic Oxidation of o/p-Cresol to o/p-Hydroxybenzaldehyde Catalyzed by Metallopoprhyrinns*; Beijing University of Chemical Technology: Beijing, China, 2013.
37. Hou, J.J.; Wu, W.Y.; Da, J.; Yao, S.; Long, H.L.; Yang, Z.; Cai, L.Y.; Yang, M.; Liu, X.; Jiang, B.H.; et al. Ruggedness and robustness of conversion factors in method of simultaneous determination of multi-components with single reference standard. *J. Chromatogr. A* **2011**, *1218*, 5618–5627. [CrossRef]
38. Huang, J.; Yin, L.; Dong, L.; Quan, H.F.; Chen, R.; Hua, S.Y.; Ma, J.H.; Guo, D.Y.; Fu, X.Y. Quality evaluation for Radix Astragali based on fingerprint, indicative components selection and QAMS. *Biomed. Chromatogr.* **2018**, *32*, e4343. [CrossRef]

Sample Availability: Samples of the compounds are available from the authors.

molecules

MDPI

Article

Protein-Based Fingerprint Analysis for the Identification of *Ranae Oviductus* Using RP-HPLC

Yuanshuai Gan [1], Yao Xiao [1], Shihan Wang [2], Hongye Guo [1], Min Liu [1], Zhihan Wang [3] and Yongsheng Wang [1,*]

[1] College of Pharmacy, Jilin University, Changchun 130021, China; ganys18@mails.jlu.edu.cn (Y.G.); xiaoyao17@mails.jlu.edu.cn (Y.X.); guohy18@mails.jlu.edu.cn (H.G.); liumin17@mails.jlu.edu.cn (M.L.)

[2] College of Chinese Herbal Medicine, Jilin Agricultural University, Changchun 130118, China; wsh8805@163.com

[3] Department of Physical Sciences, Eastern New Mexico University, Portales, NM 88130, USA; zhihan.wang@enmu.edu

* Correspondence: mikewangwys@outlook.com or wys@jlu.edu.cn

Received: 16 April 2019; Accepted: 29 April 2019; Published: 30 April 2019

Abstract: This work demonstrated a method combining reversed-phase high-performance liquid chromatography (RP-HPLC) with chemometrics analysis to identify the authenticity of *Ranae Oviductus*. The fingerprint chromatograms of the *Ranae Oviductus* protein were established through an Agilent Zorbax 300SB-C8 column and diode array detection at 215 nm, using 0.085% TFA (*v*/*v*) in acetonitrile (A) and 0.1% TFA in ultrapure water (B) as mobile phase. The similarity was in the range of 0.779–0.980. The fingerprint chromatogram of *Ranae Oviductus* showed a significant difference with counterfeit products. Hierarchical clustering analysis (HCA) and principal component analysis (PCA) successfully identified *Ranae Oviductus* from the samples. These results indicated that the method established in this work was reliable.

Keywords: *Ranae Oviductus*; identification; protein; RP-HPLC; fingerprint

1. Introduction

Rana chensinensis is mainly distributed in the Changbai Mountain area, China. *Ranae Oviductus* is the dried oviduct of female *Rana temporaria chensinensis* David. The *Ranae Oviductus* is a potent traditional Chinese medicine that has been used in clinical studies for thousands of years. Today it is widely used as a nutrient food. It has been reported that *Ranae Oviductus* has significant effects in enhancing immunity, anti-fatigue, anti-aging, and lowering blood fat [1–4]. As a precious traditional Chinese medicine, *Ranae Oviductus* has been in short supply because of its limited production [5]. Its high price and lucrative profits have tempted many counterfeit products, such as bullfrog oviduct, toad oviduct, or frog oviduct, to inundate the market, resulting in the uneven quality of *Ranae Oviductus* in the market [6,7]. Those counterfeits have a similar appearance but have less efficacy. To guarantee the quality of *Ranae Oviductus*, its authenticity identification has attracted more and more attention from the pharmacists, doctors, and medicinal scientists. The identification method of *Ranae Oviductus* is still under development. In the 2005 China Pharmacopoeia, the appearance and expansion degree were employed as discriminating items of *Ranae Oviductus* [8]. Our group has reported using UV spectra to identify *Ranae Oviductus* [9]. According to a previous study, it is difficult to identify the *Ranae Oviductus* and counterfeit products using traditional methods [10]. Therefore, it is essential to establish a highly reliable method for the identification of *Ranae Oviductus*.

More than 40% of the components in *Ranae Oviductus* are proteins and the proteins are the major bioactive components of *Ranae Oviductus* [11,12]. However, the identification of *Ranae Oviductus* and counterfeit products using HPLC based on protein has not been studied yet. In addition, reversed-phase high-performance liquid chromatography (RP-HPLC) is a simple, fast, and effective technique for protein separation and characterization, as used for protein in milk, wheat gliadin, and transgenic zein [13–15]. On the other hand, the fingerprint chromatogram is considered as a comprehensive qualitative and quantitative method for the identification of different species, especially in the quality assessment of traditional Chinese medicines [16]. The World Health Organization (WHO) has admitted the use of chromatographic fingerprints as an identification strategy for traditional Chinese medicinal preparations [17]. Many reports have employed HPLC fingerprint chromatograms to study the quality control of traditional Chinese medicines. For example, Lu et al. used the HPLC fingerprint to identify Chinese *Angelica* from related umbellifer herbs. Sun et al. analyzed polysaccharides from different *Ganoderma*. Li et al. established the fingerprint analysis of polyphenols, which were extracted from pomegranate peel, with reliable results [18–20].

In this work, the main proteins components of *Ranae Oviductus* were used as the study objects. We used RP-HPLC to establish a fingerprint method for the identification of *Ranae Oviductus*. Ten batches of *Ranae Oviductus* were collected from different main producing areas of the Changbai Mountains. A protein reference chromatogram was established using those *Ranae Oviductus*, based on protein composition similarity analysis. Furthermore, the difference between the authentic *Ranae Oviductus* and counterfeit products were investigated. The results were verified via a chemometric approach, utilizing principal component analysis and hierarchical clustering analysis. Both showed that the newly established *Ranae Oviductus* identification method was reliable.

2. Materials and Methods

2.1. Chemicals and Samples

The petroleum ether, guanidine hydrochloride, and ammonium sulfate analytical grade were purchased from Beijing Chemical Factory (Beijing, China). The dithiothreitol (DTT) and trifluoroacetic acid (TFA) were purchased from Sigma-Aldrich (St. Louis, MO, USA). The HPLC-grade acetonitrile (MeCN) and HPLC-grade methanol were purchased from Fisher (Fisher Scientific, USA). The ultrapure water was obtained from a gradient water purification system (Water Purifier, Sichuan, China).

Ranae Oviductus, bullfrog oviduct, toad oviduct and frog oviduct were provided by Jilin Province Rana Industry Association which were collected from the Changbai Mountain area in the Jilin province of China. The specific location is shown on the map in Figure 1. Ten batches of *Ranae Oviductus* samples were collected from different regions from the main producing area of the Changbai Mountain range. The specific collection information is shown in Table 1.

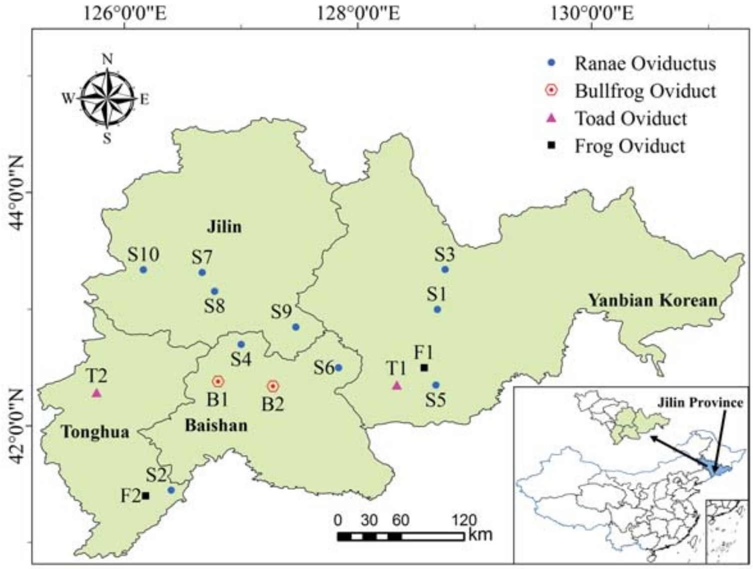


Figure 1. Distribution map of origins for *Ranae Oviductus* and its counterfeits in the Changbai mountain area.

Table 1. Origin and collecting date of the *Ranae Oviductus* samples and their counterfeits.

No.	Name of Medicine	Origin	Collection Date
S1	*Ranae Oviductus*	Yanbian, Jilin	2016.3
S2	*Ranae Oviductus*	Tonghua, Jilin	2016.1
S3	*Ranae Oviductus*	Yanbian, Jilin	2016.3
S4	*Ranae Oviductus*	Baishan, Jilin	2015.11
S5	*Ranae Oviductus*	Yanbian, Jilin	2016.3
S6	*Ranae Oviductus*	Baishan, Jilin	2015.11
S7	*Ranae Oviductus*	Jilin, Jilin	2016.12
S8	*Ranae Oviductus*	Jilin, Jilin	2016.12
S9	*Ranae Oviductus*	Jilin, Jilin	2015.11
S10	*Ranae Oviductus*	Jilin, Jilin	2015.11
B1	Bullfrog Oviduct	Baishan, Jilin	2016.12
B2	Bullfrog Oviduct	Baishan, Jilin	2016.12
T1	Toad Oviduct	Yanbian, Jilin	2016.10
T2	Toad Oviduct	Tonghua, Jilin	2016.11
F1	Frog Oviduct	Yanbian, Jilin	2016.10
F2	Frog Oviduct	Tonghua, Jilin	2016.11

2.2. Protein Extraction

The dried *Ranae Oviductus* was pulverized into a powder (passing through a 20-mesh sieve) and degreased with petroleum ether at room temperature. After filtration, the powder was placed in an oven at 55 °C for 1 h. Afterward, 0.50 g of the sample was added to PBS buffer (50 mL, 0.1 M pH 7.4). After continuously stirring for 8 h, the mixture was centrifuged at 5000 r/min for 15 min. The supernatant was collected and the precipitate was extracted again. The two centrifugal supernatants were combined. To the supernatant, an ammonium sulfate solid was slowly added until to 60% saturation [21,22]. The mixture was centrifuged at 8000 r/min for 20 min after standing at 4 °C for 1 h. The precipitate was dissolved in 6 M guanidine hydrochloride (containing 10 mM DTT) [23,24], and dialyzed in distilled water in a dialysis bag (molecular weight cutoff: 8000 Da) for 12 h [25]. The sample solution was finally scaled to 5 mL with 6 M guanidine hydrochloride (containing 10 mM DTT) in a volumetric flask and filtrated with a 0.45 µm filter membrane prior HPLC injection [26]. The preparations of bullfrog oviduct, toad oviduct and frog oviduct were the same as that for *Ranae Oviductus*.

2.3. RP-HPLC Chromatography Analysis

The samples were separated using an Agilent Technologies 1200 Series liquid chromatograph (Agilent Technologies, Pittsburgh, PA, USA) equipped with a quaternary pump, autosampler, thermostatted column compartment, diode array detector (DAD), and UV detector. The columns used were the Agilent Zorbax 300SB-C8 column (250 × 4.6 mm, 5 µm) and Agilent Zorbax SB-C18 column (250 × 4.6 mm, 5 µm) with mobile phase A (0.085% TFA in *v/v* with acetonitrile) and mobile phase B (0.1% TFA in *v/v* with ultrapure water) [27,28]. Gradient elution was adopted as follows, from 12–30% A in the first 52 min, and from 30–44% A in the next 28 min. The injection volume was 20 µL. The optimized separation conditions were tested under the different detection wavelengths, flow rates and temperatures [29]. The data were recorded and processed using the Agilent Chemstation software.

2.4. Validation of the RP-HPLC Method

Ranae Oviductus sample (S1) was used to verify the RP-HPLC method. A precision analysis was carried out by repeatedly injecting the same solution 5 times on the same day. The repeatability was assessed by injecting 5 separate solutions obtained from the same *Ranae Oviductus* sample. The stability was evaluated by analyzing the same sample solution at different time periods of 0, 2, 4, 8, 16 and 24 h at room temperature.

2.5. Establishment of the HPLC Fingerprint

The common characteristic peaks and similarities of fingerprint data of 10 batches of *Ranae Oviductus* were investigated using the professional software Similarity Evaluation System for the Chromatographic Fingerprint, according to the recommendations of the State Food and Drug Administration (SFDA). The HPLC fingerprint data of the samples were imported to the evaluation system (the solvent peaks in the first 4 min were removed and the time window was set at 0.2 s). The calibration method was multi-point calibration. The significant common peaks were labeled as mark peaks and the reference chromatogram fingerprint was generated with a mean value method. The similarity of the fingerprint data was represented by a correlation coefficient (similarity) and the higher similarity between the two samples resulted in a correlation coefficient value close to 1. The correlation coefficients of all chromatograms of 10 batches of *Ranae Oviductus* samples were calculated throughout the study and a correlation analysis was performed.

2.6. Data Analysis

Hierarchical clustering analysis (HCA) is a cluster analysis technique that reflects the similarities and differences between samples in the form of a hierarchical tree diagram [30,31]. This method is easier to observe than the complex raw data. Based on the clustering method between different groups and the Pearson correlation intervals, SPSS (version 25.0; SPSS Inc., Chicago, IL, USA) was used to group the different samples in this study.

Principal component analysis (PCA) is a classification method that uses dimensionality reduction techniques to simplify numerous original variables into several representative composite indicators [32,33]. According to the contribution rate of each comprehensive indicator, the information of the original data could be reflected when using appropriate numbers of principal components (PCs) [34]. In this study, PCA was performed using SPSS (version 25.0; SPSS Inc., Chicago, IL, USA) and the fractional scatter plot was interpreted by the relationship between PC1, PC2, and PC3 for visual analysis of the data matrix.

3. Results and Discussion

3.1. Optimization of the RP-HPLC Conditions

In order to improve the separation rate of the proteins in *Ranae Oviductus*, the *Ranae Oviductus* (S1) collected from the China Changbai mountain area were systematically investigated. The RP-HPLC chromatography method was optimized through the detection wavelength, separation column, flow rate and temperature. Three classical UV detection conditions were previously reported: 215 nm corresponding to the maximum absorption of peptide bonds; 254 nm corresponding to the maximum absorption of phenylalanine residues; and 280 nm corresponding to tyrosine and maximum absorption of tyrosine residues and tryptophan residues [35]. Figure 2a shows the UV absorption diagram of *Ranae Oviductus* using a diode array detector (DAD) with a wavelength range of 195–300 nm. The red region in the diagram indicated a larger absorption value. Although obvious solvent peaks around 215 nm were observed, the analysis of the core substance was not affected. The UV absorption diagram suggested that the separation effect at 215 nm was better than 254 nm and 280 nm.

Two types of columns (Agilent Zorbax SB-C18 column 250 × 4.6 mm, 5 μm, 80 Å and Agilent Zorbax 300SB-C8 column 250 × 4.6 mm, 5 μm, 300 Å) were used to examine the column effect on the protein separation of *Ranae Oviductus*. The results showed that the C8 column had a higher separation rate than the C18 column, which could be attributed to the large molecular weight of the proteins (Figure 2b). Therefore, the C8 column with a 300 Å pore diameter was selected for this study.

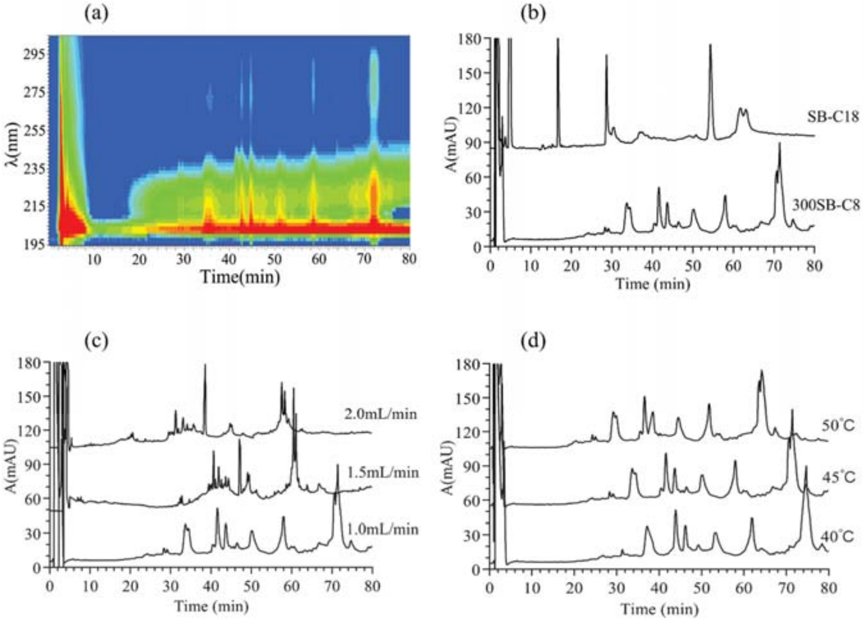


Figure 2. Optimization of reversed-phase high-performance liquid chromatography (RP-HPLC) separation method of the proteins from *Ranae Oviductus*. (**a**) The detection wavelength effect on the RP-HPLC chromatography of the *Ranae Oviductus* proteins. Diode array detector (DAD), 195–300 nm. (**b**) Column type effect on RP-HPLC chromatography of the *Ranae Oviductus* proteins (Agilent Zorbax 300SB-C8 column 250 × 4.6 mm, 5 μm, 300 Å and Agilent Zorbax SB-C18 column 250 × 4.6 mm, 5 μm, 80 Å). (**c**) Flow rate effect RP-HPLC chromatography of *Ranae Oviductus* (1.0 mL/min, 1.5 mL/min, 2.0 mL/min). (**d**) Temperature effect of RP-HPLC chromatography on the *Ranae Oviductus* proteins (40 °C, 45 °C, and 50 °C).

Since the flow rate of the mobile phase can affect the isolation efficiency, three flow rates (1.0, 1.5, 2.0 mL/min) were tested in this study. High flow rates showed that peaks overlapped (Figure 2c). The flow rate of 1.0 mL/min showed the highest separation effect and this, therefore, was chosen for the study.

On the other hand, the temperature played an important role in the RP-HPLC separation. Theoretically, high temperatures can increase the motion rate of proteins. In this study, three different temperatures (40, 45, and 50 °C) were investigated (Figure 2d). From the results, we could see that only one peak (t = 74.8 min) at 40 °C was observed, but two shoulder by shoulder peaks appeared at 45 and 50 °C. More proteins separated at 45 and 50 °C. Excessive temperature may damage the column's sorbent, therefore, 45 °C was selected as the optimum temperature.

3.2. RP-HPLC Methodology Validation

The accuracy of the RP-HPLC method was investigated through consecutive tests five times, using the same sample solution (*Ranae Oviductus* sample S1) within one day. The relative standard deviations (RSD) of the retention times and peak areas of the 12 common peaks were smaller than 2.02% and 4.23%, respectively. The repeatability was determined by injecting five separate sample solutions of the *Ranae Oviductus* sample. The results showed that the RSD of the retention time and peak area of the 12 common peaks were smaller than 2.96% and 5.62%, which suggested that the RP-HPLC method had good repeatability. The stability test was carried out at room temperature for 0, 2, 4, 8, 16 and 24 h. The RSD of the retention times and peak area were smaller than 2.62% and 5.22%. All tests indicated that the RP-HPLC method established in this work satisfied the requirements of protein fingerprinting analysis of *Ranae Oviductus*.

3.3. HPLC Fingerprint of Ranae Oviductus Protein

The protein chromatographic spectra of *Ranae Oviductus* collected from 10 sampling sites in Changbai Mountain area showed a similar profile using the optimized RP-HPLC method (Figure 3a). Based on the retention time, the 12 significant common-peaks were labeled with number 1 to 12. The 12 significant common-peaks in the *Ranae Oviductus* protein spectra were labeled as mark peaks according to the Chromatographic Fingerprint Similarity Evaluation System (2012 Edition) (Beijing, China). A reference fingerprint chromatographic spectrum of 10 batches of *Ranae Oviductus* was created (Figure 3b). The similarity was in the range of 0.779–0.980 (Table 2). The RSD value of the retention time of each common-peak was smaller than 4.70% and the RSD value of the relative peak area was smaller than 5.47%. This result pointed out that the common-peaks appearing in the chromatographic spectra were reliable in the analysis of *Ranae Oviductus*.

Table 2. Similarity values of 10 batches of *Ranae Oviductus* protein and reference chromatographic fingerprint spectra.

No.	Similarity	No.	Similarity
S1	0.779	S6	0.976
S2	0.906	S7	0.884
S3	0.967	S8	0.877
S4	0.970	S9	0.980
S5	0.970	S10	0.861

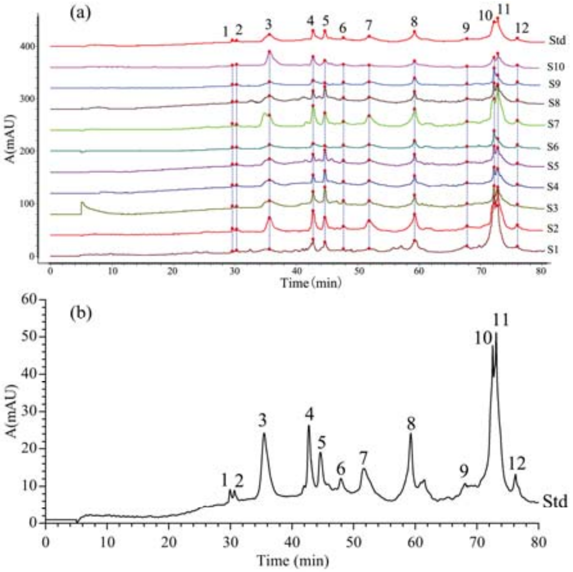


Figure 3. (**a**) HPLC fingerprint chromatographic spectra of 10 batches of *Ranae Oviductus* proteins. (**b**) The reference protein chromatographic spectra of *Ranae Oviductus*.

3.4. Fingerprint Spectra Analysis

The fingerprint spectra analysis of *Ranae Oviductus* and counterfeit products (bullfrog oviduct, toad oviduct and frog oviduct) were performed depending on the aforementioned optimized RP-HPLC method. The results showed a significant difference. By comparing Figure 4a,b, we could see that the significant common-peaks appeared at around 30 min in the reference fingerprint of *Ranae Oviductus*. In contrary, the counterfeit products, including the bullfrog oviduct, showed four common-peaks (peak A, peak B, peak C and peak D) in 0–30 min and the toad oviduct, showed three common-peaks (peak J, peak K, peak L) in the same time period. *Ranae Oviductus* showed 12 common-peaks (peak1-peak12) in 30–80 min, whereas, the bullfrog oviduct and toad oviduct only showed five common-peaks. The frog oviduct only showed four tiny common-peaks (Figure 4c), which was a finding consistent with a previous report. Huang, et al. [36] reported that the protein types in frog oviduct were less than that of other species by using the SDS-PAGE method. Both the protein extraction method and the RP-HPLC conditions were optimized according to the *Ranae Oviductus* sample, which may have not been adequate for frog oviduct. Through the comparison, we noticed that even the three counterfeit products had a significant difference (Figure 4d). The bullfrog oviduct (nine peaks) and toad oviduct (eight peaks) had more peaks than the frog oviduct (four peaks), but the retention time was different. Therefore, although *Rana chensinensis*, bullfrog, toad, and frog are similar amphibians, they are not the same species. Their genetic differences cause the expression of different types of proteins in the fallopian tubes, so that in RP-HPLC chromatographic spectra, they showed significant differences. Those differences can be used to identify *Ranae Oviductus* and counterfeit products.

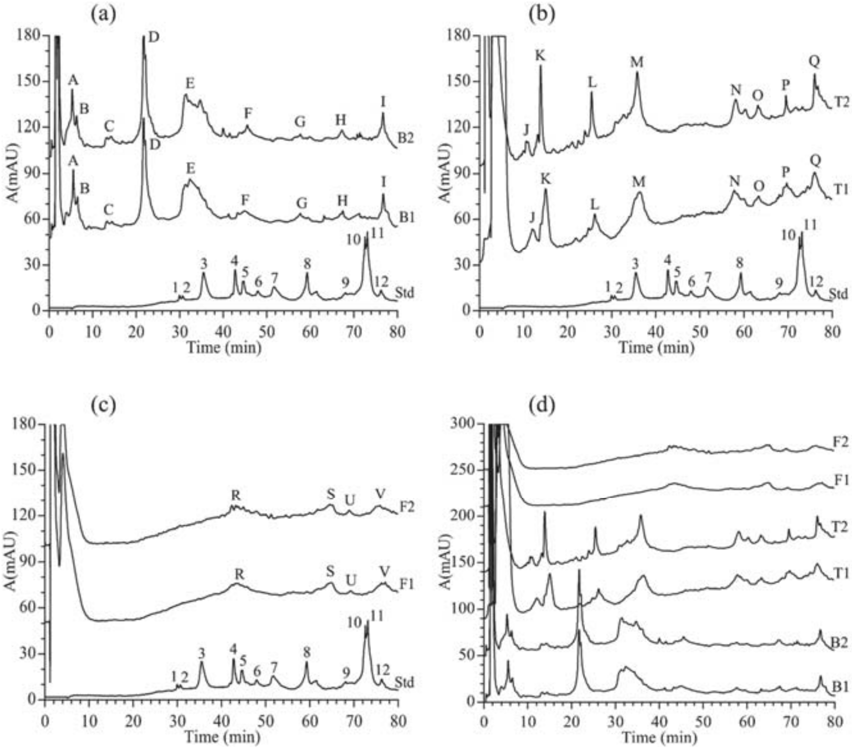


Figure 4. The comparison of *Ranae Oviductus* and counterfeit products. (**a**) Comparison of the protein HPLC fingerprint chromatogram of *Ranae Oviductus* (Std) and protein HPLC fingerprint chromatograms of the bullfrog oviduct (B1, B2). (**b**) Comparison of the protein HPLC fingerprint chromatogram of *Ranae Oviductus* (Std) and the protein HPLC fingerprint chromatograms of the toad oviduct (T1, T2). (**c**) Comparison of the protein HPLC fingerprint chromatogram of *Ranae Oviductus* (Std) and the protein HPLC fingerprint chromatograms of the frog oviduct (F1, F2). (**d**) Comparison of the protein HPLC fingerprint chromatograms of three counterfeits (bullfrog oviduct, toad oviduct, frog oviduct) of *Ranae Oviductus*.

3.5. Hierarchical Cluster Analysis (HCA)

Hierarchical cluster analysis was carried out using the relative peak areas of the characteristic peaks of *Ranae Oviductus* and counterfeit products. The 16 samples were analyzed using SPSS 25.0 software and the results are shown in Figure 5a. Obviously, there were four clusters when the interval of abscissa was 10. Cluster I, Cluster II and Cluster III were composed of the bullfrog oviduct sample, frog oviduct sample and toad oviduct sample, respectively. Cluster IV referred to the 10 samples of *Ranae Oviductus* used in the establishment of the fingerprint. The sample S1 with low similarity to *Ranae Oviductus* also showed a low correlation in Cluster IV. When the interval of abscissa was 25, the sample was divided into two clusters, one authentic and another one counterfeit.

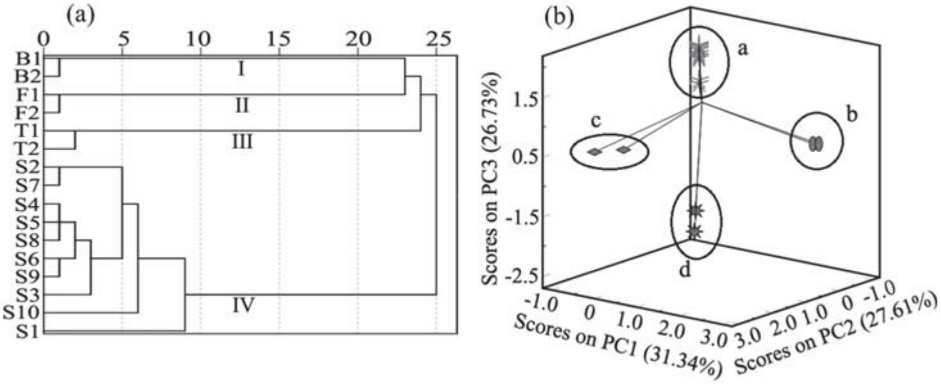


Figure 5. (**a**) The results of hierarchical cluster analysis of 10 batches of *Ranae Oveductus* and six counterfeit samples, (**b**) principal component analysis (PCA) score chart of 10 batches of *Ranae Oveductus* and six counterfeit samples in the first three principal components (PCs).

3.6. Principal Component Analysis(PCA)

As an effective data analysis technique, PCA has been used to study the classification of samples [37]. To directly reflect the difference between authentic and counterfeit products, 16 samples were used to perform the PCA analysis, based on the relative peak areas of the characteristic peaks of the samples. The variance contribution rates of the three main components (PC1, PC2, and PC3) were 31.34%, 27.61%, and 26.73%, respectively. The cumulative variance contribution rate of the three PCs was 85.68% and those variables reflected the majority of total information. To visualize the analysis results, the score charts were drawn using the three main components of PC1, PC2 and PC3 (Figure 5b). Four aggregation states are showed in Figure 5b. *Ranae Oviductus*, bullfrog oviduct, toad oviduct, and frog oviduct samples were classified in the a, b, c, and d regions, respectively. The *Ranae Oviductus* samples S1–10 could be classified in the same area (the a region), the bullfrog oviduct was classified in the b region, the toad oviduct was classified in the c region, and the frog oviduct was classified in the d region. The results were consistent with the HCA analysis, that both *Ranae Oviductus* and the counterfeit products were correctly classified. Comparing the similarity analysis with the HCA, PCA can provide a more visual comparison of the chromatograms.

4. Conclusions

This study used the RP-HPLC method and fingerprint technique to establish a chromatographic fingerprint of the proteins from *Ranae Oviductus*. Ten batches of *Ranae Oviductus* collected from the Changbai mountain area were used to analyze the protein components. The results showed 12 common-peaks in the reference fingerprint chromatographic spectrum. In combination with stoichiometry HCA and PCA, the results suggested that the method established in this work can satisfy the identification of *Ranae Oviductus* and counterfeit products. The method established in this work provides a promising approach for the identification of *Ranae Oviductus* and counterfeit products.

Author Contributions: Conceptualization, Y.W.; methodology, Y.W. and Y.G.; formal analysis, Y.G., Y.X., S.W. and H.G.; investigation, Y.G., Y.X., S.W., H.G. and M.L.; data curation, Y.G., S.W. and H.G.; writing—original draft preparation, Y.G., Y.X., S.W. and H.G.; writing—review and editing, Z.W., S.W. and Y.W.; visualization, Y.G., S.W., Z.W. and H.G.; supervision, Y.W.; project administration, S.W. and Y.W.; funding acquisition, Y.W.

Funding: This research was funded by Administration of Traditional Chinese Medicine of Jilin Province, grant number DBXM085-2018 and Jilin Bureau of Quality and Technical Supervision, grant number 2019037.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xie, C.; Zhang, L.J.; Zhang, W.Y.; Yang, X.; Fan, L.; Li, X. Immunomodulatory effect of *Oviductus Ranae* on the mice. *Chin. J. Gerontol.* **2010**, *30*, 3132–3133.
2. Li, Z.G.; Wang, C. The anti-fatigue effect of protein hydrolysate of *Oviductus Ranae* on mice and its physiological mechanism. *Chin. Sch. Phys. Educ.* **2017**, *4*, 81–87.
3. Mo, Y.; Yu, M.J.; Mo, Y.L. Protective effect of *Oviductus Ranae* on D-galactose-induced aging mice. *Chin. J. Gerontol.* **2011**, *31*, 1603–1604.
4. Peng, F.; Xu, F.; Liu, B.; Zhou, B.; Chen, X.; Zhao, Q. Effects of *Rana Temporaria Chensinensis David* egg oil on blood lipid in hyperlipemia rats. *Acad. J. Guangzhou Med. Coll.* **2003**, *31*, 57–59.
5. He, Z.; Tang, X.; Liu, J.; Yuan, Y.; Jiang, C.; Zhao, Y.; Wang, Y. Application of rapid PCR to authenticate *Ranae Oviductus*. *China J. Chin. Mater. Med.* **2017**, *42*, 2467–2472.
6. Hou, G.L.; Lu, B.Z.; Wang, C.L. Authentic identification of *Ranae Oviductus*. *Strait. Pharm. J.* **2007**, *19*, 63–64.
7. Jin, P.; Zhang, Y.; Wang, H.; Lan, M.; Zhang, H.; Sun, J.M. Advances in identification of forest frog oil. *Jilin J. Chin. Med.* **2018**, *38*, 1179–1180.
8. Liu, J.; Liu, S.; Liu, C. True or false identification of oviductus ranae. *Heilongjiang Med. Pharm.* **2009**, *32*, 20–21.
9. Wang, Y.S.; Jiang, D.C.; Bai, X.X.; Wang, E.S. Identification Research *Rana temportva Chensinensis David*'s Quality with UASLG. *Lishizhen Med. Mater. Med. Res.* **2006**, *17*, 2125–2127.
10. Zhang, W.; Wang, W.N.; Chen, F.F.; Zhang, L.; Yuan, D. Quality evaluation of *Oviductus Ranae* and similar products and fakes. *J. Shenyang Pharm. Univ.* **2012**, *29*, 951–958.
11. Hu, X.; Liu, C.B.; Chen, X.P.; Wang, L.M. Main nourishment components of *Oviductus Ranae*. *J. Jilin Agric. Univ.* **2003**, *25*, 218–220.
12. Hou, Z.H.; Zhao, H.; Yu, B.; Cui, B. Comprehensively analysis of components in *Oviductus Ranae*. *Sci. Technol. Food Ind.* **2017**, *38*, 348–352.
13. Ma, L.; Yang, Y.; Chen, J.; Wang, J.; Bu, D. A rapid analytical method of major milk proteins by reversed-phase high-performance liquid chromatography. *Anim. Sci. J.* **2017**, *88*, 1623–1628. [CrossRef]
14. Han, C.; Lu, X.; Yu, Z.; Li, X.; Ma, W.; Yan, Y. Rapid separation of seed gliadins by reversed-phase ultra performance liquid chromatography (RP-UPLC) and its application in wheat cultivar and germplasm identification. *Biosci. Biotechnol. Biochem.* **2015**, *79*, 808–815. [CrossRef]
15. Rodríguez-Nogales, J.M.; Cifuentes, A.; García, M.C.; Marina, M.L. Characterization of Protein Fractions from Bt-Transgenic and Non-transgenic Maize Varieties Using Perfusion and Monolithic RP-HPLC. Maize Differentiation by Multivariate Analysis. *J. Agric. Food Chem.* **2007**, *55*, 3835–3842. [CrossRef] [PubMed]
16. Xie, P.; Chen, S.; Liang, Y.-Z.; Wang, X.; Tian, R.; Upton, R. Chromatographic fingerprint analysis—a rational approach for quality assessment of traditional Chinese herbal medicine. *J. Chromatogr. A* **2006**, *1112*, 171–180. [CrossRef]
17. Sun, J.; Chen, P. Chromatographic fingerprint analysis of yohimbe bark and related dietary supplements using UHPLC/UV/MS. *J. Pharm. Biomed. Anal.* **2012**, *61*, 142–149. [CrossRef]
18. Lu, G.-H.; Chan, K.; Liang, Y.-Z.; Leung, K.; Chan, C.-L.; Jiang, Z.-H.; Zhao, Z.-Z. Development of high-performance liquid chromatographic fingerprints for distinguishing Chinese Angelica from related umbelliferae herbs. *J. Chromatogr. A* **2005**, *1073*, 383–392. [CrossRef]
19. Sun, X.; Wang, H.; Han, X.; Chen, S.; Zhu, S.; Dai, J. Fingerprint analysis of polysaccharides from different Ganoderma by HPLC combined with chemometrics methods. *Carbohydr. Polym.* **2014**, *114*, 432–439. [CrossRef] [PubMed]
20. Zhu, L.; Fang, L.; Li, Z.; Xie, X.; Zhang, L. A HPLC fingerprint study on Chaenomelis Fructus. *BMC Chem.* **2019**, *13*, 7. [CrossRef]
21. Harrysson, H.; Hayes, M.; Eimer, F.; Carlsson, N.-G.; Toth, G.B.; Undeland, I. Production of protein extracts from Swedish red, green, and brown seaweeds, Porphyra umbilicalis Kützing, Ulva lactuca Linnaeus, and Saccharina latissima (Linnaeus) J. V. Lamouroux using three different methods. *J. Appl. Phycol.* **2018**, *30*, 3565–3580. [CrossRef]

22. Huang, F.; Cockrell, D.C.; Stephenson, T.R.; Noyes, J.H.; Sasser, R.G. Isolation, purification, and characterization of pregnancy-specific protein B from elk and moose placenta. *Biol. Reprod.* **1999**, *61*, 1056–1061. [CrossRef] [PubMed]
23. Takakura, D.; Hashii, N.; Kawasaki, N. An improved in-gel digestion method for efficient identification of protein and glycosylation analysis of glycoproteins using guanidine hydrochloride. *Proteomics* **2014**, *14*, 196–201. [CrossRef]
24. Poulsen, J.W.; Madsen, C.T.; Young, C.; Poulsen, F.M.; Nielsen, M.L. Using Guanidine-Hydrochloride for Fast and Efficient Protein Digestion and Single-step Affinity-purification Mass Spectrometry. *J. Proteome Res.* **2013**, *12*, 1020–1030. [CrossRef]
25. Mouecoucou, J.; Villaume, C.; Sanchez, C.; Mejean, L. Effects of gum arabic, low methoxy pectin and xylan on in vitro digestibility of peanut protein. *Food Res. Int.* **2004**, *37*, 777–783. [CrossRef]
26. Vincent, D.; Rochfort, S.; Spangenberg, G. Optimisation of Protein Extraction from Medicinal Cannabis Mature Buds for Bottom-Up Proteomics. *Molecules* **2019**, *24*, 659. [CrossRef]
27. Ali, I.; Aboul-Enein, H.Y.; Singh, P.; Singh, R.; Sharma, B. Separation of biological proteins by liquid chromatography. *Saudi Pharm. J.* **2010**, *18*, 59–73. [CrossRef]
28. Esteve, C.; Del Rio, C.; Marina, M.L.; Garcia, M.C. Development of an ultra-high performance liquid chromatography analytical methodology for the profiling of olive (*Olea europaea* L.) pulp proteins. *Anal. Chim. Acta* **2011**, *690*, 129–134. [CrossRef]
29. Gao, P.; Shi, B.; Li, Z.; Wang, P.; Yin, C.; Yin, Y.; Zan, L. Establishment and Application of Infant Formula Fingerprints by RP-HPLC. *Food Anal. Method.* **2018**, *11*, 23–33. [CrossRef]
30. Cui, L.L.; Zhang, Y.Y.; Shao, W.; Gao, D.M. Analysis of the HPLC fingerprint and QAMS from Pyrrosia species. *Ind. Crop Prod.* **2016**, *85*, 29–37. [CrossRef]
31. Kannel, P.R.; Lee, S.; Kanel, S.R.; Khan, S.P. Chemometric application in classification and assessment of monitoring locations of an urban river system. *Anal. Chim. Acta* **2007**, *582*, 390–399. [CrossRef] [PubMed]
32. Wang, C.; Zhang, C.-X.; Shao, C.-F.; Li, C.-W.; Liu, S.-H.; Peng, X.-P.; Xu, Y.-Q. Chemical Fingerprint Analysis for the Quality Evaluation of Deepure Instant Pu-erh Tea by HPLC Combined with Chemometrics. *Food Anal. Method.* **2016**, *9*, 3298–3309. [CrossRef]
33. Goodarzi, M.; Russell, P.J.; Vander Heyden, Y. Similarity analyses of chromatographic herbal fingerprints: A review. *Anal. Chim. Acta* **2013**, *804*, 16–28. [CrossRef]
34. Nelson, P.R.C.; MacGregor, J.F.; Taylor, P.A. The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemom. Intell. Lab* **2006**, *80*, 1–12. [CrossRef]
35. Esteve, C.; Del Río, C.; Marina, M.L.; García, M.C. First Ultraperformance Liquid Chromatography Based Strategy for Profiling Intact Proteins in Complex Matrices: Application to the Evaluation of the Performance of Olive (*Olea europaea* L.) Stone Proteins for Cultivar Fingerprinting. *J. Agric. Food Chem.* **2010**, *58*, 8176–8182. [CrossRef] [PubMed]
36. Huang, Y.; Chang, L.; Zhang, S.W.; Yuan, D. Electrophoresis methods for the characterizaiton of *Ranae Oviductus* and its adulterants. *J. Shenyang Pharm. Univ.* **2017**, *34*, 1049–1054.
37. Fraige, K.; Pereira-Filho, E.R.; Carrilho, E. Fingerprinting of anthocyanins from grapes produced in Brazil using HPLC–DAD–MS and exploratory analysis by principal component analysis. *Food Chem.* **2014**, *145*, 395–403. [CrossRef] [PubMed]

Sample Availability: Not available.

molecules

MDPI

Article

A Quick and Efficient Non-Targeted Screening Test for Saffron Authentication: Application of Chemometrics to Gas-Chromatographic Data

Pietro Morozzi [1], Alessandro Zappi [1], Fernando Gottardi [2], Marcello Locatelli [3] and Dora Melucci [1,*]

[1] Department of Chemistry "G. Ciamician", University of Bologna, 40126 Bologna, Italy
[2] COOP ITALIA Soc. Cooperativa, Casalecchio di Reno, 40033 Bologna, Italy
[3] Department of Pharmacy, University "G. D'Annunzio" of Chieti-Pescara, 66100 Chieti, Italy
* Correspondence: dora.melucci@unibo.it; Tel.: +39-051-2099530; Fax: +39-051-2099456

Academic Editor: Thomas Letzel
Received: 10 June 2019; Accepted: 16 July 2019; Published: 17 July 2019

Abstract: Saffron is one of the most adulterated food products all over the world because of its high market prize. Therefore, a non-targeted approach based on the combination of headspace flash gas-chromatography with flame ionization detection (HS-GC-FID) and chemometrics was tested and evaluated to check adulteration of this spice with two of the principal plant-derived adulterants: turmeric (*Curcuma longa* L.) and marigold (*Calendula officinalis* L.). Chemometric models were carried out through both linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-DA) from the gas-chromatographic data. These models were also validated by cross validation (CV) and external validation, which were performed by testing both models on pure spices and artificial mixtures capable of simulating adulterations of saffron with the two adulterants examined. These models gave back satisfactory results. Indeed, both models showed functional internal and external prediction ability. The achieved results point out that the method based on a combination of chemometrics with gas-chromatography may provide a rapid and low-cost screening method for the authentication of saffron.

Keywords: saffron; adulteration; food authenticity; gas-chromatography; chemometrics

1. Introduction

The commercial product named "Saffron Powder" is a powdered spice obtained by crushing the filaments of the *Crocus sativus* L. flower [1]. Unfortunately, because of its high market price, this spice is one of the most often adulterated food products worldwide [2]. There are different kinds of possible frauds, the most frequent being the addition of foreign matter, such as derivatives from flowers of other plants, to increase the mass of the final product without adding costly pure saffron. In some cases, even total substitution of saffron powder with adulterants may be found [3].

The high market price of saffron is due to the laborious process required to obtain the spice and the limited areas of production [4]. The flower of *Crocus sativus* L. is indeed cultivated only in some regions of Asia (Kashmir, northern Iran) and Europe (Castilla la Mancha, Spain; Kozani, Greece; Abruzzo and Sardinia, Italy) [5]. Several Protected Designations of Origin (PDOs) have been created to protect the authenticity of saffron (as it has, for example, in the Italian "Zafferano dell'Aquila", one of the major areas in terms of production and global exports) [5]. Galvin-King et al. [6] report that the business volume concerning all herbs and spices is around four billion US dollars; economists soon expect growth up to 50%. As a consequence, the business volume of frauds is estimated to cause economic damage to the global food industry in the order of several tens of billions of US dollars [7].

In order to ensure the authenticity and the quality of saffron, a standard method is proposed by the International Organization for Standardization (ISO). In particular, the last international standard regulation regarding saffron quality (ISO 3632-1:2011) [1] mainly provides a UV-Vis spectrophotometric analysis to conventionally quantify the flavor strength (expressed as concentration of picrocrocin), the aroma strength (concentration of safranal), and the coloring strength (concentration of crocin) of saffron samples. However, this method has sometimes proved incapable of evaluating saffron adulteration [8] related to spectral interferences and to the impossibility to resolve chemicals present in the adulterants that show a similar UV-Vis absorbance.

Consequently, many different analytical methods have been developed to overcome this limitation; a complete and exhaustive description of all the relevant analytical techniques is given by Kiani et al. [9]. In particular, many other spectroscopic techniques [10–13], chromatographic techniques [14–16], and molecular-biological techniques [17–19] have been exploited. Among the molecular-biological techniques, the genome-based approach, usually based on DNA extraction [20], amplification, and sequencing, represents the principal strategy to ensure the food authenticity.

However, many of these procedures are time consuming and expensive, as they require highly specialized personnel and are based on destructive methodologies.

With the aim of by-passing the above-listed drawbacks, a preliminary study for a rapid, simple, and cheap screening test for the assessment of adulterated saffron is herein developed. In particular, a non-targeted approach is used.

The non-targeted approaches are increasingly used in the field of food authenticity because they allow the examining of food fingerprints, which were previously acquired by the use of spectroscopic, spectrometric, or chromatographic techniques. This check is performed holistically and without long, complicated, and problematic identification and quantification of specific and characteristic metabolites [21].

In this work, gas-chromatographic profiles are used as chemical fingerprints, because the patterns of the most volatile compounds are characteristic for odorous spices (such as saffron and their plant adulterants) and, consequently, they may represent important variables for the assessment of saffron authenticity [22–24].

In particular, this study presents a combined application of Heracles II (AlphaMos, Toulouse, France) instrumentation, a headspace flash gas-chromatography with flame ionization detection (HS-GC-FID), and chemometric techniques [25]. Heracles II provides gas-chromatographic profiles of the analyzed samples rapidly and without any chemical sample pre-treatment [25–28]. Thus, the gas-chromatographic fingerprints are subsequently submitted to chemometric modeling through a multivariate approach [29,30], allowing detection of the eventual adulteration of saffron.

The focus of this work is the evaluation of saffron adulteration by two of the most frequently used plant-derived adulterants: turmeric (*Curcuma longa* L.) and marigold (*Calendula officinalis* L.).

2. Results and Discussion

In this work, 61 samples of commercial spices were analyzed by Heracles II flash HS-GC-FID, which meant there were 244 objects or rows of the dataset matrices. Although several peaks were present in the obtained chromatograms, for the non-targeted approach used in this work it was not necessary to associate the identified chromatographic peaks with the corresponding volatile compounds.

Examples of the chromatograms of some analyzed samples are reported in Figure 1. It was evident that the discrimination of pure spices could be directly achieved by simply superimposing the GC chromatograms in Figure 1 without any need of chemometrics. Of course, pure samples are even distinguishable with eyes without any chemical analysis. What is interesting, however, is to discriminate *mixture* samples, which simulate adulterated saffron powders. This can be done only by chemometrics.

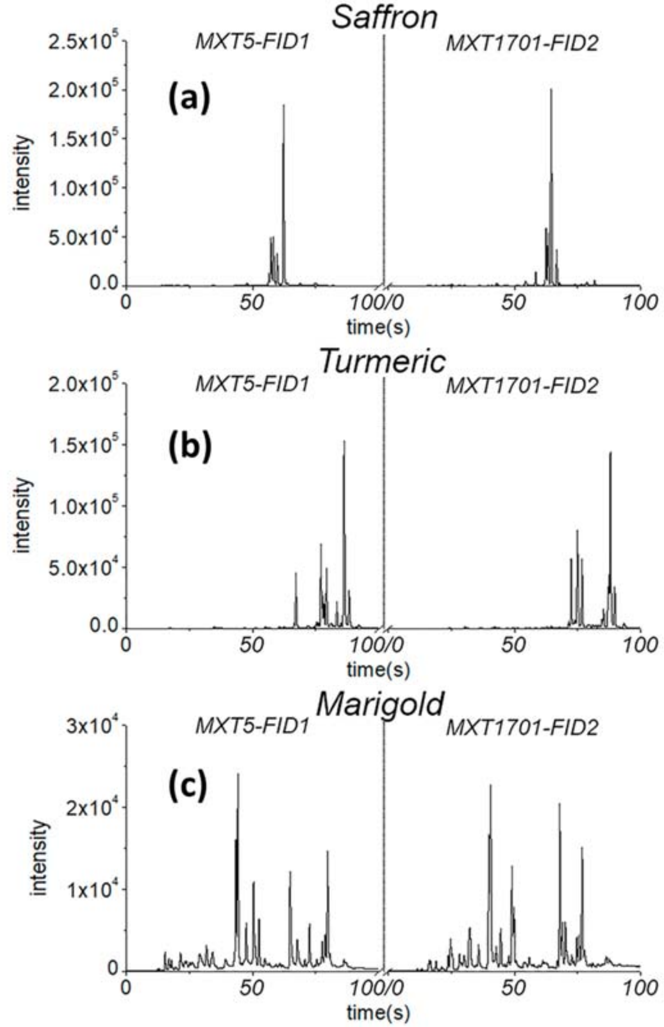


Figure 1. Representative gas-chromatographic (GC) fingerprints of saffron (**a**), turmeric (**b**), and marigold (**c**) obtained by Heracles II instrument. The chromatograms from column MXT5 are reported in the left part of the figure, while the chromatograms from column MXT1701 are reported on the right. These chromatograms were recorded simultaneously by the headspace flash gas-chromatography with flame ionization detection (HS-GC-FID).

Even if distinguishing pure samples is trivial, it is useful to create classification models based on pure standards. In fact, the models allow quantification of the dissimilarity of mixtures with respect to pure classes through parameters that are specific for each multivariate classification method.

From the obtained experimental data, two matrices were constructed: the area dataset (AD, 244 rows × 56 columns) and the intensity dataset (ID, 244 rows × 20,002 columns). More details will be given in the section Materials and Methods, paragraph 3.4 ("Working dataset").

Both matrices, as described previously, were subjected to the following chemometric elaborations (LDA and PLS-DA).

2.1. LDA Model and Results for AD

A preliminary PCA computed on the area dataset led us to find 42 outliers—20 outliers for the "Saffron" class, eight for the "Marigold" class, and 14 for the "Turmeric" class. This brought us to a dataset with dimensions 202 (objects) × 56 (variables). On this dataset, LDA was carried out. Leave-one-out cross validation (LOO-CV) was performed to internally validate the LDA model. The results of LOO-CV, in this case, could be expressed as the percentage of well-classified samples (NER), which for this LDA model was 100%. This result was obvious, since pure samples were considered.

The application of LDA produced the discriminant plot in Figure 2. Three clusters were evidenced, corresponding, as expected (100% NER), to the three a-priori classes (pure spices). In particular, the "Saffron" class was mostly discriminated from "Turmeric" along LD1 and from "Marigold" along LD2. Besides the three clusters, test samples were projected (asterisks). Table 1 summarizes all the test samples.

All the pure samples of the test set (pure_MR, pure_TR, and pure_SF) were assigned to the correct classes. They were correctly put inside the class spaces to which they were referred. What was particularly interesting was the behavior of the mixture samples; their distance from the pure spices clusters was significant. The mixture samples in Figure 2, although close to the "Saffron" class, moved away from it with an increasing percentage of adulterant. Moreover, the turmeric-adulterated samples (SFTR) got closer to the "Turmeric" class, moving along LD1, while the marigold-adulterated samples (SFMR) got closer to the "Marigold" class, moving along LD2. To quantify such behavior, the Euclidean distances between each point and each class centroid were computed, and the results are reported in Table 2. The class centroids were the points whose coordinates were the mean values of the coordinates of all the class objects. Thus, these could be considered as the "most representative" points for each class (although fictitious).

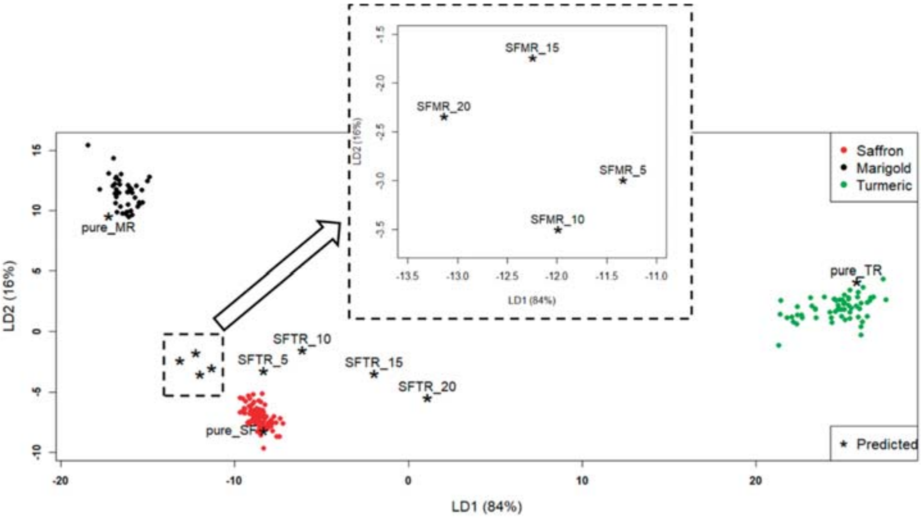


Figure 2. Linear discriminant analysis (LDA) discriminant plot, LD1 vs. LD2. The projected test samples (external validation results) are symbolized by asterisks (*). The graph portion inside the smaller dashed square is magnified into the greater dashed square.

Table 1. The test samples used for external validation: pure spices and artificial mixtures.

Test Samples	$\%_{W/W}$ of Saffron Adulteration	Code
Pure Saffron	-	pure_SF
Pure Turmeric	-	pure_TR
Pure Marigold	-	pure_MR
saffron + turmeric	5	SFTR_5
	10	SFTR_10
	15	SFTR_15
	20	SFTR_20
saffron + marigold	5	SFMR _5
	10	SFMR _10
	15	SFMR _15
	20	SFMR _20

From Table 2, it can be seen that the distances of the turmeric-adulterated samples (SFTR) from the "Saffron" class increased, and the distance from the "Turmeric" class decreased with an increasing percentage of adulteration. The situation was a bit more complicated for the SFMR samples, because their distances did not have a "linear" behavior with the adulterant percentage (in particular, SFMR_10 was farther from "Marigold" class than SFMR_5, and SFMR_20 was closer than SFMR_15), as can be seen from Figure 2. However, it is interesting to highlight that the distance of the farthest calibration saffron sample from the "Saffron" class centroid was 2.6. This distance could be considered as a sort of radius of the "Saffron" class, and all the mixture sample distances reported in Table 2 were higher than this value. This meant that, by computing the Euclidean distances of the projected samples from the class centroids, the LDA model could detect (at least qualitatively) a saffron sample adulterated by turmeric or marigold even down to the percentage of adulteration of $5\%_{W/W}$.

Table 2. Euclidean distances of the test samples reported in Table 1 from the three class centroids.

Sample Code	Saffron	Turmeric	Marigold
pure_SF	1.1	34.6	21.1
pure_TR	36.3	2.5	42.8
pure_MR	18.6	42.5	2.2
SFTR_5	3.8	33.4	16.7
SFTR_10	6.2	31.0	16.4
SFTR_15	7.6	27.2	20.7
SFTR_20	9.9	24.7	24.2
SFMR_5	4.8	36.3	15.3
SFMR_10	4.8	37.1	15.6
SFMR_15	6.4	37.1	13.8
SFMR_20	6.4	38.0	14.3

2.2. PLS-DA Model and Results for ID

A preliminary PCA computed on the intensity dataset led to finding four outliers (one sample) for the "Saffron" class and five outliers for the "Turmeric" class. Moreover, to reduce the computational cost while maintaining good data representation, one variable every ten was retained [25]. In this way, the ID dataset on which PLS-DA was carried out had dimensions of 235 × 2001. PLS-DA was chosen instead of LDA for this dataset due to the high number of variables and the high co-linearity between them. LDA requires the computation of the covariance matrix of the dataset, but it is not possible when the variables are co-linear [31]. Figure 3 shows the PLS-DA scores plot. As it can be seen in Figure 3a, Factor-1 and Factor-2 of PLS-DA together explained 82% of the X-explained variance and 50% of the Y-explained variance, which could be considered satisfactory to describe the dataset. From this scores plot, good discrimination of “Saffron” and “Turmeric” classes could be observed. The “Marigold” class, on the contrary, seemed to be overlapped to the “Saffron” class in the lower left part of the scores plot (third quadrant of the plot). However, when zooming in on this overlap zone, as it can be observed in the scores plot reported in Figure 3b, these two classes were found to be resolved.

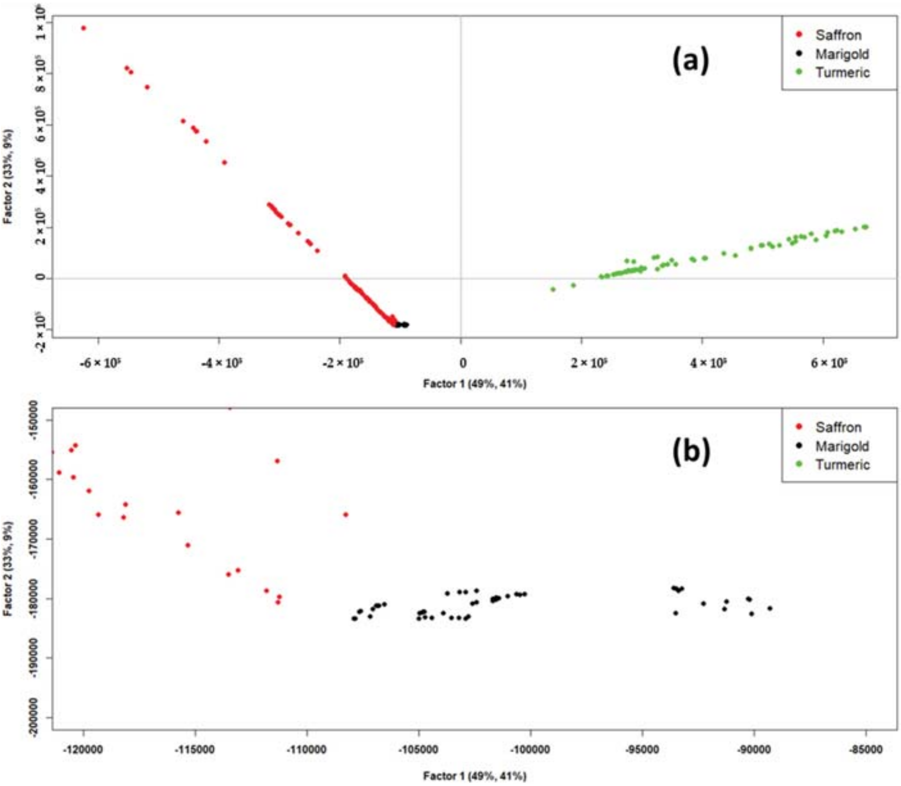


Figure 3. (**a**) Scores of partial least squares discriminant analysis (PLS-DA) model, Factor-1 vs. Factor-2. (**b**) Zoomed scores plot of the PLS-DA model, Factor-1 vs. Factor-2.

The CV was also performed to internally validate the PLS-DA model. Sensitivity and specificity for each class were computed according to Ballabio and Consonni (2013) [32] using 200 possible threshold values ranging from 0.1 to 1.1. The results are shown in Figure 4. Nine PLS-factors were used for “Saffron” and “Marigold” classes and three factors for “Turmeric” class (from Figure 3, it is easy to see that the discrimination of the “Turmeric” class was easier and required fewer factors than the discrimination of the other two). The vertical dashed lines in Figure 4 represent the chosen thresholds, which were 0.62 for “Saffron”, 0.56 for “Turmeric”, and 0.58 for “Marigold”. Thresholds were chosen as the highest value that maximized both sensitivity and specificity (1.0 or 100%) in order to have a restrictive rule for the class assignment.

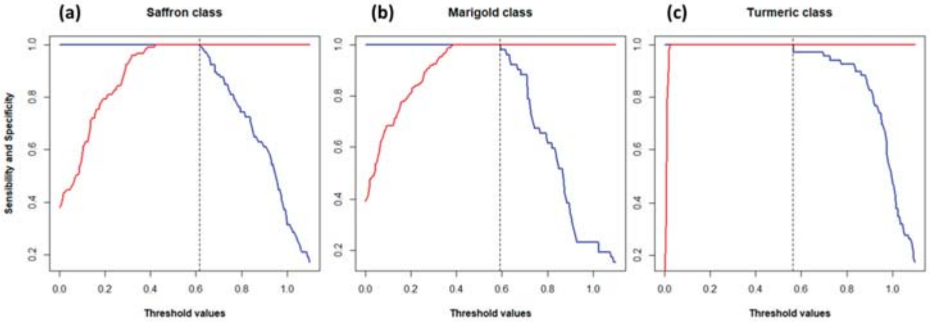


Figure 4. Sensitivity (blue lines) and specificity (red lines) for (**a**) "Saffron"; (**b**) "Marigold"; (**c**) "Turmeric" classes computed for each threshold value. Vertical dashed lines are the chosen thresholds for the corresponding class.

At this point, the test samples reported in Table 1 were projected onto the PLS-DA model to validate it. Table 3 shows the values of the dummy variables (y_marigold, y_turmeric, and y_saffron) and their corresponding standard deviation calculated by the PLS-DA model for the test samples. The pure samples (pure_MR, pure_TR, and pure_SF) could be considered well classified. Indeed, the calculated values of the dummy variables overcame the threshold values (i.e., belonging to the class considered) related to the pertaining class of each sample, while they did not overcome the thresholds (i.e., not belonging to class considered) related to the other classes. In particular, the pure_TR sample was assigned to the "Turmeric" class with a degree of 1.0, while there was still some overlap between "Saffron" and "Marigold" classes, which made the assignment of pure_MR and pure_SF samples to the corresponding class a bit more uncertain, although still satisfactory. The classification results for the adulteration mixtures (SFMR_5, SFMR_10, SFMR_15, SFMR_20, SFTR_5, SFTR_10, SFTR_15, and SFTR_20) instead showed an interesting behavior. The threshold value of 0.62 for the "Saffron" class caused the assignment of almost all the adulterated samples to the "Saffron", except for SFTR_15, SFTR_20, and SFMR_20, and none of the other predicted dummy values overcame the thresholds for the other classes. However, it is interesting to note from Table 3 that the degree of belonging to the "Saffron" class tended to decrease as the percentage of the adulterant increased. At the same time, the degree of belonging to the adulterant class tended to increase. Moreover, the calculated degrees of belonging to the "Saffron" class for all the mixtures were lower than the calculated degree obtained for pure_SF sample (although not significantly different for SFTR_5).

Table 3. External validation results (calculated Ys: degrees of belonging) of the test samples projected on the PLS-DA model. The numbers in brackets are the corresponding standard deviations.

Sample Code	y_saffron	y_turmeric	y_marigold
pure_SF	0.78 (0.03)	0.01 (0.02)	0.21 (0.04)
pure_TR	−0.1 (0.2)	1.0 (0.1)	0.1 (0.2)
pure_MR	0.34 (0.04)	0.03 (0.02)	0.63 (0.04)
SFTR_5	0.71 (0.04)	0.06 (0.02)	0.23 (0.04)
SFTR_10	0.66 (0.07)	0.12 (0.04)	0.22 (0.08)
SFTR_15	0.56 (0.06)	0.26 (0.03)	0.19 (0.06)
SFTR_20	0.51 (0.11)	0.32 (0.06)	0.17 (0.11)
SFMR_5	0.69 (0.04)	0.01 (0.02)	0.30 (0.04)
SFMR_10	0.65 (0.03)	0.02 (0.02)	0.33 (0.04)
SFMR_15	0.63 (0.04)	0.02 (0.02)	0.35 (0.04)
SFMR_20	0.59 (0.04)	0.02 (0.02)	0.39 (0.04)

This meant that the PLS-DA model, except for some uncertainties between "Saffron" and "Marigold", was able to discriminate the three studied spices and to detect both an adulteration with at least 15%$_{w/w}$ of turmeric and at least of 20%$_{w/w}$ of marigold in saffron and, at least qualitatively, some contamination in saffron with the other two spices.

2.3. Comparison between PLS-DA and LDA Models

PLS-DA and LDA models returned good results. Indeed, both models had good performances in LOO-CV, and both were able to determine the adulterations of saffron simulated with the test samples listed in Table 1.

In particular, PLS-DA showed some overlap and some uncertainties of classification between "Saffron" and "Marigold" classes. On the other side, the LDA model did not show any class overlap, and it was better than the PLS-DA model in the identification of the pure test samples. Both methods had good ability in the discrimination of the "Turmeric" class from the other two. However, it is important to underline that, even for pure_MR and pure_SF samples, the PLS-DA model was able to correctly classify them.

Regarding the artificial adulteration mixtures, PLS-DA and LDA had similar performances. In fact, for the mixture samples classified by the PLS-DA model, the calculated values of the dummy variables increased with the percentage of adulteration, although they never reached the thresholds, and some doubts persisted about the assignment to the "Saffron" class of such samples. However, the LDA model, by the calculation of the Euclidean distances between the test samples and the class centroids, showed some uncertainties between "Saffron" and "Marigold" classes, but it showed an excellent visual classification in the discriminant plot.

3. Materials and Methods

3.1. Samples

After an accurate commercial search, it was found that certified standards were not available (with the only exception of saffron pistils). Hence, the training-set samples were purchased in food retails; the reliability of these standards was subsequently verified through chemometric tools (see Paragraph 3.5, principal component analysis (PCA), and Hotelling). The spice samples were taken in the same period (April 2017) from several supermarkets, herbalist's shops, and medicinal herb gardens in Emilia Romagna (Italy). It was verified that these samples arrived at the sales centers within a month before the purchase. Twenty-eight samples of saffron, 19 samples of turmeric, and 14 samples of marigold (61 total samples, "calibration samples") were purchased by the laboratory facilities at Coop Italia. Coop Italia is one of the most important supermarket retail chains in Italy. It also has an internal food quality control laboratory in Casalecchio di Reno (Bologna, Italy), where this work was carried out.

Moreover, three samples of pure saffron, turmeric, and marigold ("test samples") were purchased for validation purposes. The pure saffron sample was taken from a supermarket and was a product certified by the SGS certification authority with the certification "Process Control IT MI. 13.P04 STP 013/24". Additionally, no further analyses by means of the ISO 3632-1:2011 [1] were necessary, because the commercially available samples had been controlled before their packaging and sales. The pure turmeric sample was purchased directly from a producer in the Agricultural fair of Santerno (Imola, Bologna, Italy). The pure marigold sample was taken from the Herb Garden of Casola Valsenio (Ravenna, Italy).

3.2. Sample Preparation

All the spice samples were stored in a dark place at low temperature until instrumental analysis. Analyses were carried out within two weeks after sample acquisition.

Regarding the calibration samples, saffron and turmeric powders did not undergo any pre-treatment, while the petals of marigold samples were powdered with Ultra Turrax Tube Drive control (IKA, Staufen im Breisgau, Germany). An aliquot of the sample was placed inside a 20-mL plastic tube with ten stainless steel spheres (5-mm diameter). The tube was subsequently sealed with the appropriate cap and was subjected to stirring at 6000 rpm for 5 min until a medium-grained powder was obtained.

Moreover, the three test samples of saffron, turmeric, and marigold (pure_SF, pure_TR, and pure_MR) were used to prepare eight artificial mixtures (SFTR_5, SFTR_10, SFTR_15, SFTR_20, SFMR_5, SFMR_10, SFMR_15, and SFMR_20) in order to simulate partial adulterations of saffron with the other spices. These samples were obtained by mixing the pure spices in different proportions to cover a wide range of adulteration degrees. In particular, four different percentages (*w/w*) of adulteration were examined: 5%, 10%, 15%, and 20%. These pure samples and mixtures did not undergo the chemometric procedure described later but were used to validate the final partial least squares discriminant analysis (PLS-DA) and linear discriminant analysis (LDA) models.

3.3. Flash Gas-Chromatography (Flash-GC)

All samples from both the calibration set (training set) and the test set were analyzed according to the following procedure.

For GC analysis, an aliquot of (30 ± 3) mg of each powdered sample was placed in a 20-mL glass vial sealed with a magnetic cap. Each sample was prepared in quadruplicate to assess the repeatability and the reproducibility of the method as well as to increase the degrees of freedom of statistical problems. The replicate measurements generated four objects (rows of the dataset-matrix) for each sample. Flash HS-GC-FID analysis was performed by Heracles II instrument at Coop Italia Laboratories.

In particular, this instrument was equipped with two capillary chromatographic columns working in parallel, namely a non-polar column (MXT5: 5% diphenyl, 95% methylpolysiloxane, 10 m length, and 180 μm diameter) and a slightly polar column (MXT1701: 14% cyanopropylphenyl, 86% methylpolysiloxane, 10 m length, and 180 μm diameter) and two flame ionization detectors (FIDs) at the end of each column. GC operation, auto sampling, and chromatographic output were managed by Alphasoft V12.4 software (AlphaMos, Toulouse, France).

The parameters of the chromatographic analysis were chosen after an optimization step to avoid significant problems such as low sensitivity, overcoming of full-scale, and low peaks resolution.

The instrument was also equipped with an auto-sampler HS100 (CTC Analytics AG, Zwingen, Switzerland), which managed up to 96 samples in the same program. The sample vials were placed in a shaker oven at 50 °C and 500 rpm for 20 min. Then, the auto-sampler syringe took 5000 μL of the head-space (by piercing the silicone septum of the vial plug). The sample was injected at 100 $\mu L\ s^{-1}$ (the injector temperature was 200 °C). The carrier gas was molecular hydrogen (H_2) produced by an Alliance High Purity Hydrogen generator (F-dgsi, Évry, France). A solid adsorbing trap Tenax TA 60/80 (Tenax SPA, Verona, Italy) was placed before the chromatographic columns and was maintained at 40 °C and 60 kPa for 65 s while carrier gas was flowing and then heated at 240 °C. This allowed for absorption of the volatile molecules onto the trap and removal of excess air and moisture to concentrate the analytes. Analytes were then introduced into the GC columns by a rotatory valve. The column's initial temperature was 40 °C, which was maintained at such a value for 2 s and then increased by 3 $°C\ s^{-1}$ until reaching 270 °C, then it was kept at this value for 21 s. The total acquisition time was 100 s, and the signal was digitalized every 0.01 s. While a sample was injected, other samples were shaken; the entire process was automated and managed by the instrument in the absence of personnel. As a result, if 96 samples were analyzed in the same program, the overall time needed was not 20 × 96 min but about 180 min.

3.4. Working Datasets

After flash GC analysis, the gas-chromatographic data obtained were tabled into a source matrix (dataset). The dataset rows represented the replicates of the 61 samples (244 rows or objects, 4 replicates for each sample). The labels of the dataset columns corresponded to GC variables, which were the acquisition times derived from the digitalization of the GC signal. Each dataset cell reported the FID signal registered at the corresponding GC time for the relevant object. A further column was the class variable reporting the *a priori* class to which the relevant object belonged. Objects were grouped into classes based on their labeled identity (saffron, turmeric, and marigold).

In particular, two different datasets were created: the "area dataset" and the "intensity dataset". The "area dataset" (AD) variables corresponded to peak areas (56 columns); these variables corresponded to the chromatographic peaks identified by the automatic integration tool of AlphaSoft. The "intensity dataset" (ID) variables were the full chromatograms recorded by Heracles II (20002 columns); cell values were the electric current intensities of FIDs. The signal was digitalized every 0.01s for 100 s (10,001 signals), and the chromatogram of the second column was appended to the one of the first column.

Both datasets were obtained from the chromatograms elaborated by Alphasoft V12.4 software.

3.5. Chemometrics

Before applying any of the chemometric techniques used in this work, all the data were standardized [33]. In particular, two different scaling methods were applied to the datasets: autoscaling for the "area dataset" and centering for the "intensity dataset".

Two models for the determination of partial or total adulteration of saffron with turmeric and marigold were created and evaluated, LDA [30] and PLS-DA [29,32]. In particular, the LDA model was computed for AD, while the PLS-DA model was computed for ID.

For each dataset, the following chemometric procedure was carried out in parallel. First, for each class, the elimination of the outliers was performed by PCA and Hotelling analysis [34] at a confidence level of 95%, as already described in a previous work [25].

Then, the refined datasets including only statistically significant samples were subsequently subjected to LDA and PLS-DA. Both chemometric models were then validated by internal cross-validation (CV) [29,30] and by projecting the eleven test samples (not used for model creation) [29]. CV is a statistical technique that allowed evaluating the prediction ability of a model (i.e., the ability to determine the values of the response variables from the predictors for the test samples). CV performed the following steps iteratively: exclude some samples (randomly selected) from the training set, build the model without the excluded samples, and classify the excluded samples with this model. During this procedure, each sample of the training set was used as a test sample at least one time. However, the results of CV were different for LDA and PLS-DA.

LDA computed a model characterized by the definition of new variables starting from the original variables (in the case of AD, chromatographic peak areas) as well as in PCA. However, LDA, unlike PCA, defined linear discriminant functions (LDs) rather than principal components (PCs) that were more effective in separating the examined classes [29]. Such a model could classify unknown samples by projecting them in the LDs space. An unknown sample was always assigned to the class for which the calculated posterior probability [35] was higher; however, the distance of objects from the classes needed to be taken into account in order to finely evaluate the degree of membership to a class.

For LDA, the CV output was represented by the confusion matrix. In this matrix, the lines represented the "a priori" classes, and the columns represented the calculated "a posteriori" classes, to which CV reassigned the samples. The ideal situation was a diagonal matrix (i.e., the matrix in which the entries outside the main diagonal were all zero) because it was the situation in which all of the samples were correctly assigned to the corresponding "a priori" classes. Subsequently, starting from the confusion matrix, it was possible to compute the "non-error-rate" (NER) as the ratio between the objects correctly classified and the total number of objects, which represented the ability of the model to correctly recognize its objects.

PLS-DA [32] is instead a regression method in which the predictor variables (X-matrix) were the experimental ones (in the ID case, the full chromatograms), while the responses (Y-matrix) were the so-called "dummy variables". These dummy variables were the degrees of belonging to the examined classes (in this work, saffron, turmeric, and marigold) and assumed the values for calibration objects to be 0 and 1 (where 1 represented the certainty of belonging to the considered class, while 0 represented the certainty of not belonging to the considered class). The projection of an external sample onto a PLS-DA model returned a set of values for the dummy variables that could be considered as "degrees of belonging" to each class.

CV results for a PLS-DA model were represented by the calculated values of dummy variables for each sample, which meant the predicted degree of belonging of each sample to each class. These values could be used to calculate a threshold value for each class that optimized both sensitivity and specificity for the classification. The procedure for computing such threshold values is described by Ballabio and Consonni (2013) [32]. The projected samples of the test set could then be assigned to a class if their corresponding calculated value of the dummy variable overcame the threshold.

Outliers elimination was carried out by the software The Unscrambler V10.4 (Camo, Oslo, Norway), while LDA and PLS-DA were carried out (with relative CV and projections) by the software R V3.4.3 (R Core Team, Vienna, Austria) with the packages "MASS" [31] and "pls" [35].

4. Conclusions

The achieved results illustrate that the herein proposed, non-targeted strategy based on the combined application of chemometrics with Heracles II flash HS-GC-FID may provide a rapid and low-cost screening method for the authentication of saffron.

The samples were analyzed without any preparation or after a rapid grinding operation, allowing us to avoid expensive pre-treatments and any contamination before analysis by gas-chromatography. Furthermore, once the sample is put into the auto-sampler of the instrument, this instrumental analysis is entirely automated and requires a short analysis time (overall, less than 20 min for a single sample and a couple of minutes *per* sample for 96 samples simultaneously put in the auto-sampler).

Finally, with chemometrics, it was possible to use the GC data both as they are produced by the instrument (chromatograms) and by integrating the chromatographic peaks to build classification models (PLS-DA and LDA). These models had good calibration ability, evaluated by cross-validation (CV) and, most of all, good prediction ability, evaluated by projecting external test samples that simulated adulterations of saffron with turmeric and marigold. Moreover, for adulterant additions below 33%$_{w/w}$, the official UV-VIS spectrophotometry method was not able to detect adulteration [8]. On the contrary, Heracles II combined with chemometrics allowed us to go far below this limit; a PLS-DA model able to detect down to 15 ÷ 20%$_{w/w}$ of adulteration was validated. Moreover, a discriminant plot obtained through LDA showed significant differences between pure samples and adulterated samples down to 5 ÷ 10%$_{w/w}$.

Another important characteristic of the chemometric approach is that it does not require the identification of the volatile compounds to create a model able to find an adulterated saffron sample. The use of the entire chromatograms ensures that all the possible markers for turmeric or marigold adulteration are taken into account in the model construction.

Author Contributions: Conceptualization, F.G. and D.M.; Methodology, F.G. and P.M.; Software, A.Z. and P.M.; Validation, P.M. and A.Z.; Formal Analysis, P.M. and F.G.; Investigation, P.M.; Resources, F.G. and M.L.; Data Curation, A.Z., P.M. and D.M.; Writing—Original Draft Preparation, P.M.; Writing—Review & Editing, A.Z., M.L. and D.M.; Visualization, F.G.; Supervision, D.M.; Project Administration, F.G.

Funding: This research received no external funding.

Acknowledgments: The authors thank Coop Italia for providing the Heracles II flash HS-GC-FID instrument and Paolo De Giorgi for the assistance in the experimental activity.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Organization for Standardization. *ISO 3632-1. Spices—Saffron (Crocus sativus L.)*; ISO: Geneva, Switzerland, 2011.
2. Moore, J.C.; Spink, J.; Lipp, M. Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *J. Food Sci.* **2012**, *77*, R118–R126. [CrossRef] [PubMed]
3. Nazari, S.H.; Keifi, N. Saffron and various fraud manners in its production and trades. *Acta Hortic.* **2007**, *739*, 411–416. [CrossRef]
4. Johnson, R. *Food Fraud and "Economically Motivated Adulteration" of Food and Food Ingredients*; Congressional Research Service Report; University of North Texas Libraries: Denton, TX, USA, 2014; pp. 1–40.
5. Bosmali, I.; Ordoudi, S.A.; Tsimidou, M.Z.; Madesis, P. Greek PDO saffron authentication studies using species specific molecular markers. *Food Res. Int.* **2017**, *100*, 899–907. [CrossRef] [PubMed]
6. Galvin-King, P.; Haughey, S.A.; Elliott, C.T. Herb and spice fraud; the drivers, challenges and detection. *Food Control* **2018**, *88*, 85–97. [CrossRef]
7. PwC & SSAFE. Food Fraud Vulnerability Assessment. 2016. Available online: http://www.pwc.com/gx/en/services/food-supply-integrity-services/assets/pwc-food-fraud-vulnerability-assessment-and-mitigation-november.pdf (accessed on 1 July 2019).
8. Sabatino, L.; Scordino, M.; Gargano, M.; Belligno, A.; Traulo, P.; Gagliano, G. HPLC/PDA/ESI-MS evaluation of saffron (*Crocus sativus* L.) adulteration. *Nat. Prod. Commun.* **2011**, *6*, 1873–1876. [CrossRef] [PubMed]
9. Kiani, S.; Minaei, S.; Ghasemi-Varnamkhasti, M. Instrumental approaches and innovative systems for saffron quality assessment. *J. Food Eng.* **2018**, *216*, 1–10. [CrossRef]
10. Petrakis, E.A.; Cagliani, L.R.; Polissiou, M.G.; Consonni, R. Evaluation of saffron (*Crocus sativus* L.) adulteration with plant adulterants by1H NMR metabolite fingerprinting. *Food Chem.* **2015**, *173*, 890–896. [CrossRef]
11. Petrakis, E.A.; Polissiou, M.G. Assessing saffron (*Crocus sativus* L.) adulteration with plant-derived adulterants by diffuse reflectance infrared Fourier transform spectroscopy coupled with chemometrics. *Talanta* **2017**, *162*, 558–566. [CrossRef]
12. Zalacain, A.; Ordoudi, S.A.; Díaz-Plaza, E.M.; Carmona, M.; Blázquez, I.; Tsimidou, M.Z.; Alonso, G.L. Near-infrared spectroscopy in saffron quality control: Determination of chemical composition and geographical origin. *J. Agric. Food Chem.* **2005**, *53*, 9337–9341. [CrossRef]
13. Ordoudi, S.A.; De Los Mozos Pascual, M.; Tsimidou, M.Z. On the quality control of traded saffron by means of transmission Fourier-transform mid-infrared (FT-MIR) spectroscopy and chemometrics. *Food Chem.* **2014**, *150*, 414–421. [CrossRef]
14. Rubert, J.; Lacina, O.; Zachariasova, M.; Hajslova, J. Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food Chem.* **2016**, *204*, 201–209. [CrossRef] [PubMed]
15. Nenadis, N.; Heenan, S.; Tsimidou, M.Z.; Van Ruth, S. Applicability of PTR-MS in the quality control of saffron. *Food Chem.* **2016**, *196*, 961–967. [CrossRef] [PubMed]
16. Aliakbarzadeh, G.; Parastar, H.; Sereshti, H. Classification of gas chromatographic fingerprints of saffron using partial least squares discriminant analysis together with different variable selection methods. *Chemom. Intell. Lab. Syst.* **2016**, *158*, 165–173. [CrossRef]
17. Torelli, A.; Marieschi, M.; Bruni, R. Authentication of saffron (*Crocus sativus* L.) in different processed, retail products by means of SCAR markers. *Food Control* **2014**, *36*, 126–131. [CrossRef]
18. Gismondi, A.; Fanali, F.; Martínez Labarga, J.M.; Caiola, M.G.; Canini, A. *Crocus sativus* L. Genomics and different DNA barcode applications. *Plant Syst. Evol.* **2013**, *299*, 1859–1863. [CrossRef]
19. Babaei, S.; Talebi, M.; Bahar, M. Developing an SCAR and ITS reliable multiplex PCR-based assay forsafflower adulterant detection in saffron samples. *Food Control* **2014**, *35*, 323–328. [CrossRef]
20. Danezis, G.P.; Tsagkaris, A.S.; Camin, F.; Brusic, V.; Georgiou, C.A. Food authentication: Techniques, trends & emerging approaches. *TrAC Trends Anal. Chem.* **2016**, *85*, 123–132.
21. Esslinger, S.; Riedl, J.; Fauhl-Hassek, C. Potential and limitations of non-targeted fingerprinting for authentication of food in official control. *Food Res. Int.* **2014**, *60*, 189–204. [CrossRef]

22. Matsushita, T.; Zhao, J.J.; Igura, N.; Shimoda, M. Authentication of commercial spices based on the similarities between gas chromatographic fingerprints. *J. Sci. Food Agric.* **2018**, *98*, 2989–3000. [CrossRef]
23. Heidarbeigi, K.; Mohtasebi, S.S.; Foroughirad, A.; Ghasemi-Varnamkhasti, M.; Rafiee, S.; Rezaei, K. Detection of adulteration in saffron samples using electronic nose. *Int. J. Food Prop.* **2015**, *18*, 1391–1401. [CrossRef]
24. Carmona, M.; Zalacain, A.; Salinas, M.R.; Alonso, G.L. A new approach to saffron aroma. *Crit. Rev. Food Sci. Nutr.* **2007**, *47*, 145–159. [CrossRef] [PubMed]
25. Melucci, D.; Bendini, A.; Tesini, F.; Barbieri, S.; Zappi, A.; Vichi, S.; Conte, L.; Gallina Toschi, T. Rapid direct analysis to discriminate geographic origin of extra virgin olive oils by flash gas chromatography electronic nose and chemometrics. *Food Chem.* **2016**, *204*, 263–273. [CrossRef] [PubMed]
26. Wiśniewska, P.; Śliwińska, M.; Namieśnik, J.; Wardencki, W.; Dymerski, T. The Verification of the Usefulness of Electronic Nose Based on Ultra-Fast Gas Chromatography and Four Different Chemometric Methods for Rapid Analysis of Spirit Beverages. *J. Anal. Methods Chem.* **2016**, *2016*. [CrossRef] [PubMed]
27. Wojtasik-Kalinowska, I.; Guzek, D.; Górska-Horczyczak, E.; Głabska, D.; Brodowska, M.; Sun, D.W.; Wierzbicka, A. Volatile compounds and fatty acids profile in Longissimus dorsi muscle from pigs fed with feed containing bioactive components. *LWT Food Sci. Technol.* **2016**, *67*, 112–117. [CrossRef]
28. Górska-Horczyczak, E.; Wojtasik-Kalinowska, I.; Guzek, D.; Sun, D.W.; Wierzbicka, A. Differentiation of chill-stored and frozen pork necks using electronic nose with ultra-fast gas chromatography. *J. Food Process Eng.* **2017**, *40*, e12540. [CrossRef]
29. Berrueta, L.A.; Alonso-Salces, R.M.; Héberger, K. Supervised pattern recognition in food analysis. *J. Chromatogr. A* **2007**, *1158*, 196–214. [CrossRef]
30. Bevilacqua, M.; Nescatelli, R.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. Chemometric classification techniques as a tool for solving problems in analytical chemistry. *J. AOAC Int.* **2014**, *97*, 19–28. [CrossRef] [PubMed]
31. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; Volume 53, ISBN 0387954570.
32. Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal. Methods* **2013**, *5*, 3790–3798. [CrossRef]
33. Van den Berg, R.A.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genom.* **2006**, *7*, 142. [CrossRef]
34. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2002; ISBN 0387954422.
35. Mevik, B.H.; Wehrens, R.; Liland, K.H. pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0. *J. Stat. Softw.* **2015**. Available online: https://www.researchgate.net/deref/http%3A%2F%2Fmevik.net%2Fwork%2Fsoftware%2Fpls.html (accessed on 16 July 2019).

Sample Availability: Samples are available from the authors.

molecules

Article

Untargeted Metabolomic Profile for the Detection of Prostate Carcinoma—Preliminary Results from PARAFAC2 and PLS–DA Models

Eleonora Amante [1,2], Alberto Salomone [1,2], Eugenio Alladio [1,2], Marco Vincenti [1,2,*], Francesco Porpiglia [3] and Rasmus Bro [4]

[1] Dipartimento di Chimica, Università degli Studi di Torino, Via P. Giuria 7, 10125 Torino, Italy
[2] Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Regione Gonzole 10/1, 10043 Orbassano, Italy
[3] Division of Urology, San Luigi Gonzaga Hospital and University of Torino, 10043 Orbassano, Italy
[4] Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg, Denmark
* Correspondence: marco.vincenti@unito.it

Received: 24 July 2019; Accepted: 20 August 2019; Published: 22 August 2019

Abstract: Prostate-specific antigen (PSA) is the main biomarker for the screening of prostate cancer (PCa), which has a high sensibility (higher than 80%) that is negatively offset by its poor specificity (only 30%, with the European cut-off of 4 ng/mL). This generates a large number of useless biopsies, involving both risks for the patients and costs for the national healthcare systems. Consequently, efforts were recently made to discover new biomarkers useful for PCa screening, including our proposal of interpreting a multi-parametric urinary steroidal profile with multivariate statistics. This approach has been expanded to investigate new alleged biomarkers by the application of untargeted urinary metabolomics. Urine samples from 91 patients (43 affected by PCa; 48 by benign hyperplasia) were deconjugated, extracted in both basic and acidic conditions, derivatized with different reagents, and analyzed with different gas chromatographic columns. Three-dimensional data were obtained from full-scan electron impact mass spectra. The PARADISe software, coupled with NIST libraries, was employed for the computation of PARAFAC2 models, the extraction of the significative components (alleged biomarkers), and the generation of a semiquantitative dataset. After variables selection, a partial least squares–discriminant analysis classification model was built, yielding promising performances. The selected biomarkers need further validation, possibly involving, yet again, a targeted approach.

Keywords: untargeted metabolomics; PARAFAC2; alignment; gas chromatography–mass spectrometry (GC–MS); prostate carcinoma

1. Introduction

Prostate cancer (PCa) is the most common non-skin cancer in men [1,2] and the second most frequently diagnosed malignancy in males worldwide [3]. The first biomarker for PCa detection was prostatic acid phosphatase (PAP), which was introduced in the 1930s [1]. In the 1980s, PAP was replaced by prostate-specific antigen (PSA) [1,4], a secreted protein encoded by a prostate-specific gene and member of the tissue kallikrein family [1], which is produced almost exclusively in the prostate [5,6]. After the introduction of PSA, more men were diagnosed with PCa, with the majority having the early-stage, clinically indolent form of the disease. However, a large number of patients affected by a benign pathology, such as inflammation or hyperplasia, exhibited abnormal PSA values, which lead to the execution of useless biopsies and demonstrate the low specificity of this biomarker [1,6]. This phenomenon was generally designated as "overdiagnosis" or "overtreatment" [1,4,7,8].

Considerable effort has been devoted to improving the PSA-test performance, including the introduction of PSA density, PSA velocity (and doubling time), the dosage of free or complexed PSA, and the quantitation of its isoforms [1,2]. A combination of these parameters yields the Prostate Health Index (PHI) [3].

Meanwhile, intensive research has been devoted to the search for different biomarkers, mainly by applying omics methods (e.g., genomics, proteomics, transcriptomics, and metabolomics) [1], and several authors have reviewed the emerging biomarkers, among which the most prominent are the urinary prostate cancer antigen 3 (PCA3) [1,3,5] and transmembrane protease, serine 2 (TMPRSS2-ERG) (sometimes combined together) [1,3,5]. Alpha-methylacyl-CoA Racemase (AMACR) demonstrated high sensitivities and specificities on prostate biopsy, but it is not suitable for non-invasive detection in urine [1,5]. Increased diagnostic performances were obtained by the serum dosage of human kallikrein-related peptidase 2 (KLK2) in combination with total and free PSA [5]. An evolution of the application of kallikreins consists in a blood measurement of the four existing isoforms which, combined with clinical information, allows the probability calculation of PCa incidence [3]. Significantly increased levels of prostasomes were found in blood samples from patients with PCa [9], while elevated levels of urinary sarcosine were found to be associated with aggressive forms of prostate cancer [1].

Studies conducted in the 1970s and 1980s highlighted the correlation between increased urinary excretion of polyamines (i.e., spermine, spermidine, and putrescine) and several types of cancer [10,11]. However, anomalous oxidative degradation reactions of these polyamines resulted in low concentrations of these biomarkers in approximately 20% of the patients, leading to false-negative prediction and consequently limiting their application as diagnostic biomarkers [10].

The correlation between altered steroidal biosynthesis and PCa is well known [12–14]. For this reason, in a previous study, we carried out a targeted analysis of urine samples, addressed to a large panel of androgens, including testosterone and its principal phase I metabolites. The multivariate statistical interpretation of these steroid profiles produced satisfactory results in terms of sensitivity, specificity, and area under the curve (AUC) [15].

In this study, the search for new urinary biomarkers was undertaken by using untargeted methods. In perspective, emerging biomarkers could possibly be combined with the most discriminating steroid biomarkers to improve their screening performances further, without altering the inherent simplicity of the instrumental procedure. In fact, the ideal biomarker should be cheap to determine, non-invasive, easily accessible, and quickly quantifiable [1,2]. Taking into account the abovementioned considerations, gas chromatography–electron impact mass spectrometry (GC–EIMS) would give a more suitable solution than the other commonly used analytical techniques to provide a three-dimensional pattern for untargeted analysis. Urine was chosen as the election matrix, as it is easily available in large volumes and involves non-invasive sampling.

2. Materials and Methods

2.1. Chemicals and Reagents

Tert-butyl methyl ether (TBME), ethyl acetate, dithioerythritol, ammonium iodide (NH_4I), *N*-Methyl-*N*-(trimethylsilyl)trifluoroacetamide (TMSTFA), and trifluoroacetic anhydride (TFAA) were provided by Sigma-Aldrich (Milan, Italy). β-glucuronidase from *Escherichia coli* was purchased from Roche Life Science (Indianapolis, IN, USA). Ultra-pure water was obtained using a Milli-Q® UF-Plus apparatus (Millipore, Bedford, MA, USA).

2.2. Samples Collection

The subjects involved in this study were recruited in the ambulatory of the Department of Urology at the San Luigi Hospital of Orbassano (TO, Italy), after approval of the protocol by the reference Ethical

Committee (protocol number 0019267). A total of 91 subjects were enrolled, including 43 affected by prostate carcinoma (PCa, confirmed by a positive biopsy) and 48 diagnosed with benign prostatic hyperplasia (BPH, with a PSA lower than the European cut-off of 4 ng/mL or with a PSA above the threshold but a negative biopsy result). In a previous study, the progressive modification of the urinary steroidal profile with age was investigated [16]. From this study, we decided to enroll only individuals older than 60 years, when the bias effect due to aging became negligible [16]. Moreover, since ethnicity represents another important bias factor, only Caucasian individuals were recruited. Finally, diabetes, other carcinoma, metabolic diseases, and therapies suspected to alter the urinary steroid profile (such as steroid therapy) were considered as exclusion criteria.

Body mass index (BMI), alcohol consumption, medical therapy, digital rectal examination, PSA value, and biopsy Gleason Score (GS) were recorded. In detail, the group's mean age and standard deviation was 70 ± 10 years for BPH and 70 ± 8 years for PCa. BMI was within the range of normality for all individuals (between 18.5 and 25), and PSA was 3.8 ± 2.3 ng/mL for BPH and 11.0 ± 9.5 ng/mL for PCa. The PCa class was distributed as low risk (GS = 3 + 3, 15 patients), middle risk (GS = 3 + 4 and 4 + 3, 21 patients), and high risk (GS = 4 + 4 and 4 + 5, seven patients).

2.3. Sample Treatment and GC–MS Analysis

Firstly, the protein components of the urinary samples were precipitated by centrifugation at 4000 rpm for 5 min. Two aliquots (A and B) of 5 mL each were taken from each sample. The urine pH was adjusted between 6.8 and 7.4 by adding 2 mL of phosphate buffer and a few drops of NaOH 1M or HCl 1M whenever necessary. Enzymatic hydrolysis of the glucuronide metabolites was conducted with 100 µL of β-glucuronidase from *Escherichia coli* (equivalent to 83 enzymatic units) by heating it in the oven for 1 h. After cooling to room temperature, the two aliquots were subjected to different liquid–liquid extraction (LLE) with 5 mL of TBME each, at basic (pH ≥ 10) and acid (pH ≤ 1) conditions, respectively, obtained by the introduction of some drops of NaOH 1M and HCl 1M. Both aliquots were dried under a gentle nitrogen stream at room temperature. The dried aliquot A was derivatized using 50 µL TFAA at 65 °C for 1 h. Then, the solvent was dried and the residue was dissolved in 50 µL TBME and injected into the GC–MS. The chromatographic separation was achieved with a J&W Scientific HP-5, 17 m × 0.2 mm (i.d.) × 0.33 µm (f.t.) capillary column. The oven temperature was programmed as follows: The starting temperature of 90 °C was held for 1 min. Then, the temperature of 180 °C was reached with a rate of 30 °C/min and held for 7 min. A final heating rate of 15 °C/min was applied until the temperature of 325 °C was reached (held for 3 min). The chromatographic run lasted 22.20 min.

Aliquot B was derivatized using 50 µL of TMSTFA/NH_4I/dithioerythritol (1.000:2:4 *v*/*w*/*w*), at 70 °C for 30 min and then injected into a GC–MS equipped with a J&W Scientific HP-1, 17 m × 0.2 mm (i.d.) × 0.11 µm (f.t.) capillary column. The oven temperature was programmed to heat up from 120 to 177 °C at a rate of 70 °C/min, and from 177 to 236 °C at a rate of 5 °C/min. A final heating rate of 30 °C/min was applied until the temperature of 315 °C was reached. The chromatographic run lasted 18.25 min. Both the runs were performed in full-scan mode, in the interval 40–650 *m*/*z* at a scan rate of 2.28 scans/s.

Because the samples were analyzed in five analytical sections performed on different days, it was important to monitor the occurrence of a data structure due to the different analytical sections. The exploratory unsupervised data analysis can serve to this scope, and principal component analysis (PCA) was employed. No clustering or trend related to the day of the analysis was detected.

2.4. Statistical Analysis

The main steps of the statistical analysis are reported in Figure 1.

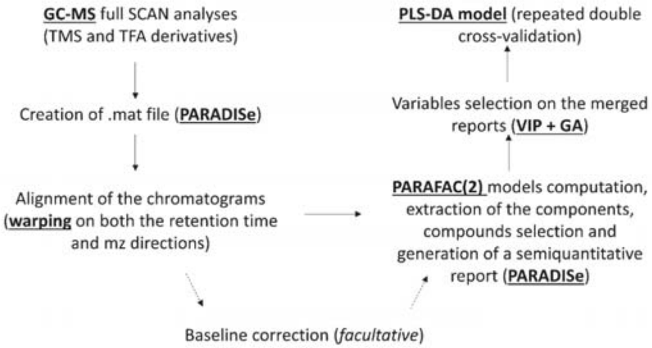

GC-MS full SCAN analyses (TMS and TFA derivatives)
Creation of .mat file (PARADISe)
Alignment of the chromatograms (warping on both the retention time and mz directions)
Baseline correction (*facultative*)
PARAFAC(2) models computation, extraction of the components, compounds selection and generation of a semiquantitative report (PARADISe)
Variables selection on the merged reports (VIP + GA)
PLS-DA model (repeated double cross-validation)

Figure 1. Statistical analysis workflow.

2.5. Pre-Treatment of the Raw Data

The .AIA files of the chromatographic runs were downloaded using the software ChemStation®.

The PARADISe version 3 software was employed to convert the files in a form suitable for MATLAB (extension .mat). The alignment procedure, both propaedeutic and mandatory for the following steps of data analysis, was executed over the three-way (samples × retention time × *m*/*z*) array of size 91 × 3099 × 612 (over 172 million data) and 91 × 1640 × 652 (over 97 million data) for the trifluoroacetyl (TFA) and trimethylsilyl (TMS) derivatives, respectively. The correlation optimized warping (COW) was performed along the retention time and the *m*/*z* dimensions [17]. The two matrices were segmented along the retention time dimension to improve the performances of the COW algorithm, and for each slice the computation was iterated until a visually satisfying result was obtained. Lastly, to improve the visualization of the data, the baseline was subtracted. It is important to highlight that the latter computational step only served to improve the data visualization by the operator, because PARAFAC2 is able to recognize the baseline and the noise components, allowing their automatic exclusion [18–21].

2.6. PARAFAC2 Models Computation and Molecular Identification

The aligned dataset was analyzed in the PARADISe software, to proceed with the computation of PARAFAC2 models; the operating procedure consists in the manual identification of intervals along the chromatogram (with each interval ideally containing approximately one peak). The PARAFAC2 models were built introducing the non-negativity constraint and performing 10,000 iterations for interval [18–21].

Within the software, the operator can label the components as (i) baseline, (ii) noise, or (iii) compounds. All the mass spectra of the components belonging to the third category are automatically compared with the NIST database. A report is produced, including the relative concentrations of the detected compounds (assuming a uniform response factor of 1), and the *n* (number subjectively chosen by the user) most likely identifications for each compound. Finally, the relative concentrations were normalized using the urinary creatinine values.

2.7. Classification Models

The dataset composed by the relative concentrations of the detected metabolites for each sample was used to perform partial least squares–discriminant analysis (PLS–DA) [22], classifying the samples into having prostate carcinoma or not. Firstly, the dataset was log10 transformed (with the aim of achieving a more even distribution of each of the variables) and autoscaled. Then, the variable importance in projection (VIP) method (using a threshold of 1) [23] and genetic algorithms (GAs) [24] were run to select the most relevant variables. The reduced dataset was finally used to build the

PLS–DA classification model. The model was then validated using the repeated double cross-validation (dCV) approach [25]. The PLS_Toolbox version 8.5 (Eigenvector Research, Inc., Manson, WA, USA) was used to perform this part of the analysis [26].

3. Results and Discussion

The preliminary PARAFAC2 model extracted a total of 329 relevant compounds (184 from the chromatographic run after TMS derivatization and 145 after TFA derivatization). Of these, 89 were selected using the VIP algorithm, and a further 58 substances were discarded by one cycle of GAs. The final dataset, consisting in a 91 × 32 (subjects × variables) matrix, was employed to build a PLS–DA classification model. Due to the heterogeneity of the patients enrolled (in terms of pathology staging, prostatic volume, and PSA values), the model was validated using repeated double cross-validation (30 repetitions were performed) [25] instead of the standard external validation, in which the use of a limited and heterogeneous population may result in significant bias. The plot reporting the Y-value predicted in cross-validation (CV) in one of the several classification models produced during the repeated double cross-validation process is shown in Figure 2A. The corresponding receiver operating characteristic (ROC) curve is depicted in Figure 2B. The high values of the area under the curve (AUC) for both the estimated and cross-validated ROCs are an indicator of high performances and robustness of the model. In detail, using a discriminating Y-value of 0.5, the model provides 92.5 ± 2.2% sensitivity and 88.7 ± 3.9% specificity for the cancer-affected population. On average, misclassification occurred on about 3 ± 2 patients affected by carcinoma out of 43 and 5 ± 2 patients with hyperplasia out of 48.

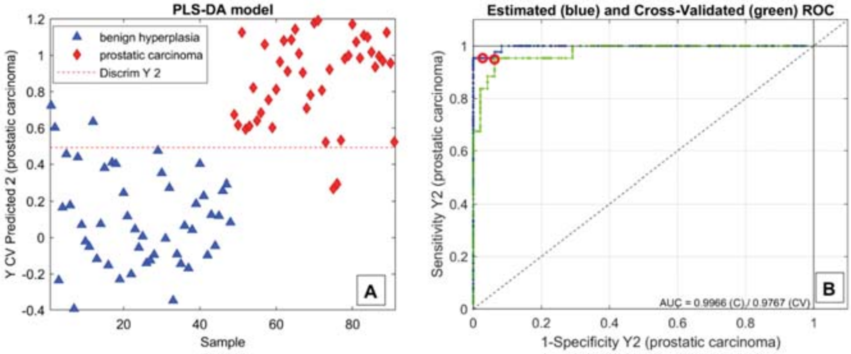


Figure 2. Y-value predicted in cross-validation (CV) (**A**) and receiver operating characteristic (ROC) curves (**B**) of one of the several classification models built during the repeated double cross-validation procedure.

Of the 32 compounds selected by the dedicated algorithms to build the model, 17 were not found in the available NIST libraries, while for seven other compounds, the identification provided by automatic spectral matching was deemed incorrect. On the other hand, manual mass spectra interpretation was made difficult by the structural similarity of many candidate biomarkers, as well as the effect of the derivatizations, that introduced functional groups (e.g., $-Si(CH_3)_3$) yielding prevalent fragment ions in the spectrum. The mass spectra of the 32 metabolites are provided as Supplementary Materials (Supplementary Figure S1). Table 1 reports the eight identified compounds, accompanied by their Human Metabolome Database (HMDB) and Kyoto Encyclopedia of Genes and Genomes (KEGG) identification numbers, when available. Since the PARADISe output provides only a rough semiquantitative report based on the total ion current (TIC) without any external calibration, the real physiological concentration of each metabolite could not be evaluated. However, these absolute TIC values can be evaluated in relative terms to provide an averaged qualitative comparison between the two populations for all the analytes. The overexpression and underexpression of these metabolites allegedly linked to the occurrence of PCa are reported in Table 1.

Table 1. List of the 32 selected metabolites. The kind of derivatization, retention time, and expression in prostate cancer (PCa)-affected individuals are reported. Moreover, metabolites with a putative identification are accompanied by the match score with NIST library and the relative identification (ID) number in the Human Metabolome Database (HMDB) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. The mass spectra of the 32 metabolites are reported in Supplementary Materials—Figure S1.

		Compound	**Derivatization**	**Retention Time (min)**	**Match with NIST**	**HMDB ID**	**KEGG ID**	**Expression in PCa Patients**
TMS derivatives	1	5-Hydroxyindoleacetic acid	TMS	5.26	893	HMDB0000763	C05635	overexpression
	2	Unknown 1	TMS	5.86	-	-	-	overexpression
	3	Unknown 2	TMS	7.44	-	-	-	underexpression
	4	Androsterone	TMS	8.14	912	HMDB0000031	C00523	overexpression
	5	16-Hydroxydehydroisoandrosterone	TMS	9.23	888	HMDB0000352	C05139	overexpression
	6	Unknown 3	TMS	9.84	-	-	-	comparable
	7	Unknown 4	TMS	10.31	-	-	-	underexpression
	8	Unknown 5	TMS	10.61	-	-	-	underexpression
	9	Unknown 6	TMS	11.29	-	-	-	underexpression
	10	Unknown 7	TMS	11.32	-	-	-	comparable
	11	Enterodiol	TMS	12.19	826	HMDB0005056	C18166	underexpression
	12	5β-pregnanediol	TMS	12.53	853	HMDB0005943	Not available	underexpression
	13	Unknown 8	TMS	13.6	-	-	-	overexpression
	14	Unknown 9	TMS	13.67	-	-	-	comparable
	15	Pregnanetriol	TMS	13.73	904	HMDB0006070	Not available	underexpression
	16	Unknown 10	TMS	14.03	-	-	-	underexpression
	17	Unknown 11	TMS	14.50	-	-	-	underexpression
	18	Unknown 12	TMS	14.53	-	-	-	underexpression
	19	Unknown 13	TMS	14.6	-	-	-	overexpression
	20	Unknown 14	TMS	14.66	-	-	-	underexpression
	21	Unknown 15	TMS	15.04	-	-	-	underexpression
TFA derivatives	22	Unknown 16	TFA	1.63	-	-	-	underexpression
	23	Unknown 17	TFA	1.71	-	-	-	comparable
	24	Vanillyl alcohol	TFA	3.37	860	HMDB0032012	C06317	overexpression
	25	Unknown 18	TFA	4.97	-	-	-	comparable
	26	Unknown 19	TFA	5.71	-	-	-	underexpression
	27	Unknown 20	TFA	3.32	-	-	-	underexpression
	28	Epiandrosterone	TFA	15.61	925	HMDB0000365	C07635	comparable
	29	Unknown 21	TFA	16.32	-	-	-	underexpression
	30	Unknown 22	TFA	17.87	-	-	-	underexpression
	31	Unknown 23	TFA	18.11	-	-	-	overexpression
	32	Unknown 24	TFA	18.24	-	-	-	underexpression

It is interesting to note that among the eight identified compounds, five (63%) are involved in steroidal biosynthesis, confirming their potential in the detection and diagnosis of PCa. Similarly, Choi et al. found elevated levels of 16-hydroxy-dehydroepiandrosterone, epiandrosterone, etiocholanolone, and pregnanetriol in patients diagnosed with papillary thyroid carcinoma [27]. The first steroid appears to also be overexpressed in the present case for patients with PCa, but pregnanetriol was underexpressed in the same patients and epiandrosterone was found in comparable concentrations in the two populations. Dehydroepiandrosterone is involved in the expression of insulin-like growth factor 1, whose dysregulation is implicated in certain colon, liver, prostate, and breast cancers [28]. This observation may justify the inclusion of 16-hydroxy-dehydroepiandrosterone among the potential biomarkers for PCa. Pregnanetriol, together with 5 β-pregnanediol, is also known to be dysregulated in adrenal syndromes, such as adrenal tumors or Cushing's syndrome [29,30]. Increased androsterone levels were found in a cohort of PCa-affected individuals within a multivariate investigation of the urinary steroidal profile, and the present findings are in accordance with our previous study [16]. Other steroids that proved useful to discriminate PCa from BPH [27] were possibly overlooked in the present untargeted selection because of their low concentration in urine.

The overexpression of serotonin and its biomarkers (among which, 5-hydroxyindoleacetic acid) represents a potential urinary biomarker for neuroblastic and carcinoid tumors [31]. While there is no evidence in the literature of an association between 5-hydroxyindoleacetic acid and PCa, the present data suggest such a hypothesis, as its overexpression is clearly evident in the PCa-affected population considered.

Phytoestrogens are a class of substances accredited to prevent the onset of cancer [32,33]. In accordance with this hypothesis, enterodiol is underexpressed in the present PCa population.

Among the 32 selected biomarkers, different contributions to the overall discrimination achieved by the PLS–DA model (Figure 1) were expected. A rough estimation of the relative importance of these biomarkers is expressed by their selectivity ratios [34], reported in Figure 3. Nine biomarkers exhibit selectivity ratios higher that 0.1, while, for five others, values between 0.07 and 0.1 were found. Interestingly, out of the nine biomarkers with the highest selectivity ratio, eight are underexpressed in the PCa patients, apparently suggesting them as protective substances. The expression of the 14 metabolites is represented in the form of boxplots in Figure 4. Tentative PLS–DA models were built with only these 9 and 14 biomarkers, but their overall efficiency significantly dropped with respect to the model of 32 biomarkers, demonstrating that the relative contribution of the remaining biomarkers is not negligible. In particular, the specificity index was considerably reduced in the models of 9 biomarkers and 14 biomarkers, while the sensitivity score remained relatively high.

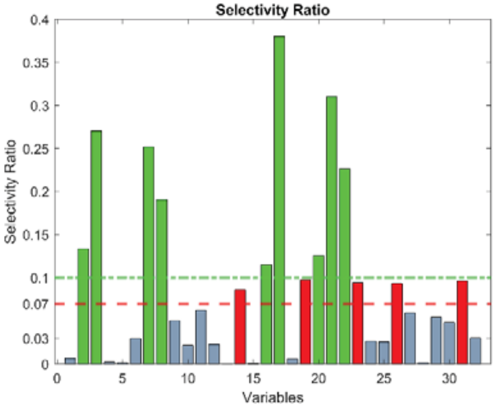


Figure 3. Selectivity ratio of the 32 selected features. The variables above the threshold of 0.1 are reported in green, and the ones between the thresholds 0.07–0.1 are reported in red.

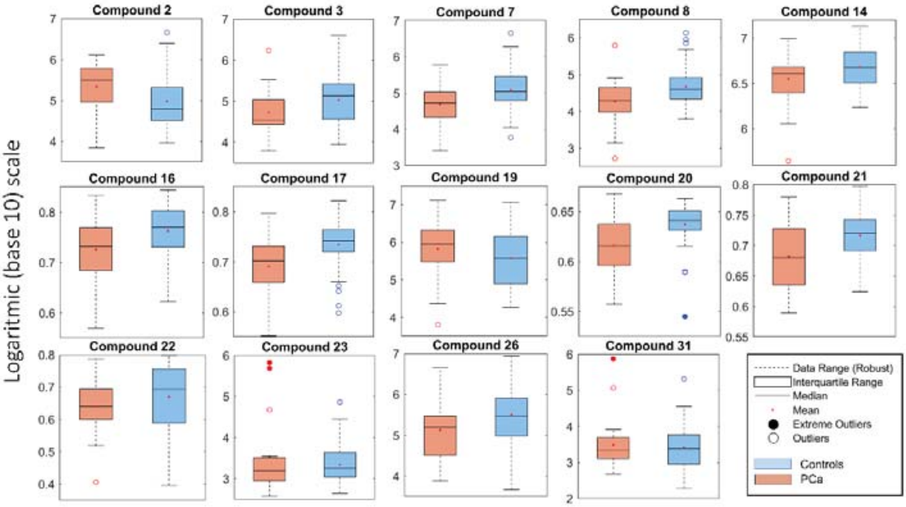


Figure 4. Boxplot, in logarithmic (base-10) scale, of the 14 compounds above the selectivity ratio threshold of 0.07 (see Figure 3).

Further testing was also conducted on the six biomarkers showing comparable mean intensity for the two populations. One variable at a time was removed, and a new classification model was computed using a simple cross-validation with each reduced dataset. Five of the seven new models yielded decreased sensitivity and specificity, while the other two models provided comparable performance, substantially confirming the choice of the 32 biomarker model.

4. Conclusions

The preliminary results reported in the present study support the premise that GC–MS tridimensional data can be profitably exploited in untargeted metabolomics studies devoted to prostatic carcinoma diagnosis. Compared to the more resource-demanding ultra-high-performance liquid chromatography–tandem mass spectrometer (UHPLC–MS/MS) and ultra-high-performance liquid chromatography–high-resolution mass spectrometry (UHPLC–HRMS) approaches frequently presented in the literature, GC–MS offers comparable chromatographic resolution and structured spectroscopic information, as is generated by the fragment ion pattern typical of electron impact ionization. On these complex data arrays, the ultimate performance in the extraction of crucial information relies on the software purposely adopted and PARAFAC2 combined with VIP and GA methods of variables selection proved to produce highly efficient models of class discrimination, allowing us to distinguish prostatic carcinoma from benign hyperplasia with good sensitivity and specificity scores.

A common limitation of untargeted metabolomics methods, including the present one, is that the most abundant components of the screened samples are preferentially isolated as potential biomarkers with respect to minor constituents, possibly present at trace levels. This explains the differences in the selected biomarkers with respect to the targeted approach that we previously tested [27]. On the other hand, complementary sets of biomarkers are extracted and then evaluated from targeted and untargeted approaches, to be subsequently combined to achieve optimal performance. More work has to be done on large populations of PCa-affected patients and controls to confirm the present findings, and further effort is necessary to reveal the identity of the most valuable biomarkers and possibly confirm their real value as interesting biomarkers by univariate statistics. Despite these limitations to be overcome in the subsequent investigations, the strategy adopted in the present study, based on non-invasive urine sampling, cheap instrumentation, and advanced data treatment by PARADISe software, appears to be extremely promising in PCa screening.

Supplementary Materials: The following are available online.

Author Contributions: All the Authors participated to the preliminary study design and planning. Recruitment, participation criteria, clinical evaluations, F.P.; Development of the analytical method, E.A. (Eleonora Amante) and A.S.; Sample processing, E.A. (Eleonora Amante); Method validation, E.A. (Eleonora Amante), E.A. (Eugenio Alladio), and M.V.; Chemometrics and statistical analysis, R.B., E.A. (Eleonora Amante), and E.A. (Eugenio Alladio); Mass spectra interpretation, M.V., E.A. (Eleonora Amante); Writing—Original Draft Preparation, E.A. (Eleonora Amante); Writing—Review & Editing, M.V. All Authors read and approved the final manuscript.

Funding: This research was funded by the Italian "Ministero dell'Istruzione, dell'Università e della Ricerca" within a PRIN 2017 call for bids (Research Projects of Relevant National Interest—grant 2017Y2PAB8).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Prenser, J.R.; Rubin, M.A.; Wei, J.T.; Chinnaiyan, A.M. Beyond PSA: The next generation of prostate cancer biomarkers. *Sci. Transl. Med.* **2012**, *4*, 127rv3.
2. Velonas, V.M.; Woo, H.H.; Dos Remedios, C.G.; Assinder, S.J. Current status of biomarkers for prostate cancer. *Int. J. Mol. Sci.* **2013**, *14*, 11034–11060. [CrossRef] [PubMed]
3. Hendriks, R.J.; Van Oort, I.M.; Schalken, J.A. Blood-based and urinary prostate cancer biomarkers: A review and comparison of novel biomarkers for detection and treatment decisions. *Prostate Cancer Prostatic Dis.* **2017**, *20*, 12–19. [CrossRef] [PubMed]
4. Etzioni, R.; Penson, D.F.; Legler, J.M.; Di Tommaso, D.; Boer, R.; Gann, P.H.; Feuer, E.J. Overdiagnosis due to prostate-specific antigen screening: Lessons from U.S. prostate cancer incidence trends. *J. Natl. Cancer Inst.* **2002**, *94*, 981–990. [CrossRef] [PubMed]
5. Sardana, G.; Dowell, B.; Diamandis, E.P. Emerging biomarkers for the diagnosis and prognosis of prostate cancer. *Clin. Chem.* **2008**, *54*, 1951–1960. [CrossRef] [PubMed]
6. Kramer, B.S.; Brown, M.L.; Prorok, P.C.; Potosky, A.L.; Gohagan, J.K. Prostate cancer screening: What we know and what we need to know. *Ann. Intern. Med.* **1993**, *119*, 914–923. [CrossRef] [PubMed]
7. Moyer, V.A. Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **2012**, *157*, 120–134. [CrossRef]
8. Gigerenzer, G.; Mata, J.; Frank, R. Public knowledge of benefits of breast and prostate cancer screening in Europe. *J. Natl. Cancer Inst.* **2009**, *101*, 1216–1220. [CrossRef]
9. Tavoosidana, G.; Ronquist, G.; Darmanis, S.; Yan, J.; Carlsson, L.; Wu, D.; Conze, T.; Ek, P.; Semjonow, A.; Eltze, E.; et al. Multiple recognition assay reveals prostasomes as promising plasma biomarkers for prostate cancer. *Expert Rev. Anticancer Ther.* **2011**, *11*, 1341–1343. [CrossRef]
10. Bachrach, U. Polyamines and cancer: Minireview article. *Amino Acids* **2004**, *26*, 307–309. [CrossRef]
11. Schipper, R.G.; Romijn, J.C.; Cuijpers, V.M.; Verhofstad, A.A. Polyamines and prostatic cancer. *Biochem. Soc. Trans.* **2003**, *31*, 375–380. [CrossRef] [PubMed]
12. Lévesque, E.; Huang, S.P.; Audet-Walsh, E.; Lacombe, L.; Bao, B.Y.; Fradet, Y.; Laverdière, I.; Rouleau, M.; Huang, C.Y.; Yu, C.C.; et al. Molecular markers in key steroidogenic pathways, circulating steroid levels, and prostate cancer progression. *Clin. Cancer Res.* **2013**, *19*, 699–709. [CrossRef] [PubMed]
13. Gnanapragasam, V.J.; Robson, C.N.; Leung, H.Y.; E Neal, D. Androgen receptor signalling in the prostate. *BJU Int.* **2000**, *86*, 1001–1013. [CrossRef] [PubMed]
14. Kelloff, G.J.; Lieberman, R.; Steele, V.E.; Boone, C.W.; Lubet, R.A.; Kopelovich, L.; Malone, W.A.; Crowell, J.A.; Higley, H.R.; Sigman, C.C. Agents, biomarkers, and cohorts for chemopreventive agent development in prostate cancer. *Urology* **2001**, *57*, 46–51. [CrossRef]
15. De Luca, S.; Fiori, C.; Manfredi, M.; Amante, E. Preliminary results of prospective evaluation of urinary endogenous steroid profile and prostatic carcinoma-induced deviation. *J. Urol.* **2019**, *201*, e263–e264. [CrossRef]
16. Amante, E.; Alladio, E.; Salomone, A.; Vincenti, M.; Marini, F.; Alleva, G.; De Luca, S.; Porpiglia, F. Correlation between chronological and physiological age of males from their multivariate urinary endogenous steroid profile and prostatic carcinoma-induced deviation. *Steroids* **2018**, *139*, 10–17. [CrossRef] [PubMed]

17. Tomasi, G.; Berg, F.V.D.; Andersson, C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom.* **2004**, *18*, 231–241. [CrossRef]
18. Johnsen, L.G.; Skou, P.B.; Khakimov, B.; Bro, R. Gas chromatography—Mass spectrometry data processing made easy. *J. Chromatogr. A* **2017**, *1503*, 57–64. [CrossRef]
19. Amigo, J.M.; Skov, T.; Bro, R.; Coello, J.; Maspoch, S. Solving GC-MS problems with PARAFAC2. *TrAC Trends Anal. Chem.* **2008**, *27*, 714–725. [CrossRef]
20. Bro, R.; Andersson, C.A.; Kiers, H.A.L. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. *J. Chemom.* **1999**, *13*, 295–309. [CrossRef]
21. Amigo, J.M.; Popielarz, M.J.; Callejón, R.M.; Morales, M.L.; Troncoso, A.M.; Petersen, M.A.; Toldam-Andersen, T.B.; Morales, M.L. Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *J. Chromatogr. A* **2010**, *1217*, 4422–4429. [CrossRef] [PubMed]
22. Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal. Methods* **2013**, *5*, 3790–3798. [CrossRef]
23. Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial Least Squares Projections to Latent Structures. In *3D QSAR in Drug Design: Theory, Methods and Applications*; KLUWER ESCOM Science Publisher: Heidelberg, Germany, 1993; pp. 523–550.
24. Zou, W.; Tolstikov, V.V. Probing genetic algorithms for feature selection in comprehensive metabolic profiling approach. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1312–1324. [CrossRef] [PubMed]
25. Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. *J. Chemom.* **2009**, *23*, 160–171. [CrossRef]
26. Wise, B.; Gallagher, N.; Bro, R. *PLS_Toolbox 8.5*; Eigenvector Research, Inc.: Manson, WA, USA, 2017; Available online: http://eigenvector.com/software/pls-toolbox/ (accessed on 22 August 2019).
27. Choi, M.H.; Moon, J.-Y.; Cho, S.-H.; Chung, B.C.; Lee, E.J. Metabolic alteration of urinary steroids in pre- and post-menopausal women, and men with papillary thyroid carcinoma. *BMC Cancer* **2011**, *11*, 342. [CrossRef] [PubMed]
28. Miller, K.K.M. The Biological Actions of Dehydroepiandrosterone. *Drug Metab. Rev.* **2006**, *38*, 89–116.
29. Arlt, W.; Biehl, M.; Taylor, A.E.; Hahner, S.; Libé, R.; Hughes, B.A.; Schneider, P.; Smith, D.J.; Stiekema, H.; Krone, N.; et al. Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. *J. Clin. Endocrinol. Metab.* **2011**, *96*, 3775–3784. [CrossRef] [PubMed]
30. Arlt, W.; Stewart, P.M. Adrenal corticosteroid biosynthesis, metabolism, and action. *Endocrinol. Metab. Clin. N. Am.* **2005**, *34*, 293–313. [CrossRef] [PubMed]
31. Lionetto, L.; Lostia, A.M.; Stigliano, A.; Cardelli, P.; Simmaco, M. HPLC-mass spectrometry method for quantitative detection of neuroendocrine tumor markers: Vanillylmandelic acid, homovanillic acid and 5-hydroxyindoleacetic acid. *Clin. Chim. Acta* **2008**, *398*, 53–56. [CrossRef]
32. Stephens, F.O.; Unit, O. Phytoestrogens and prostate cancer: Possible preventive role. *Med. J. Aust.* **1997**, *167*, 138–140. [CrossRef]
33. Hedelin, M.; Klint, Å.; Chang, E.T.; Bellocco, R.; Johansson, J.E.; Andersson, S.O.; Heinonen, S.M.; Adlercreutz, H.; Adami, H.O.; Grönberg, H.; et al. Dietary phytoestrogen, serum enterolactone and risk of prostate cancer: The Cancer Prostate Sweden Study (Sweden). *Cancer Causes Control* **2006**, *17*, 169–180. [CrossRef] [PubMed]
34. Rajalahti, T.; Arneberg, R.; Berven, F.S.; Myhr, K.-M.; Ulvik, R.J.; Kvalheim, O.M. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35–48. [CrossRef]

Sample Availability: Samples of the compounds are available from the authors.

molecules

Article

Comparison and Identification for Rhizomes and Leaves of *Paris yunnanensis* Based on Fourier Transform Mid-Infrared Spectroscopy Combined with Chemometrics

Yi-Fei Pei [1,2], Qing-Zhi Zhang [2], Zhi-Tian Zuo [1,*] and Yuan-Zhong Wang [1,*]

[1] Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, Kunming 650200, China; feifei950222@gmail.com

[2] College of Traditional Chinese Medicine, Yunnan University of Traditional Chinese Medicine, Kunming 650500, China; ynkzqz@126.com

* Correspondence: yaaszztian@126.com (Z.-T.Z.); boletus@126.com (Y.-Z.W.); Tel.: +86-871-6503-3575 (Y.-Z.W.); Fax: +86-871-6503-3441 (Y.-Z.W.)

Academic Editor: Marcello Locatelli
Received: 29 November 2018; Accepted: 14 December 2018; Published: 17 December 2018

Abstract: *Paris polyphylla*, as a traditional herb with long history, has been widely used to treat diseases in multiple nationalities of China. Nevertheless, the quality of *P. yunnanensis* fluctuates among from different geographical origins, so that a fast and accurate classification method was necessary for establishment. In our study, the geographical origin identification of 462 *P. yunnanensis* rhizome and leaf samples from Kunming, Yuxi, Chuxiong, Dali, Lijiang, and Honghe were analyzed by Fourier transform mid infrared (FT-MIR) spectra, combined with partial least squares discriminant analysis (PLS-DA), random forest (RF), and hierarchical cluster analysis (HCA) methods. The obvious cluster tendency of rhizomes and leaves FT-MIR spectra was displayed by principal component analysis (PCA). The distribution of the variable importance for the projection (VIP) was more uniform than the important variables obtained by RF, while PLS-DA models obtained higher classification abilities. Hence, a PLS-DA model was more suitably used to classify the different geographical origins of *P. yunnanensis* than the RF model. Additionally, the clustering results of different geographical origins obtained by HCA dendrograms also proved the chemical information difference between rhizomes and leaves. The identification performances of PLS-DA and the RF models of leaves FT-MIR matrixes were better than those of rhizomes datasets. In addition, the model classification abilities of combination datasets were higher than the individual matrixes of rhizomes and leaves spectra. Our study provides a reference to the rational utilization of resources, as well as a fast and accurate identification research for *P. yunnanensis* samples.

Keywords: *Paris polyphylla* Smith var. *yunnanensis*; multivariate analysis; chemometrics; Fourier transform infrared

1. Introduction

The perennial herb plant *Paris* is a genus in the Liliaceae family. *Paris* is one of more than 2000 medicinal plants described in the Chinese Pharmacopoeia (2015 edition), and it has utmost important medicinal effects on treating diseases, including snake bite and insect sting, innominate toxin swelling, and various inflammatory and traumatic injuries with ancient history in China. In addition, *Paris* is also used as an ethnobotanical medicinal herb in Nepal and India, which export *Paris* raw materials every year to China to meet the Chinese traditional medicine (TCM) market demand [1]. *Paris* medicinal plants sold in today's TCM markets were both of wild and cultivated types, with the number of wild *Paris* gradually decreasing, with a long-term growth cycle, immoderate harvesting, and huge commercial activities [2]. Additionally, amongst almost 28 species and varieties of *Paris*, only *Paris polyphylla* Smith var. *chinensis* (Franch.) Hara (*P. chinensis*) and *P. polyphylla* var. *yunnanensis* (Franch.) Hand. -Mazz (*P. yunnanensis*) are officially described by the Chinese Pharmacopoeia (2015 edition), which further restricted the number of *Paris* medicinal plants [3–5]. Hence, substitutes with similar medicinal effects and chemical compounds are considered for selection from the closely related species of *P. yunnanensis* and *P. chinensis*, and other parts of the plants, such as stems and leaves.

A serious problem is that many number of leaves of *Paris* medicinal plants were abandoned every year, with the rhizomes being unable to meet the market demand. Thus, the use of *P. yunnanensis* and *P. chinensis* leaves as substitutes for the primary choice was to be considered. Currently, Qin et al. have reviewed the feasibility for whether renewable above-ground parts (leaves and stems) of *P. yunnanensis* could be used as an alternative source to rhizomes [6]. They concluded that the above-ground parts can be the substitute source for the rhizomes of *P. yunnanensis*, in that similar pharmacological properties, including antimicrobial, hemostatic, cytotoxic, and other effects. A variety of quality of Paridis Rhizomes in TCM markets may affect the quality of Chinese patent medicines based on *P. yunnanensis* rhizomes. On these basis, it is necessary and meaningful to quickly assess the quality of *P. yunnanensis* rhizomes and leaves.

Yunnan possesses complex climatic conditions, which means that the quality of TCM plants varies with different climatic conditions of different geographical origins in Yunnan. A variety of analytic techniques have been applied to determine the active chemical components and fingerprints to assess the quality of *P. yunnanensis* samples, including ultra-high performance liquid chromatography-mass spectrometry (UHPLC-MS) [7,8], ultraviolet-visible (UV-Vis) [9], high performance liquid chromatography (HPLC) [10], and Fourier transform mid infrared (FT-MIR) [8,11,12], and so on. Up to now, chemometrics has been widely applied to herbal medicines and plant spectral analyses [13,14]. For example, principal component analysis (PCA) often was used to research Chinese herbal medicines of multiple tissues and geographical origins [15,16]. Partial least squares discriminant analysis (PLS-DA) and random forest (RF) have been gradually applied to the field of traditional Chinese herbs in recent years, such as *Panax notoginseng*, *Dendrubium officinale*, etc. [17,18]. Our previous studies have demonstrated that all of these techniques have obtained better identification abilities for *P. yunnanensis* from different geographical origins. Compared with chromatography, the better classification abilities, more convenience, and time-saving techniques were displayed using spectroscopy techniques. To date, combined with various analytical techniques, chemometrics methods have been successfully applied to assess *P. yunnanensis* samples with better classification and identification abilities, including support vector machine [19], RF [11,12], hierarchical cluster analysis (HCA) [10,12,20,21], PLS-DA [9,12,22], and PCA [9,12,22]. However, they failed to analyze other parts of *P. yunnanensis* to fast assess their quality, as well as comparing and combining rhizomes and leaves to identify *P. yunnanensis* from a variety of geographical origins. Hence, the purpose of our study is to assess the quality of *P. yunnanensis* medicinal materials by determining their rhizomes and leaves in FT-MIR spectra, combined with chemometrics.

In this study, to further obtain better, faster, and reliable identification methods for *P. yunnanensis* raw materials from different geographical origins, we investigated *P. yunnanensis* samples from six regions from Yunnan Province by FT-MIR spectroscopy, combined with four chemometrics methods, including PCA, PLS-DA, RF, and HCA. The influence on the fast-quality assessment effects of different parts, including leaves and rhizomes of *P. yunnanensis* were compared. The results may demonstrate the importance of the leaves of *P. yunnanensis*, and they can provide direction for the future development of *P. yunnanensis* medicinal plants.

2. Results and Discussion

2.1. Comparison Analysis between Rhizomes and Leaves

The raw and SD FT-MIR spectra of rhizomes and leaves of *P. yunnanensis* samples from six geographical regions are showed in Figure 1. The peaks height, character, and position among different geographical origins samples are similarly shown in Figure 1a. Characteristic peaks appeared at ~3328 cm^{-1}, and were assigned to O–H absorption, at ~2726, 1414, and 1370 cm^{-1} to methylene and methyl stretching, and bending vibration. Absorption at ~1742 cm^{-1} was endorsed to C=O stretching vibration, at ~1650 cm^{-1} it was attributed to C=C and C=O stretching vibration, which may be attributed to oils, saccharides, steroid saponins, and flavonoids. Besides, absorption at ~1244 cm^{-1} was assigned to C–O stretching vibration, while ~1151, 1078, and 1020 cm^{-1} were endorsed to C–C, C–O stretching vibration and C–OH bending vibration, as well as the main attribute to saccharides and glycosides. Absorption at ~929 cm^{-1} was assigned to the sugar skeleton. These attributes for characteristic peaks were in accordance with studies by Sun et al. and Yang et al. [23,24]. Absorption at ~2855, 1547, 1340, 862, 765, 708, 611, and 580 cm^{-1} were also showed in these FT-MIR spectra. Absorption at ~1650 cm^{-1} and ~1020 cm^{-1} were the key peaks among all absorption peaks of the raw FT-MIR spectra of rhizomes. Additionally, many details of spectral information were shown by standard normal variate–second derivative (SNV-SD) FT-MIR rhizomes spectra in Figure 1c. In detail, among the peaks regions of 1200–900 cm^{-1}, the peaks absorptions were at 1173, 1135, 1093, 1065, 1050, 1035, 996, 976, and 950 cm^{-1}, which are not shown in the raw FT-MIR spectra of rhizomes.

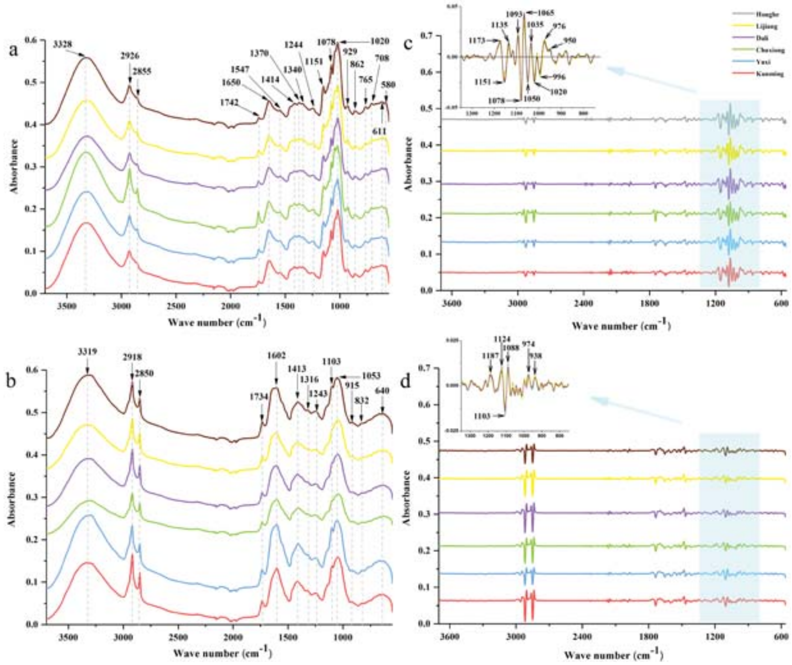


Figure 1. The FT-MIR spectra of Kunming, Yuxi, Chuxiong, Dali, Lijiang, and Honghe, Yunnan: (**a**) the raw spectra of rhizomes, (**b**) the raw spectra of leaves, (**c**) the best preprocessing spectra of rhizomes, (**d**) the best preprocessing spectra of leaves.

The raw FT-MIR spectra of leaves showed different peak heights, characters and positions and numbers of the characteristic peaks for those of rhizomes, which are shown in Figure 1b. Compared with the raw rhizomes FT-MIR spectra, absorption for the raw leaves spectra exhibited a red-shift at

1750–1290 cm^{-1}, and a blue-shift at 1290–950 cm^{-1}. In other words, various differences of chemical information was reflected by the raw rhizomes and leaf FT-MIR spectra. Similar to the raw rhizome FT-MIR spectra, the absorption was mainly attributed to oils, saccharides, steroid saponins, flavonoids saccharides, and glycosides. Namely, absorption at 1602 cm^{-1} and 1053 cm^{-1} are the two key peaks of the raw leaf FT-MIR spectra. Similarly, certain details from the spectral information are shown in SNV-SD leaf FT-MIR spectra in Figure 1d. In detail, among peaks regions of 1200–900 cm^{-1}, the peak absorptions at 1187, 1124, 1088, 974, and 938 cm^{-1} are proven, which are not shown in the FT-MIR spectra of raw leaves.

The PCA score plot and loading plot based on the total FT-MIR spectra are shown in Figure 2. Besides, 72.9% and 17% FT-MIR spectra information were exhibited by PC 1 and PC 2, respectively. Two parts (rhizomes and leaves) were well separated by the first two principal components (PCs) in the PCA score plot. Absorption at 1300–550 cm^{-1} by PC 1 contributed to a higher importance than that of PC 2. In other words, the bands of this region are more important to PC 2.

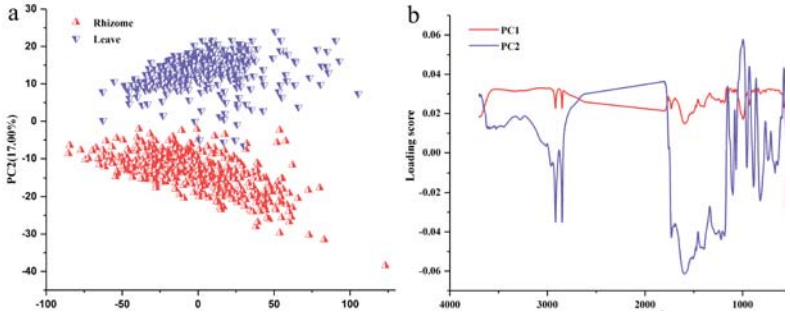


Figure 2. Principal component analysis (PCA) result based on Fourier transform mid infrared (FT-MIR) spectra: (**a**) Score plot, (**b**) Loading plot.

2.2. Origin Traceability Based on Chemometrics

2.2.1. Using Rhizome FT-MIR Spectra Datasets

Raw FT-MIR rhizomes spectra were pretreated by SNV, standard normal variate-first-derivative (SNV-FD), SNV-SD, and SD preprocessing methods, and to select the best pretreatment method. All parameters for these pretreatment methods are shown in Table S1. Comparing parameters to the raw FT-MIR spectra, all parameters are better after preprocessing. Among them, SNV-SD was defined as the optimal preprocessing method for the fundamental for the larger values of cumulative interpretation ability (R^2), cumulative prediction ability (Q^2), and accuracy of the calibration set, as well as the lower values of the root mean square error of estimation (RMSEE) and the root mean square error of cross-validation (RMSECV). Despite SD obtaining a better accuracy, SNV-SD obtained a lower RMSEE, RMSECV, and latent variables (LVs). In our following study for rhizomes, models established by raw and the best preprocessing (SNV-SD) FT-MIR spectra data will be compared.

The variable importance for the projection (VIP) scores for values greater than 1 of the raw rhizome FT-MIR data are shown in Figure 3a. The regions of 1750–1500 cm^{-1} and 1200–750 cm^{-1} are important variables regions for differentiating six geographical origins of *P. yunnanensis* by FT-MIR spectra. The bonds at 1750–1500 cm^{-1} are mainly attributed to oils, saccharides, steroid saponins, and flavonoids compounds. Besides, the bands at 1200–750 cm^{-1} are mainly endorsed to saccharides and glycosides compounds. The two key peaks of raw rhizome FT-MIR spectra were contained in these two bands. What's more, there were also some peaks that were not clearly identified, and these peaks are equally important for the identification of *P. yunnanensis* samples from different origins. On the basis of the SNV-SD rhizome FT-MIR data, the VIP scores for values greater than 1 are shown in Figure 3b. The degrees of important variables regions from 1750–750 cm^{-1} seem to be similar in importance for the differentiation of six geographical origins of *P. yunnanensis* by FT-MIR spectra. It was further demonstrated that each peak was important for distinguishing *P. yunnanensis* samples from different geographical origins.

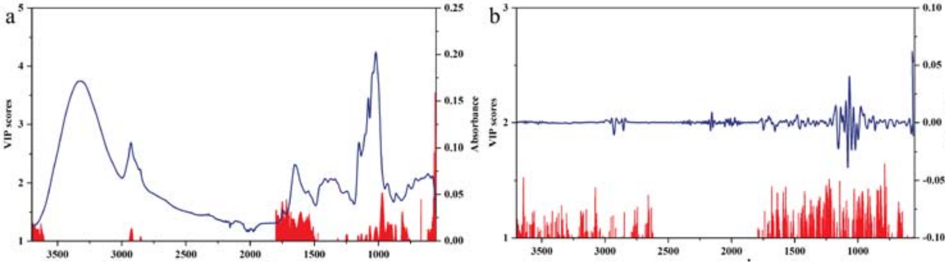


Figure 3. Variable importance for the projection (VIP) scores of the FT-MIR data of rhizomes for regional differences: (**a**) raw dataset, (**b**) standard normal variate–second derivative (SNV-SD) dataset.

RF models were established on raw and SNV-SD rhizome FT-MIR spectra data matrixes. The 1207 and 1202 variables were contained in raw and SNV-SD rhizome FT-MIR spectra datasets, respectively. For the two RF models of raw and SNV-SD rhizomes FT-MIR spectra, the initial number of trees (n_{tree}) were set as 2000 trees. The suitable value of n_{tree} was selected, based on the lowest total value, and the need to be assured of the lower values of the most classes. The 1328–1392 trees and 650–740 trees are the lowest ranges for n_{tree} of raw and SNV-SD rhizomes datasets, respectively, which are shown in Figure 4a,b. Besides, the optimal values 1383 and 951 trees were obtained for further selection of the suitable number of variable (m_{try}) values of the RF models, based on raw and SNV-SD rhizomes FT-MIR datasets, respectively. As shown in Figure 4c,d, the optimal m_{try} were calculated to be 33 and 36, according to the lowest out-of-bag (OOB) values for the raw and SNV-SD datasets, respectively. The suitable n_{tree}, combined with the optimal m_{try}, were used to select the most important variables.

To start with, all variables of the raw and SNV-SD datasets were sorted from the least important variables, to the most important variables, respectively. The 10-fold cross validation error rates of the RF model, based on raw and SNV-SD FT-MIR datasets of rhizomes *P. yunnanensis* samples are shown in Figure 5a,b. It was reduced sequentially by five variables for each step for the initial variables of 1207 and 1202, for raw and SNV-SD datasets, respectively. In both the range of 1–1207 and 1–1202 variables numbers, all important variables were divided into three regions. Among these regions, the 10-fold cross validation error rate values showed a reduced or incremental trend. When the 10-fold cross validation error rate shows a drop trend and then an upward trend, that number of variables at the turning point is likely to be the optimal number for the most importance variables. Hence, variable numbers of 207 and 292 with a lower than 10-fold cross validation error rate for 0.34202 and 0.08143 were selected, to establish the RF models of raw and SNV-SD rhizome FT-MIR spectra, respectively.

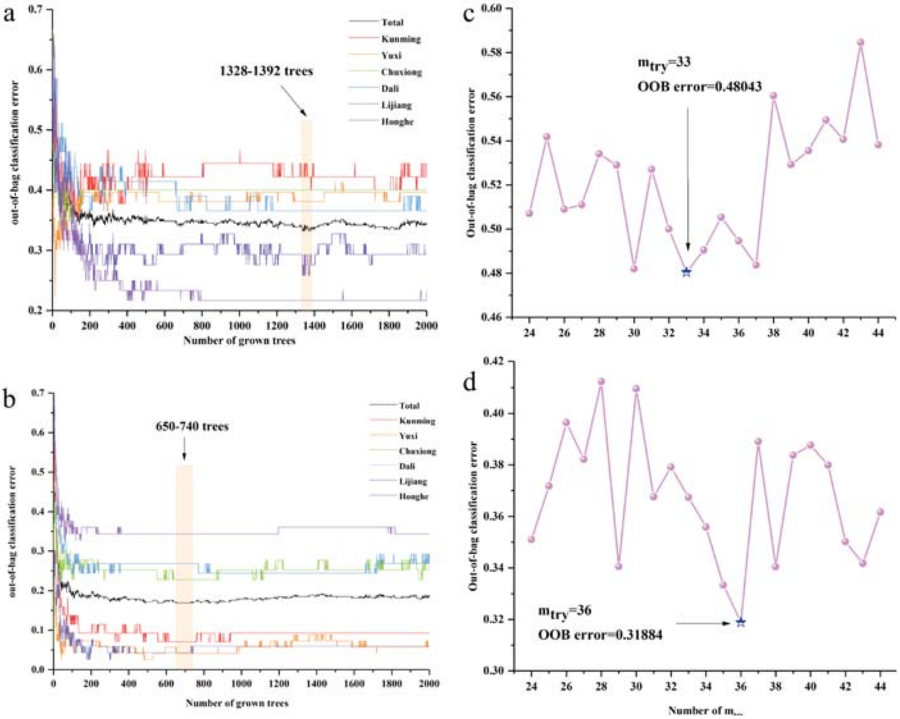


Figure 4. The n_{tree} and m_{try} screening of random forest (RF) models of *P. yunnanensis* samples before variables ranked by permutation accuracy importance: (**a**) n_{tree} of the raw rhizomes dataset, (**b**) n_{tree} of the SNV-SD rhizomes dataset, (**c**) m_{try} of the raw rhizomes dataset, (**d**) m_{try} of the SNV-SD rhizomes dataset.

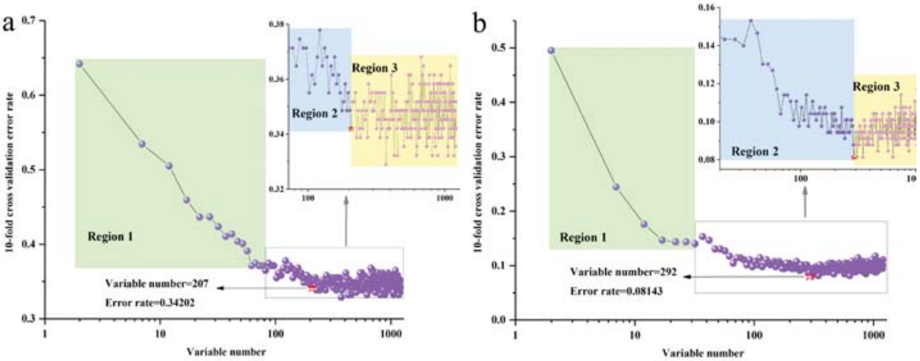


Figure 5. The 10-fold cross validation error rates of the RF model (sequentially reduce each five variables) based on *P. yunnanensis* samples: (**a**) raw rhizomes dataset, (**b**) SNV-SD rhizomes dataset.

When the most important variables were re-selected, forming the new data matrix, it was necessary for the reconstruction of optimal n_{tree} and m_{try} values for raw and SNV-SD FT-MIR spectra. The selecting process was the same as above. As shown in Figure 6, the 1011–1201 trees and 788–880 trees are the lowest ranges for n_{tree} of raw and SNV-SD rhizomes dataset, respectively. Finally, 1110 and 820 trees are selected for the optimal n_{tree}, as well as 19 and 26, are selected for the best m_{try} of raw and SNV-SD FT-MIR rhizome spectra of *P. yunnanensis* samples, respectively. These optimal n_{tree} and m_{try} were used to establish the RF model, and they obtained the accuracy of the calibration set and the validation set, respectively. It is undeniable that the variable selection process is important. The error rate for calibration set of raw datasets was reducing from 36.16% to 33.88%, and it was decreasing from 10.42% to 8.79% for the SNV-SD dataset. In addition, the geographical origin classification ability of the RF model, based on SNV-SD FT-MIR spectra of rhizome *P. yunnanensis* samples, was significantly better than that of the raw spectra.

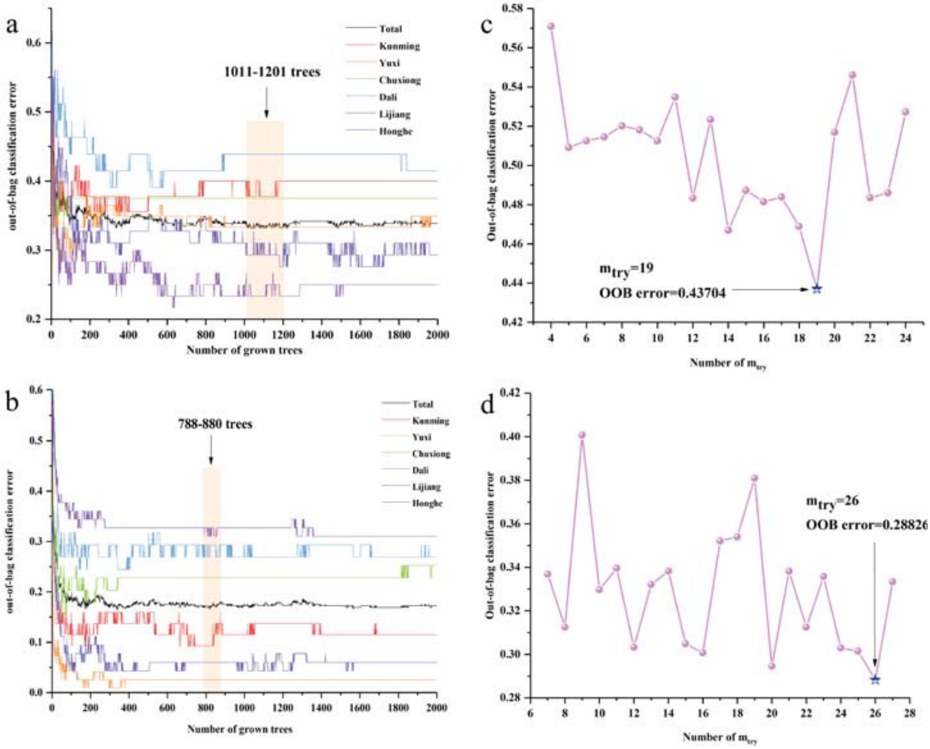


Figure 6. The n_{tree} and m_{try} screening of RF models of the *P. yunnanensis* samples after variables are ranked by permutation accuracy importance: (**a**) n_{tree} of the raw rhizomes dataset, (**b**) n_{tree} of the SNV-SD rhizomes dataset, (**c**) m_{try} of the raw rhizomes dataset, (**d**) m_{try} of the SNV-SD rhizomes dataset.

The parameters for each class of calibration set and validation set of the PLS-DA and RF models, based on raw and SNV-SD rhizomes FT-MIR spectra data matrixes, are shown in Table S2. The values for all parameters of each class of calibration set and the validation set for the PLS-DA model, based on raw FT-MIR data matrixes, were higher than that of the RF model, and they differ greatly. Additionally, all parameters for the RF model based on SNV-SD FT-MIR data matrixes were greatly enhanced and close to that of the PLS-DA model. Obviously, the parameters of two models for the SNV-SD data matrixes based on FT-MIR spectra were higher than those of raw data matrixes. However, the identification abilities and accuracy for two models based on rhizome FT-MIR spectra were needed for improvement.

2.2.2. Using Leaf FT-MIR Spectra Datasets

Raw leaf FT-MIR spectra dataset was preprocessed by SNV, SNV-FD, SNV-SD, and SD preprocessing methods to select the best pretreatment method. All parameters for these four kinds of preprocessing methods are displayed in Table S3. Similarity to rhizomes, all parameters for the preprocessed model of FT-MIR spectra for leaves are better than those of the raw data matrix. Besides, the SNV-SD pretreatment among all preprocessing methods was the best one for classifying the different origins of *P. yunnanensis* leaf samples, which possessed values of R^2, Q^2, RMSEE, RMSECV, accuracy and LVs that were more satisfactory than other pretreatment methods. For the following study of leaves, models established by raw data, and the best pretreatment (SNV-SD) FT-MIR spectra data were selected to study.

The VIP scores for values greater than 1 of the raw leaf FT-MIR data are shown in Figure S1a. The region of 1800–1700 cm^{-1} is the most important variable region for differentiating six geographical origins of *P. yunnanensis* by leaf FT-MIR spectra. The regions of 1700–1300 cm^{-1}, 1250–1100 cm^{-1}, and 1200–750 cm^{-1} almost possessed equally important degrees for differentiating various geographical origins of *P. yunnanensis* by leaf FT-MIR spectra. The bonds at these regions are also mainly assigned to oils, saccharides, steroid saponins, and flavonoids, saccharides, and glycoside compounds. What's more, the number of important variables of leaf VIP scores were more than those of rhizome VIP scores, which reflected the difference in chemical information in classifying *P. yunnanensis* samples from different regions. Based on the SNV-SD leaf FT-MIR data, the VIP scores for values greater than 1 are displayed in Figure S1b. Compared with the other three regions, variables important for the region of 1800–1700 cm^{-1} show greater importance. Similar, it was also demonstrated that each peak of leaf FT-MIR spectra was important to distinguish *P. yunnanensis* samples from a variety of geographical origins. However, a number of peaks were non-identified chemical compounds in the leaf FT-MIR spectra.

RF models were established on raw and SNV-SD leaf FT-MIR spectra datasets. To start with, the 1207 and 1201 variables were contained in the raw and SNV-SD leaf FT-MIR spectra matrixes, respectively. Similar to the rhizomes, the initial n_{tree} were set as 2000 trees for the RF models of raw and SNV-SD leaf FT-MIR spectra. As shown in Figure S2a,b, 947–961 trees and 980–1008 trees were selected to be the lowest ranges for n_{tree} of raw and SNV-SD leaf datasets, respectively. Additionally, the optimal values of 951 and 982 trees were selected for further selection of the suitable m_{try} values of RF models, based on raw and SNV-SD leaf FT-MIR datasets, respectively. As shown in Figure S2c,d, the optimal m_{try} were calculated to be 42 and 31, respectively.

Like rhizomes, all variables of the raw and SNV-SD matrixes of leaves were ranked from to the least important variables to the most important variables, respectively. The 10-fold cross-validation error rates of the RF model, based on the raw and SNV-SD FT-MIR datasets of leaf *P. yunnanensis* samples are shown in Figure S3a,b. In addition, in both the range of 1–1207 and 1–1201 variables numbers, all important variables, were also divided into three regions. Moreover, variable numbers of 157 and 441 with lower than 10-fold cross validation error rates for 0.36808 and 0.02280 were selected to establish the RF models of the raw and SNV-SD FT-MIR spectra, respectively. The 10-fold cross validation error rate of the SNV-SD matrix was far below that of the raw dataset.

Similar to rhizomes, the most important variables were as the new data matrixes, and meanwhile, the optimal n_{tree} and m_{try} values for raw and SNV-SD datasets were re-selected, respectively. The selection process was the same as above. As shown in Figure S4, the 1527–1607 trees and 898–966 trees were the lowest ranges for n_{tree} of raw and SNV-SD leaf datasets, respectively. Then, 1570 and 900 trees were selected for the optimal n_{tree}, as well as 18, and 18 were selected for the best m_{try} of the raw and SNV-SD datasets, respectively. Furthermore, these optimal n_{tree} and m_{try} were used to establish high-performance RF models. The error rate for the calibration set of raw datasets was reduced from 40.07% to 38.11%, and it decreased from 3.26% to 2.93% for the SNV-SD dataset. In addition, not only was the geographical origin classification ability of the RF model based on SNV-SD FT-MIR leaves spectra significantly better than that of the raw spectra, but higher performances were also obtained by the RF models of leaves than those of rhizomes.

Parameters of sensitivity (SENS), specificity (SPEC), accuracy (ACC), and the Matthews correlation coefficient (MCC) for each class of calibration set and validation set of PLS-DA and RF model, based on raw and SNV-SD leaf FT-MIR spectra data matrices are displayed in Table S4. Similar to the performance of parameters for the models of rhizomes, the values for all parameters of each class of calibration set and validation set for the PLS-DA model, based on raw leaf FT-MIR data matrices, were higher than that of the RF model. The identification ability of the SNV-SD PLS-DA model of the leaf data matrix almost reached the best ratings, and only samples collected from Yuxi and Dali were misclassified. Additionally, all parameters of validation set for the RF model based on the SNV-SD FT-MIR data matrixes were close, to the best, and only samples collected from Dali and Lijiang were misclassified. Additionally, parameters of two models for the SNV-SD data matrices based on FT-MIR spectra were higher than those of raw data matrixes. However, the classification performance for the PLS-DA and RF models on the basis of the leaf FT-MIR spectra were required for enhancement.

2.3. Regional Differences between VIP and Important Variables

The VIP and important variables of the RF and PLS-DA models of *P. yunnanensis* samples are displayed in Figure 7. In detail, Figure 7a,b are based on the raw FT-MIR spectra of rhizomes and leaves, respectively. Figure 7c,d are based on the SNV-SD FT-MIR spectra of rhizomes and leaves, respectively. The important variable numbers of the RF model of raw datasets for rhizomes and leaves were far more than those of the SNV-SD RF models. The variables with VIP values greater than 1 showed greater concentrations for several regions in the VIP scores based on raw rhizome and leaf matrixes, than those of the VIP scores of the SNV-SD datasets. From a comparison of the scatter of the most important variables between rhizomes and leaves, the number and distribution of important variables are different. It was demonstrated that various and different chemical profiles were contained between the rhizomes and leaves of *P. yunnanensis*. From the higher accuracy rate and the more uniform distribution of important variables of rhizomes or leaves in the PLS-DA model than those of rhizomes or leaves in the RF model, it was found that the PLS-DA was more suitable for the identification of geographical origins for *P. yunnanensis*.

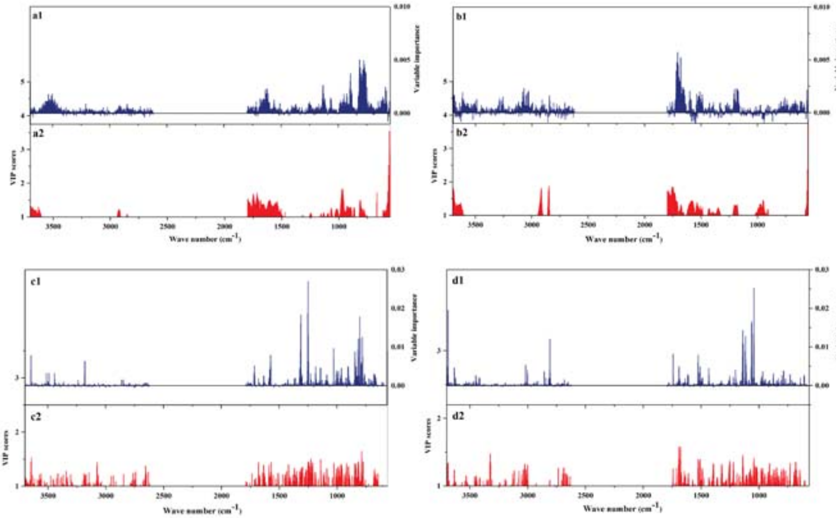


Figure 7. The importance variables (1) of RF models and the VIP values (2) of partial least squares discriminant analysis (PLS-DA) models of the *P. yunnanensis* samples: (**a**) the raw rhizomes dataset, (**b**) the raw leaves dataset, (**c**) the SNV-SD rhizomes dataset, (**d**) the SNV-SD leaves dataset.

2.4. Data Fusion Strategy

Despite the high performance obtained by PLS-DA, and the RF classification models of leaves of the FT-MIR spectra of *P. yunnanensis* samples, the 100% identification accuracy of the calibration set and the validation set were not acquired, and models' abilities needed further enhancement. Hence, the data fusion strategy was used to further improve the prediction abilities of PLS-DA and RF models. Data fusion were concatenated variables of FT-MIR spectra from different parts, forming a single matrix where row numbers were the analyzed sample quantities, and columns consisted of variables. In other words, the rhizome and leaf datasets were combined to establish the classification models.

The process for establishing the data fusion RF model was similar to the individual dataset. RF models were established based on raw and SNV-SD data fusion FT-MIR spectra datasets. A total of 2414 and 2403 variables were contained in the two matrices, respectively. As shown in Figure S5a,b, 356–399 trees and 375–404 trees were the lowest ranges for n_{tree} of raw and SNV-SD matrices, respectively. Additionally, the optimal values of 377 and 393 trees were selected for further selection of the suitable m_{try} values for raw and SNV-SD datasets, respectively. As shown in Figure S5c,d, the optimal m_{try} were 51 and 18, respectively. Similar to the individual dataset, all variables were in ascending order with importance. The 10-fold cross-validation error rates of the RF model for raw and SNV-SD data fusion datasets are shown in Figure S6a,b, respectively. Additionally, variable numbers of 69 and 288 with the lower 10-fold cross-validation error rates of 0.34853 and 0.02606 were selected to establish the data fusion RF models. Besides, the most important variables were the new data matrices, while re-selecting the optimal n_{tree} and m_{try} values for raw and SNV-SD data fusion datasets, respectively. As shown in Figure S7, the 1609–1660 trees and 98–125 trees were the lowest ranges for n_{tree} of the two datasets, respectively. Besides, 1652 and 104 trees, as well as 10 and 18, are selected for the best n_{tree} and $m_{try,}$ respectively. Compared to the accuracy of the RF models between the raw and SNV-SD data fusion matrixes, the error rate for the calibration set of the raw dataset decreased from 37.46% to 37.13%, and decreased from 2.61% to 1.63% for the SNV-SD dataset. The classification abilities in the rhizome and lead data fusion RF model were better than in the individual dataset RF model.

From a comparison of parameters for SENS, SPEC, ACC, and MCC between the PLS-DA and RF models, based on data fusion strategy, the PLS-DA model had a better classification ability than that of the RF model. As shown in Table 1, the geographical origins identification abilities reached the best of each class calibration set and validation set for the PLS-DA model of the SNV-SD FT-MIR spectra. However, the parameter values were close to 100% for most classes of RF model. Hence, it could be demonstrated that the PLS-DA model was more suitable for tracing the different geographical origins of cultivated *P. yunnanensis*.

Table 1. The major parameters of PLS-DA and RF models of each class, based on the data fusion SNV-SD FT-MIR spectra datasets of *P. yunnanensis* samples.

Preprocessing	Set	Classes [a]	PLS-DA				RF			
			SENS	SPEC	ACC	MCC	SENS	SPEC	ACC	MCC
SNV-SD	Calibration set	1	1	1	1	1	1	0.996	0.997	0.987
		2	1	1	1	1	0.984	0.996	0.993	0.98
		3	1	1	1	1	0.975	1	0.997	0.986
		4	1	1	1	1	0.951	0.992	0.987	0.944
		5	1	1	1	1	0.9831	1	0.997	0.989
		6	1	1	1	1	1	0.996	0.997	0.99
	Validation set	1	1	1	1	1	1	1	1	1
		2	1	1	1	1	1	1	1	1
		3	1	1	1	1	1	1	1	1
		4	1	1	1	1	0.95	1	0.994	0.971
		5	1	1	1	1	1	0.992	0.994	0.979
		6	1	1	1	1	1	1	1	1

[a] 1: Kunming, 2: Yuxi, 3: Chuxiong, 4: Dali, 5: Lijiang, 6: Honghe. Sensitivity (SENS), specificity (SPEC), accuracy (ACC) and the Matthews correlation coefficient (MCC).

2.5. Hierarchical Clustering Analysis

HCA dendrograms based on average SNV-SD FT-MIR spectra datasets of rhizomes and leaves of *P. yunnanensis* from different geographical origins are presented in Figure 8a,b, respectively. It is obviously that all the six classes are grouping into two main clusters, both in the two HCA dendrograms. However, the clustering results among Kunming, Yuxi, Chuxiong, Dali, Lijiang and Honghe were obtained based on rhizomes and leaves FT-MIR spectral matrixes were different. As shown in Figure 9, the altitude is decreasing gradually from Northwest Yunnan to Southeast Yunnan. In addition, the two main clusters are influenced to some extent by the topography including altitude. Nevertheless, Kunming was cluster with Honghe and Yuxi in HCA dendrograms based on rhizomes dataset but cluster with Lijiang, Dali and Chuxiong of HCA plot based on leaves. It is demonstrated that the different chemical information between rhizomes and leaves of *P. yunnanensis* were influenced on the results of clustering.

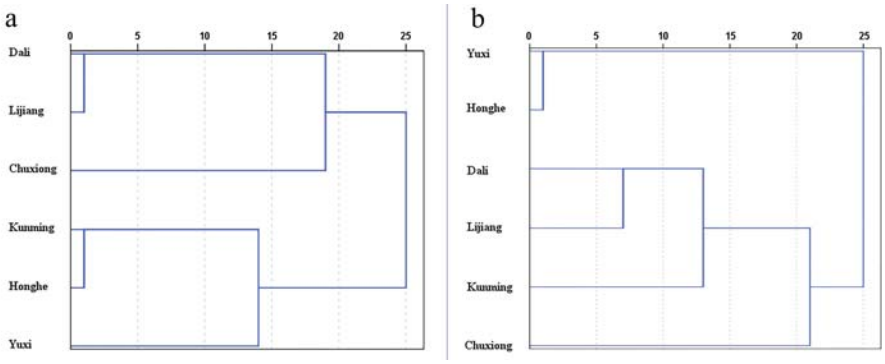


Figure 8. Dendrograms resulting of hierarchical cluster analysis (HCA) based on six geographical origins of *P. yunnanensis* samples: (**a**) the rhizomes dataset, (**b**) the leaves dataset.

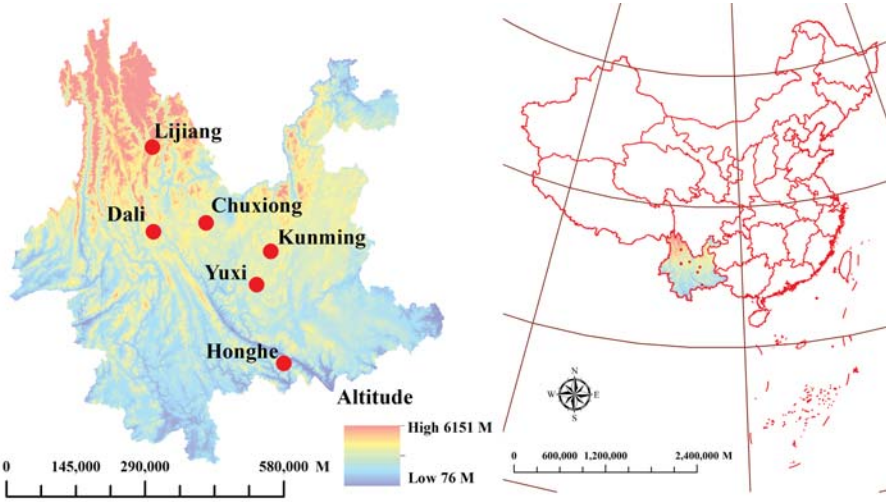


Figure 9. Location distribution of cultivated *P. yunnanensis* samples in Kunming, Yuxi, Chuxiong, Dali, Lijiang and Honghe, Yunnan Province.

3. Materials and Methods

3.1. Plant Material Preparation

In our experiment, rhizomes and leaves of 462 cultivated *P. yunnanensis* samples were collected from Kunming, Yuxi, Chuxiong, Dali, Lijiang, and Honghe cities in Yunnan Province; the collection locations and detailed information are shown in Figure 9 and Table S5. All samples were identified as *P. polyphylla* Smith var. *yunnanensis* (Franch.) Hand.-Mazz. by Professor Hang Jin (Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, China). To start with, the different parts for each *P. yunnanensis* samples were separated and washed, then dried at 50 degrees Celsius. In addition, both rhizome and leaf samples were sifted through 100 mesh sieves, and stored in a relatively dry environment.

3.2. FT-MIR Spectral Acquisition

FT-MIR analysis uses a FTIR spectrometer with a DTGS detector equipped, combined with a ZnSe attenuated total reflectance accessory (Perkin Elmer, Norwalk, CT, USA). The FT-MIR spectra collection parameters and methods are referenced in our previous experiment [14]. The FT-MIR spectra recorded ranges of 4000–550 cm^{-1} with 4 cm^{-1} resolution and 16 scans, both for rhizomes and leaves of each of the *P. yunnanensis* samples. Three scans were repeated for all rhizomes and leaves samples. Moreover, it was required that a relatively constant temperature and humidity was provided during the assessment of the FT-MIR spectra.

3.3. Chemometrics Methods

3.3.1. Principal Component Analysis

PCA is an exploratory data analysis method and an unsupervised pattern recognition technique, which seeks for the optimum data distribution in a multivariate space [25–27]. The fundamental of PCA is that all the raw data are projected onto a two-dimensional sub-space, to ensure that information loss is minimized. The higher the front PCs, the higher the proportion of important variables represented. Generally, the first few PCs represent the most information. The first two or three PCs of all samples can be shown in two- or three-dimensional scores plots, and they further show the regularities of distribution for all the samples. Moreover, the relationship between the first two PCs and wave numbers can be shown by the loading plot.

3.3.2. Partial Least Squares Discriminant Analysis

PLS-DA, a binary classification algorithm from 0 to 1, is based on the PLS algorithm, to add category labels to achieve the effect of classification prediction, and it shows the relationship by multivariate projection between independent and dependent variables, which are expressed by X and Y, respectively [28,29]. Besides, LVs were one feature variable that were produced by an intermediate process in the PLS-DA method [30]. LVs are useful for us to analyze the important variables and information. The X matrix and target and important values in Y are more closely correlated than the noise or unimportant values in Y. Additionally, the VIP plot summarizes the importance of the variables, both to explain X, and to correlate to Y, meaning that variables with a VIP value greater of than 1 are important; as well, those that are greater than 0.5 and less than 1 may be important, depending on the circumstances. Hence, classifying samples by PLS-DA requires that variables possess numbers that are greater than the classification sample numbers, and there should be some correlation among the identified samples.

3.3.3. Random Forest

RF model, developed by Breiman in 2001, has been widely used to resolve classification problems in the field of food, and so on [31,32]. The RF model is based on the assembly classification or regression trees algorithm, and it shows a higher ability to resolve binary classification or regression issues [31]. The operational steps of the RF model can be roughly divided into the following five steps. Firstly, a spectra dataset was separated into two parts according to the ratio of 2 to 1, by the Kennard-stone (KS) algorithm by MATLAB 2017a (MathWorks, Natick, MA, USA) [33,34]. Two-thirds of the dataset was assigned as the calibration set (bootstrap samples), and one-third as the validation set (out-of-bag samples). The calibration set was used to obtain the optimal classification trees, and the validation set was applied to evaluate the ability of the FR model. Besides, the initial values of n_{tree} and m_{try} were defined as 2000, and the square root of the number of all variables, respectively. The optimal n_{tree} and m_{try} were both selected according to the lowest OOB classification error values. Thirdly, the most important variables were selected by a lower 10-fold cross-validation error rate, and as a new data matrix reimport. Fourth, the optimal n_{tree} and m_{try} were reselected according to the fundamental of step 2. Finally, the establishment of the final RF discrimination model was performed by using the optimized n_{tree} and m_{try} parameters. Step two to five were completed by R package (version 4.6–14).

3.3.4. Hierarchical Cluster Analysis

HCA clusters different categories at a certain distance, according to the degree of similarity of each class, which means that it could preliminarily identify a classification trend for each category [35]. Besides, the Person correlation coefficient was applied to measure the linear relationship between the distance variables. These analyses were completed by SPSS 20.0 software (IBM Corp., Armonk, NY, USA).

3.4. Data Analysis

The purpose of data analysis involves the reduction of the influence by noise and other factors from experiments and instruments on the raw FT-MIR spectra data. Firstly, the raw FT-MIR spectra were pretreated by advancing ATR (attenuated total reflection) correction, and absorbance was transformed from transmittance by OMNIC 9.7.7 (Thermo Fisher Scientific, Madison, WI, USA). Secondly, the best preprocessing method was selected among a combination of various pretreatment methods, including SNV, FD and SD, which can enhance the accuracy and feasibility for identification study [36,37]. SNV and its derivatives could decrease a part of the irrelevant interferences, such as high frequency random noise, the interference of light scattering, baseline drift, and unequal concentration, and so on, to improve the classification ability of the models. All these preprocessing methods were completed by SIMCA-P^+ 13.0 (Umetrics, Umea, Sweden). The datasets were separated into a calibration set and a validation set, with a rate of 2 to 1 by the KS algorithm, using MATLAB 2017a (The MathWorks), which was also used to establish the PLS-DA and RF models. In other words, the FTIR spectra of samples were divided into a calibration set (307 samples) and validation set (155 samples), as shown in Table S6.

Generally, parameters including RMSEE, RMSECV, and the accuracy of calibration sets Q^2 and R^2 were used to estimate the identification ability of the calibration model [38,39]. The optimal preprocessing model required lower values of RMSEE and RMSECV, as well as higher values of accuracy for the calibration sets R^2 and Q^2. Besides, the model may have poor robustness and over-fitting when the values of the root mean square error of prediction (RMSEP) are greater than that of RMSECV [12]. In addition, due to both the PLS-DA and RF models being able to obtain the vote matrices, the two models could calculate the values of true negative (TN), true positive (TP), false negative (FN), and false positive (FP), respectively. SENS (Equation (1)), SPEC (Equation (2)), ACC (Equation (3)), and MCC (Equation (4)) were the four parameters for each class, resulting in identification effects for different geographical origins of *P. yunnanensis* samples of PLS-DA and RF models. Obviously, this led to the higher values of these four parameters and a better identification ability for each class.

$$\mathrm{SENS} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{1}$$

$$\text{SPEC} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{2}$$

$$\text{ACC} = \frac{(\text{TN} + \text{TP})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{3}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

4. Conclusions

In our article, the geographical origin identification of 462 *P. yunnanensis* samples from Kunming, Yuxi, Chuxiong, Lijiang, Dali, and Honghe were analyzed by rhizome and leaf FT-MIR spectra, combined with PLS-DA, RF, and HCA methods. The chemical information differences between rhizomes and leaves were directly displayed on the FT-MIR spectra and the results of models. PLS-DA was more suitable for use in classifying the different geographical origins of *P. yunnanensis* than the RF model, in that it had the best identification ability and more uniformly distributed important variables. Besides, the order of classification ability from strong to weak is the data fusion dataset > leaves dataset > rhizomes dataset, which means that leaves can be used quickly and accurately to identify the geographical origin of *P. yunnanensis*, and more comprehensive information can be showed by multiple sources of chemical information.

Supplementary Materials: The following are available online. Figure S1: VIP scores of FT-MIR data of leaves for regional differences, Figure S2: The ntree and mtry screening of RF models of *P. yunnanensis* samples before variables are ranked by permutation accuracy importance, Figure S3: The 10-fold cross validation error rates of the RF model (sequentially reduced each five variables), based on *P. yunnanensis* samples, Figure S4: The ntree and mtry screening of RF models of *P. yunnanensis* samples after variables are ranked by permutation accuracy importance, Figure S5: The n_{tree} and m_{try} screening of RF models of *P. yunnanensis* samples before variables are ranked by permutation accuracy importance, Figure S6: The 10-fold cross validation error rates of the RF model (sequentially reduced each five variables) based on *P. yunnanensis* samples, Figure S7: The n_{tree} and m_{try} screening of RF models of *P. yunnanensis* samples after variables are ranked by permutation accuracy importance.

Author Contributions: Y.-F.P. and Y.-Z.W. developed the concept of the manuscript, Y.-F.P. performed the experiments, analyzed the data, and discussed the results, Q.-Z.Z. and Z.-T.Z. performed final corrections for this manuscript.

Funding: This work was sponsored by the National Natural Science Foundation of China (Grant number: 81460584).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cunningham, A.B.; Brinckmann, J.A.; Bi, Y.F.; Pei, S.J.; Schippmann, U.; Luo, P. *Paris* in the spring: A review of the trade, conservation and opportunities in the shift from wild harvest to cultivation of *Paris polyphylla* (Trilliaceae). *J. Ethnopharmacol.* **2018**, *222*, 208–216. [CrossRef] [PubMed]
2. Lu, H.; Xu, J.H.; Chen, R.P.; Yang, H.; Liu, Y.P. Status of the genus *Paris* L. re-sources of Yunnan and countermeasures for protection. *J. Yunnan Univ.* **2006**, *28*, 307–310.
3. State Pharmacopoeia Commission. *Chinese Pharmacopoeia*; Chemistry and Industry Press: Beijing, China, 2015.
4. Li, H. *The Genus Paris (Trilliaceae)*; Science Press: Beijing, China, 1998; pp. 12–16.
5. Yang, J.; Wang, Y.H.; Li, H. *Paris* qiliangiana (Melanthiaceae), a new species from Hubei, China. *Phytotaxa* **2017**, *329*, 193–196. [CrossRef]
6. Qin, X.J.; Yu, M.Y.; Ni, W.; Yan, H.; Chen, C.X.; Cheng, Y.C.; Li, H.; Liu, H.Y. Steroidal saponins from stems and leaves of *Paris polyphylla* var. *yunnanensis*. *Phytochemistry* **2016**, *121*, 20–29. [CrossRef] [PubMed]
7. Dai, X.W.; Feng, L.L.; Li, H.F. Analysis of differences and correlation of steroidal saponins in rhizomes and leaves of *Paris polyphylla* var. *yunnanensis* from different planting base. *Chin. J. Exp. Tradit. Med. Form.* **2018**, *24*, 41–48.
8. Yang, Y.G.; Zhang, J.; Zhao, Y.L.; Zhang, J.Y.; Wang, Y.Z. Quantitative determination and evaluation of *Paris polyphylla* var. *yunnanensis* with different harvesting times using UPLC-UV-MS and FT-IR spectroscopy in combination with partial least squares discriminant analysis. *Biomed. Chromatogr.* **2017**, *31*. [CrossRef] [PubMed]

9. Yang, Y.G.; Jin, H.; Zhang, J.; Zhang, J.Y.; Wang, Y.Z. Quantitative evaluation and discrimination of wild *Paris polyphylla* var. *yunnanensis* (Franch.) Hand.-Mazz from three regions of Yunnan Province using UHPLC-UV-MS and UV spectroscopy couple with partial least squares discriminant analysis. *J. Nat. Med.* **2017**, *71*, 148–157. [CrossRef] [PubMed]
10. Chen, T.Z.; Wen, F.Y.; Zhang, T.; Yang, Y.X.; Fang, Q.M.; Zhang, H.; Xue, D. Evaluation of saponins in *Paris Polyphylla* var. *chinensis* from twenty-one growing areas. *Chin. Tradit. Patent Med.* **2017**, *39*, 2345–2350.
11. Wu, X.M.; Zhang, Q.Z.; Wang, Y.Z. Traceability of wild *Paris polyphylla* Smith var. *yunnanensis* based on data fusion strategy of FT-MIR and UV-Vis combined with SVM and random forest. *Spectrochim. Acta A* **2018**, *205*, 479–488. [CrossRef] [PubMed]
12. Pei, Y.F.; Wu, L.H.; Zhang, Q.Z.; Wang, Y.Z. Geographical traceability of cultivated *Paris polyphylla* var. *yunnanensis* using ATR-FTMIR spectroscopy with three mathematical algorithms. *Anal. Methods* **2018**. [CrossRef]
13. Gad, H.A.; El-Ahmady, S.H.; Abou-Shoer, M.I.; Al-Azizi, M.M. Application of chemometrics in authentication of herbal medicines: A review. *Phytochem. Anal.* **2012**, *24*, 1–24. [CrossRef] [PubMed]
14. Biancolillo, A.; Marini, F. Chemometrics applied to plant spectral analysis. In *Vibrational Spectroscopy for Plant Varieties and Cultivars Characterization*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 69–104.
15. Li, J.; Zhang, J.; Zhao, Y.L.; Huang, H.Y.; Wang, Y.Z. Comprehensive quality assessment based specific chemical profiles for geographic and tissue variation in *Gentiana rigescens* using HPLC and FTIR method combined with principal component analysis. *Front. Chem.* **2017**, *5*. [CrossRef] [PubMed]
16. Qi, L.M.; Liu, H.G.; Li, J.Q.; Li, T.; Wang, Y.Z. Feature fusion of ICP-AES, UV-Vis and FT-MIR for origin traceability of *Boletus Edulis* mushrooms in combination with chemometrics. *Sensors* **2018**, *18*, 241. [CrossRef] [PubMed]
17. Li, Y.; Zhang, J.Y.; Wang, Y.Z. FT-MIR and NIR spectral data fusion: A synergetic strategy for the geographical traceability of *Panax notoginseng*. *Anal. Bioanal. Chem.* **2018**, *410*, 91–103. [CrossRef] [PubMed]
18. Wang, Y.; Huang, H.Y.; Zuo, Z.T.; Wang, Y.Z. Comprehensive quality assessment of *Dendrubium officinale* using ATR-FTIR spectroscopy combined with random forest and support vector machine regression. *Spectrochim. Acta A* **2018**, *205*, 637–648. [CrossRef] [PubMed]
19. Yang, Y.G.; Wang, Y.Z. Characterization of *Paris polyphylla* var. *yunnanensis* by infrared and ultraviolet spectroscopies with chemometric data fusion. *Anal. Lett.* **2018**, *51*, 1730–1742. [CrossRef]
20. Xie, J.D.; Sun, L. An overall quality evaluation of Paridis Rhizoma by multiple components determination based on the chemometrics. *Chin. J. Pharm. Anal.* **2015**, *35*, 1585–1590.
21. Zhang, S.S.; Liu, X.; Wang, J.F.; Yu, M.J.; Huang, Z.J.; Liu, Y.; Zhang, H. Determination of seven steroidal saponins in Paridis Rhizoma and polygerm varieties from different regions in Yunnan Province by UPLC and establishment of fingerprint. *Chin. Tradit. Herb. Drugs* **2016**, *47*, 4257–4263.
22. Zhang, J.Y.; Wang, Y.Z.; Zhao, Y.L.; Yang, S.B.; Zhang, J.; Yuan, T.J.; Wang, J.J.; Jin, H. Ultraviolet absorption spectrum analysis and identification of medicinal plants of *Paris*. *Spectrosc. Spectr. Anal.* **2012**, *32*, 2176–2180.
23. Sun, S.Q.; Zhou, Q.; Chen, J.B. *Analysis of Traditional Chinese Medicine by Infrared Spectroscopy*; Chemical Industry Press: Beijing, China, 2010.
24. Yang, L.F.; Ma, F.; Zhou, Q.; Sun, S.Q. Analysis and identification of wild and cultivated Paridis Rhizoma by infrared spectroscopy. *J. Mol. Struct.* **2018**, *1165*, 37–41. [CrossRef]
25. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometr. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
26. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef] [PubMed]
27. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
28. Ståle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemometr.* **1987**, *1*, 185–196.
29. Indahl, U.G.; Martens, H.; Naes, T. From dummy regression to prior probabilities in PLS-DA. *J. Chemometr.* **2007**, *21*, 529–536. [CrossRef]
30. Nocairi, H.; Qannari, E.M.; Vigneau, E.; Bertrand, D. Discrimination on latent components with respect to patterns. Application to multicollinear data. *Comput. Stat. Data Anal.* **2005**, *48*, 139–147. [CrossRef]
31. Breiman, L. Random forests. *Mach Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Amjad, A.; Ullah, R.; Khan, S.; Bilal, M.; Khan, A. Raman spectroscopy based analysis of milk using random forest classification. *Vib. Spectrosc.* **2018**, *99*, 124–129. [CrossRef]

33. Saptoro, A.; Tadé, M.O.; Vuthaluru, H. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chem. Prod. Process Model.* **2012**, *7*, 1–14. [CrossRef]
34. Rajer-Kanduč, K.; Zupan, J.; Majcen, N. Separation of data on the training and test set for modelling: A case study for modelling of five colour properties of a white pigment. *Chemometr. Intell. Lab. Syst.* **2003**, *65*, 221–229. [CrossRef]
35. Jain, A.K.; Dubes, R.C. Algorithms for clustering data. In *Technometrics*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, USA, 1988.
36. Barnes, R.J.; Dhanoa, M.S.; Lister, S.J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777. [CrossRef]
37. Savitzky, A.; Golay, M.J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
38. Xie, L.J.; Ye, X.Q.; Liu, D.H.; Ying, Y.B. Quantification of glucose, fructose and sucrose in bayberry juice by NIR and PLS. *Food Chem.* **2009**, *114*, 1135–1140. [CrossRef]
39. Qi, L.M.; Zhang, J.; Liu, H.G.; Li, T.; Wang, Y.Z. Fourier transform mid-infrared spectroscopy and chemometrics to identify and discriminate *Boletus edulis* and *Boletus tomentipes* mushrooms. *Int. J. Food Prop.* **2017**, *20*, S56–S68. [CrossRef]

Sample Availability: Samples of the compounds are not available from the authors.

molecules

MDPI

Article

ATR-FTIR Spectroscopy, a New Non-Destructive Approach for the Quantitative Determination of Biogenic Silica in Marine Sediments

Dora Melucci [1,*], Alessandro Zappi [1], Francesca Poggioli [1], Pietro Morozzi [1], Federico Giglio [2] and Laura Tositti [1]

[1] Department of Chemistry “G. Ciamician”, University of Bologna, 40126 Bologna, Italy; alessandro.zappi4@unibo.it (A.Z.); francesca.poggioli3@studio.unibo.it (F.P.); pietro.morozzi2@unibo.it (P.M.); laura.tositti@unibo.it (L.T.)

[2] Polar Science Institute-National Research Council ISP-CNR, Via P. Gobetti 101, 40129 Bologna, Italy; federico.giglio@cnr.it

* Correspondence: dora.melucci@unibo.it; Tel.: +39-051-209-9530

Received: 30 August 2019; Accepted: 29 October 2019; Published: 31 October 2019

Abstract: Biogenic silica is the major component of the external skeleton of marine micro-organisms, such as diatoms, which, after the organisms death, settle down onto the seabed. These micro-organisms are involved in the CO_2 cycle because they remove it from the atmosphere through photosynthesis. The biogenic silica content in marine sediments, therefore, is an indicator of primary productivity in present and past epochs, which is useful to study the CO_2 trends. Quantification of biosilica in sediments is traditionally carried out by wet chemistry followed by spectrophotometry, a time-consuming analytical method that, besides being destructive, is affected by a strong risk of analytical biases owing to the dissolution of other silicatic components in the mineral matrix. In the present work, the biosilica content was directly evaluated in sediment samples, without chemically altering them, by attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy. Quantification was performed by combining the multivariate standard addition method (MSAM) with the net analyte signal (NAS) procedure to solve the strong matrix effect of sediment samples. Twenty-one sediment samples from a sediment core and one reference standard sample were analyzed, and the results (extrapolated concentrations) were found to be comparable to those obtained by the traditional wet method, thus demonstrating the feasibility of the ATR-FTIR-MSAM-NAS approach as an alternative method for the quantification of biosilica. Future developments will cover in depth investigation on biosilica from other biogenic sources, the extension of the method to sediments of other provenance, and the use higher resolution IR spectrometers.

Keywords: diatoms; biogenic silica; ATR-FTIR; chemometrics; NAS

1. Introduction

In the present work, we introduce an innovative and non-destructive method for the quantification of biogenic silica in marine sediments through the use of infrared spectroscopy combined with chemometrics. The proposed method was applied to sediment samples coming from Terra Nova Bay, Antarctica.

Antarctica is a unique natural laboratory because it is the coldest, driest, highest, windiest, and most isolated continent. Therefore, it is almost unaffected by anthropogenic influence [1]. The Southern Ocean allows the diffusion of atmospheric carbon dioxide into the deep sea, which is partially used by sea plants for growth and for the production of organic matter [2]. Therefore, this region is one of the most important for the study of climate changes and conditions of the ocean [3].

In particular, an important tool to control the chemical composition of seawater and to reconstruct paleo-ocean conditions is represented by marine sediments, which are a reservoir and a sink of chemical species involved and cycled in the marine food chain [1]. Among nutrients, silicon is an essential element in the ocean ecosystems, because it is responsible for the growth of Radiolaria, Sponges, Phaeodaria, and particularly Diatoms, which represent a major portion of planktonic primary producers [4]. Diatoms are planktonic unicellular microalgae, known to form an external skeleton called frustule, constituted by amorphous silica and organic components (usually including long-chain polyamines and silaffins) [5,6]. After their death, the diatom siliceous skeleton settles down through the water column. The extent of diatom deposition in the sediments will be a function of the sea bottom depth and of the degree of solubilization of opal silica in the water column [7]. Siliceous microfossils, therefore, can represent a large part of the mass of biogenic sediments accumulating on the deep-sea floor [8].

In the whole Southern Ocean, the Ross Sea is the region of the most widely extensive algal blooms, usually initiating in the Ross Sea polynya [9], an ice-free area of enhanced bio-productivity that can be considered as a biological "hot spot" compared with the surrounding waters. This area extends to the open sea surface as soon as the austral summer develops and the sea ice melts [10,11]. It plays a key climatic role on a global scale. Indeed, the Ross Sea is one of the main sink areas for the tropospheric CO_2, widely contributing to counterbalancing its budget and the associated role in climate change [12,13]. In the western Ross Sea, the polynya of Terra Nova Bay (TNB) is an area of high accumulation of biogenic silica in the sediments [14,15].

Biogenic silica (BSi) content in marine sediment can be considered as a good proxy to characterize the bio-productivity of the Southern Ocean [16,17]. However, the quantification of BSi is complicated by the presence of lithogenic silica, which is chemically equivalent to BSi (SiO_2), with the only difference being crystalline, while BSi is amorphous. Several methods have been proposed to estimate BSi in marine sediments: (1) X-ray diffraction after the conversion of opal to cristobalite at a high temperature [18]; (2) direct X-ray diffraction of amorphous silica [19]; (3) direct infrared spectroscopy of amorphous opal [20]; (4) elemental partitioning of sediment chemistry [21,22]; (5) microfossil counts [23,24]; and (6) several wet-alkaline extraction methods [7,25,26].

Among the above-mentioned techniques, the wet alkaline methods are the most popular because they are the most sensitive techniques for BSi assessment. According to these methods, BSi is extracted and distinguished from lithogenic silica based on hot alkaline solutions [7]. Wet methods exploit a different rate of dissolution of lithogenic and biogenic silica in alkaline solution, with BSi dissolving more quickly than the mineral component. Solubilized BSi can, therefore, be collected in the supernatant of the solution, and subsequently determined by spectrophotometry. Such separations are extremely demanding and time-consuming, and above all, they do not ensure the quantitative recovery of BSi, owing to inherent systematic problems; that is, dependence on matrix effects, incomplete opal recovery, and contamination by non-biogenic silica [17,23].

The increasing success of chemometric tools applied to basic spectrophotometric techniques such as Fourier transform infrared (FTIR), together with the compelling need for understanding key biogeochemical processes of global importance, have recently inspired the introduction of an alternative approach to solve the problem of BSi assessment. In particular, FTIR spectroscopy has been applied to lacustrine sediments for the analysis of silica and other minerals by Rosén et al. [27,28]. Vogel et al. and Rosén et al. [29,30] showed that FTIR spectroscopy in the mid-infrared region is highly sensitive to chemical components present in minerogenic and organic material, such as sediments; this fact provides an efficient tool for quali- and quantitative characterization of these fundamental, but complex environmental matrices.

Moreover, a method based on attenuated total reflectance (ATR)-FTIR measurements has also been proposed in the literature to quantify inorganic components in marine sediments [31,32]. ATR-FTIR spectroscopy is particularly appealing for the analysis of sediments because no chemical sample pre-treatment is required: it may in principle by-pass all the drawbacks of the wet-chemical method;

moreover, it works with small amounts of sample material (0.05–0.1 g, dry weight) and it is rapid, inexpensive, and efficient. Besides, ATR-FTIR is a non-destructive method, allowing to recover the sample for further analyses, and it can be carried out even off the lab.

In the present work, we developed and present an analytical method based on ATR-FTIR for the quantitative determination of biosilica content in marine sediments. The feasibility of the method was evaluated by quantifying BSi in a series of sediment samples collected in the Ross Sea. Optimization of the experimental procedures such as the drying process, homogenization, and deposition of the sample on the ATR crystal are discussed in detail, in order to provide a reliable background useful to solve reproducibility problems, which may constitute a drawback of such a simple instrumental approach. Furthermore, the strong matrix effect intrinsic to environmental samples is faced and solved by applying a multivariate standard addition method (MSAM) [33], improved by net analyte signal (NAS) computation [34,35].

2. Results and Discussion

2.1. ATR Spectra

For each of the 22 analyzed sediment samples (21 coming from Mooring D and one 53%$_{w/w}$ reference standard), four standard-added (add.x) samples were prepared: the zero-added sample (add.0) is the pure sample, add.1 has an added concentration of diatomite at 5%$_{w/w}$, add.2 at 10%$_{w/w}$, and add.3 at 15%$_{w/w}$. All added samples (and a pure diatomite sample) were analyzed by ATR-FTIR.

In the ATR spectra of marine sediments, the contribution of silica, both biogenic and lithogenic, is dominant. Such spectra exhibit four characteristic vibrational bands. The two main bands at 1100 and 471 cm^{-1} are attributed to triply degenerated stretching and bending vibration modes, respectively, of the [SiO_4] tetrahedron [36]. The band at 800 cm^{-1} corresponds to an inter-tetrahedral Si–O–Si bending vibration mode, and the band near 945 cm^{-1} to an Si–OH vibration mode [37]. Previous studies have shown that the absorbance centered around 1640 cm^{-1} and between 3000 and 3750 cm^{-1} can be attributed to hydroxyl vibrations because hydroxyl ions are major constituents of clay minerals, opal, and organic compounds present in marine sediments [38]. However, these bands are not specific for silica, their intensity is generally low (about one-tenth of the main band); moreover, they are overlapped with the residual absorption bands of H_2O. Therefore, to reduce the noise in the spectral data acquired, we decided to discard the IR region between 4000 and 1300 cm^{-1} and to apply the chemometric procedure only in the region between 1300 and 400 cm^{-1}.

Figure 1 shows the raw spectra of sample D10 (as an example of the spectra obtained for all sediment samples) and the replicates of a pure-diatomite sample. In Figure 1a, spectra obtained by instrumental analysis are shown, while Figure 1b highlights the effect of the spectral pre-treatments: uninformative-band removal and MSC.

On the spectra reported in Figure 1b, the two chemometric procedures described in Section 3.4 were carried out for all sediment samples, and the results are reported in Table 1. The expected values reported in Table 1 are the BSi concentrations obtained by wet analyses that were carried out only on five sediment samples (and on 53%$_{w/w}$-standard): D4, D6, D9, D18, and D21.

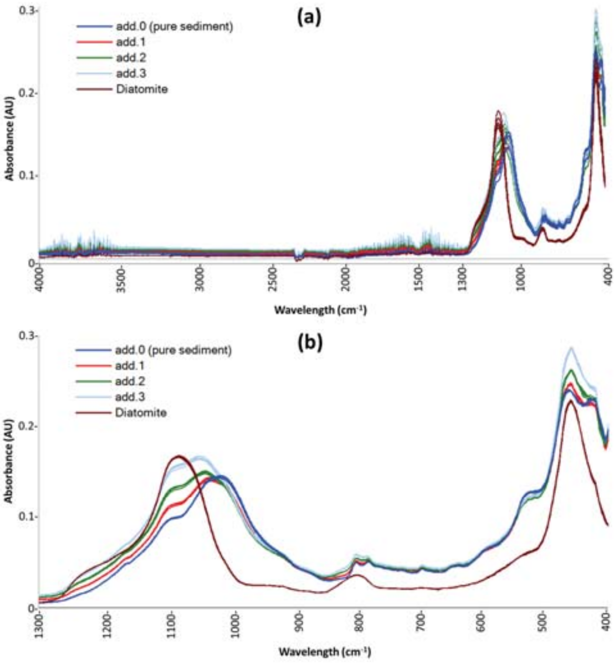


Figure 1. (**a**) ATR raw spectra of sample D10, as obtained from the spectrophotometer; (**b**) the same spectra after band removal (discarding the IR range 4000–1300 cm^{-1}) and after multiplicative scatter correction (MSC) pre-treatment. "add." in the legends indicates standard added samples.

Table 1. Net analyte signal (NAS) results for the two pre-processing methods. All the numbers are formatted with three significant digits to allow for a detailed comparison. LoD, limit of detection.

		Procedure 1				*Procedure 2*			
SAMPLE CODE	**Expected Value ± Standard Deviation (%$_{w/w}$)**	**NAS Extr. C (%$_{w/w}$)**	**Standard Deviation**	**R^2**	**LoD**	**NAS Extr. C (%$_{w/w}$)**	**Standard Deviation**	**R^2**	**LoD**
Std 53%	53 ± 3	53.6	6.02	0.992	1.01	50.2	2.74	0.988	7.76
D0	-	3.23	0.903	0.996	0.0265	8.64	3.01	0.956	2.39
D1	-	3.24	1.92	0.993	0.190	-	-	-	-
D2	-	9.88	1.70	0.994	0.416	9.28	1.23	0.990	6.24
D3	-	7.55	0.315	0.993	0.743	8.40	0.436	0.999	4.08
D4	13 ± 2	12.0	1.16	0.988	0.469	12.8	2.01	0.988	3.58
D5	-	5.41	1.10	0.993	0.0214	5.38	1.75	0.995	2.88
D6	14 ± 2	14.2	1.54	0.989	0.00946	14.2	2.71	0.997	1.41
D7	-	5.87	1.27	0.999	0.646	-	-	-	-
D8	-	9.45	0.671	0.993	0.380	-	-	-	-
D9	8 ± 1	8.26	3.16	0.986	0.654	8.91	0.514	0.982	4.68
D10	-	12.0	1.18	0.997	0.515	12.1	2.12	0.995	0.267
D11	-	2.94	0.646	0.994	0.0666	3.84	0.803	0.993	1.76
D12	-	-	-	-	-	10.6	2.32	0.995	2.32
D13	-	-	-	-	-	3.72	1.81	0.987	3.82
D14	-	9.00	0.146	0.998	0.133	10.9	0.433	0.998	7.62
D15	-	3.25	0.184	0.993	0.0784	5.19	0.122	0.996	2.22
D16	-	-	-	-	-	5.63	1.38	0.989	3.60
D17	-	2.71	0.535	0.994	0.0647	13.7	1.41	0.984	1.93
D18	4.2 ± 0.6	4.80	1.67	0.996	0.0900	5.16	1.70	0.992	4.34
D19	-	4.04	0.332	0.998	0.0459	3.22	0.345	0.997	1.23
D20	-	2.36	0.535	0.999	0.610	4.26	1.19	0.991	8.46
D21	3.4 ± 0.5	3.12	1.86	0.998	0.653	4.12	0.296	0.993	3.93

2.2. *Results: Procedure 1*

To assess the reliability of the new methodology proposed here, the BSi results obtained here can be compared to those achieved through the traditional wet method, taken as a reference. Table 1 shows that the results obtained with *Procedure 1* (band removal and MSC) are in good agreement with the expected values obtained when samples were analyzed with the wet method. Indeed, the confidence intervals obtained for these five sediment samples by NAS are not significantly different from the ones obtained by the wet method (at 0.05 significance). Also, the result obtained for the standard sample (Std 53%$_{w/w}$) is in agreement with the expected concentration. The coefficients R^2 of the NAS standard addition lines are all higher than 0.98, indicating a good correlation between added concentrations (dependent variable) and the pseudo-univariate NAS values calculated by the chemometric procedure. The LoD values are also very good, being, in general, in the order of magnitude of one-tenth (or even lower) with respect to the corresponding extrapolated BSi concentration.

Moreover, Frignani et al. [15] reported that BSi concentration in surface sediments in this area is usually relatively low, <10%$_{w/w}$; the results obtained by NAS are in agreement with this consideration, as D0, D1, D2, and D3 extrapolated concentrations are lower than the indicated value.

All these considerations confirm that NAS applied to ATR spectra of the standard added samples can be a valuable and reliable alternative to the time-consuming wet method for the quantification of BSi in marine sediments.

The main drawbacks concern the three samples for which no results were obtained: D12, D13, and D16. In these cases, the NAS standard addition lines had, for all PLS-factors, either a negative slope or intercept, giving negative extrapolated values, or not acceptable R^2 (lower than 0.7), that make any possible result unreliable. The reason for such behavior is still under study, but we can hypothesize that there is still some source of noise in ATR analysis that was not taken into account, although several precautions were taken during instrumental analyses, as described in Section 3.3. We, therefore, decided to proceed with further chemometric assessments, also to test the hypothesis of a possible defect in the NAS procedure.

2.3. Results: Procedure 2

As described in Section 3.4, a variable selection was carried out on baseline-corrected spectra. Correlation loadings on PLS-factor 1 were used to select the most important variables to describe the regression model. Although a different variable selection was carried out for each sediment sample, not always giving the same variables, a general description of the selected IR bands can be drawn and is resumed in Figure 2. High correlation loading values in the PLS-factor 1 are computed in the regions of 1260–1060 cm^{-1}, 830–800 cm^{-1}, and 467–436 cm^{-1}. These regions of the IR spectra correspond to the characteristic SiO_2 absorbance maxima as reported by Vogel et al. [29]. On these selected variables, NAS computation was carried out and the results are reported in the last vertical section of Table 1.

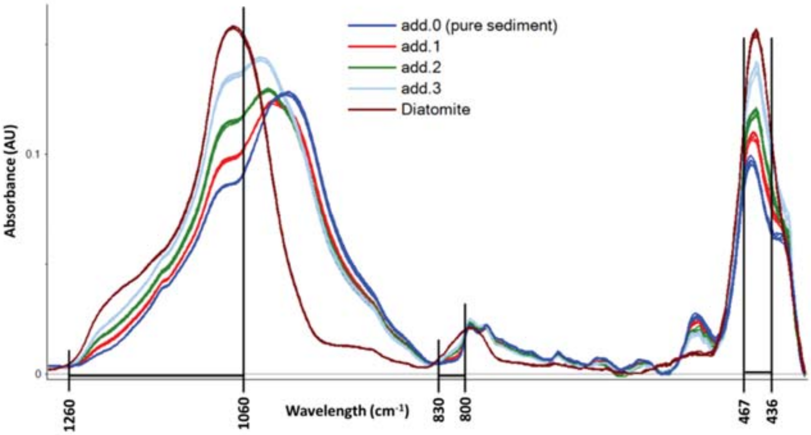


Figure 2. Baseline-corrected ATR spectra of sample D10. Black lines indicate the variables considered most important by partial least square (PLS) correlation loadings.

Again, concentrations extrapolated by NAS are not significantly different from the "wet method-based" values, with high R^2 (>0.95). After the application of this second procedure (baseline correction and variable selection before NAS), significant differences were detected only for D0 and D17, which, in this case, also have a lower R^2 compared with the other samples. In this case, some problems arise from LoDs, which, in most cases, are comparable to the extrapolated value (and also higher than that for D20). Such a drawback might, therefore, be because of the spectral pre-treatment; in order to calculate LoD, a blank spectrum is necessary. However, such a blank spectrum has to be pre-treated as all the other spectra, and in this case, it has to be baseline corrected. In this way, the pre-treatment can likely produce some spikes in the blank spectrum (that is, a noisy signal oscillating around the zero), thus affecting the computation of LoD.

The three samples that did not give results with the computation by *Procedure 1* (D12, D13, and D16) in this case have an acceptable extrapolated concentration. However, there are again three samples (D1, D7, and D8) with no result. This strengthens the hypothesis of the presence of a noise source that was not taken into account. Indeed, variable selection may reduce the noise present in the whole spectrum, but, at the same time, if noise is present in the selected variables, its effect may be enhanced. Therefore, the two chemometric methods presented in this work may be considered to be complementary for this study.

3. Materials and Methods

3.1. Study Area

Sediment samples for the present study were collected in "mooring D" (or "site D"), which is located in Antarctica, in the western sector of the Ross Sea continental shelf within the polynya of Terra Nova Bay at 75°06′ S and 164°28′5′′ E (Figure 3). The box-core, from which the sediments were collected, was sampled at a depth of 972 m during the 2003–2004 Italian PNRA (Programma Nazionale di Ricerca in Antartide) Campaign [39], whose basis was situated in the "Mario Zucchelli" station.

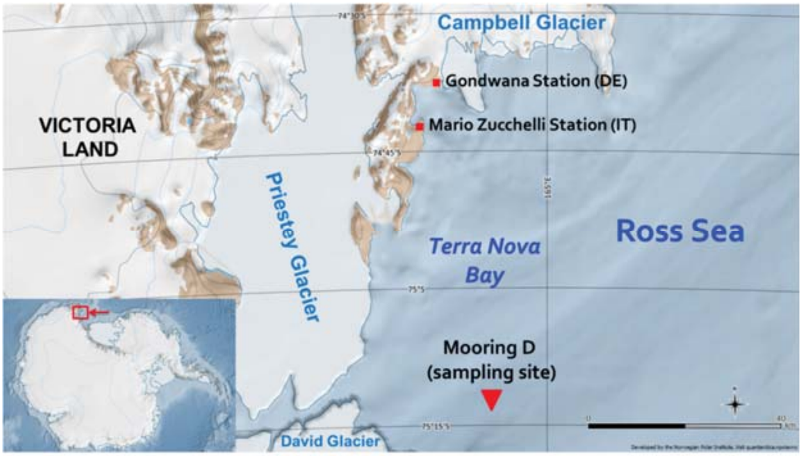


Figure 3. Sampling site (Mooring D) in Terra Nova Bay, Antarctica.

In the Ross Sea, surface sediments are generally composed of unsorted ice-rafted debris, terrigenous silts and clays, and siliceous and calcareous biogenic debris [40]. In site D, in particular, coarse terrigenous deposits are predominant, owing to the proximity of Priestley, David, and Campbell glaciers [15].

3.2. Samples

The sediment collected in site D was sampled using a 1T Oceanic box corer. A sub core 22 cm long was collected by means of a polyvinyl chloride (PVC) liner. The short core was subsampled with a resolution of 1 cm [39]. Twenty-two sediment sections were thus obtained and named with a two-digit code: a letter, "D", indicating the sampling place; and a number, from 0 to 21, indicating the core height, with D0 being the top, corresponding to the sediment surface. Sediments were then stored at −21 °C in a polycarbonate Petri capsule and oven-dried at 50 °C just prior to the analyses. The BSi content of five of these samples was also quantified by a wet method analysis, according to the DeMaster method [7,15], thus providing some comparison values for the ATR analyses. In the absence of a commercial certified reference material for BSi, the "internal reference standard" used in the Polar Science Institute-National Research Council (CNR-ISP) laboratory was adopted for the purpose of this paper. This sample consists of an Antarctic marine sediment analyzed repeatedly both in CNR and other biogeochemical laboratories, resulting in a BSi content of (53 ± 3)$\%_{w/w}$ and a remaining 47$\%_{w/w}$ of alkaline halide.

For the sake of readability, a flowchart concerning the sample preparation is reported in Figure 4. Before sample preparation, all samples were manually ground in an agate mortar for approximately 15 min and heated in a ventilated oven at 105 °C for 1 h to remove atmospheric moisture. Afterward, each sample was split into four aliquots, three of which were added with known amounts (5$\%_{w/w}$, 10$\%_{w/w}$, 15$\%_{w/w}$) of Diatomite (Celite® 545 AW, Sigma-Aldrich, Darmstadt, Germany), in order to apply the multivariate standard addition method. The total weight of each standard-added sample was 200 mg. Diatomite was chosen as a proxy of standard biogenic silica, because it is composed of frustulae of biogenic silica, similar to what we wanted to quantify in marine sediments. Such a similarity was visually evaluated by analyzing some samples with a scanning electron microscope (SEM) Philips 515B (Philips, Amsterdam, Netherlands), equipped with an EDAX DX4 microanalytical device (EDAX Inc., Mahwah, NJ, USA). Figure 5 shows the pictures obtained by SEM. From Figure 5c,d, it can be seen that samples D1 and D4 contain the same radiolaria present in the Diatomite (Figure 5a) used as a proxy of BSi.

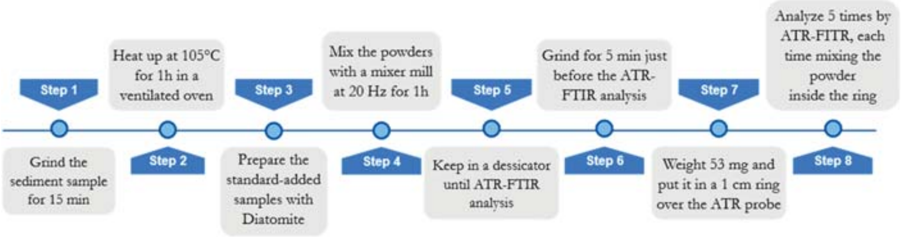

Step 1
Grind the sediment sample for 15 min
Step 2
Heat up at 105°C for 1h in a ventilated oven
Step 3
Prepare the standard-added samples with Diatomite
Step 4
Mix the powders with a mixer mill at 20 Hz for 1h
Step 5
Keep in a dessicator until ATR-FTIR analysis
Step 6
Grind for 5 min just before the ATR-FTIR analysis
Step 7
Weight 53 mg and put it in a 1 cm ring over the ATR probe
Step 8
Analyze 5 times by ATR-FITR, each time mixing the powder inside the ring

Figure 4. Sample preparation flowchart. ATR-FTIR, attenuated total reflection Fourier transform infrared.

To ensure better homogenization of the powders, a Mixer Mill "MM20" (Retsch Inc., Düsseldorf, Germany) was used. Each added sample was placed in stainless steel cylinders of 1.5 mL volume and left in the ball mill for 60 min at 20 Hz. Before the instrumental analysis, samples were kept in a desiccator filled with silica gel to prevent the absorption of atmospheric moisture. Standard added samples were then analyzed by ATR-FTIR spectroscopy.

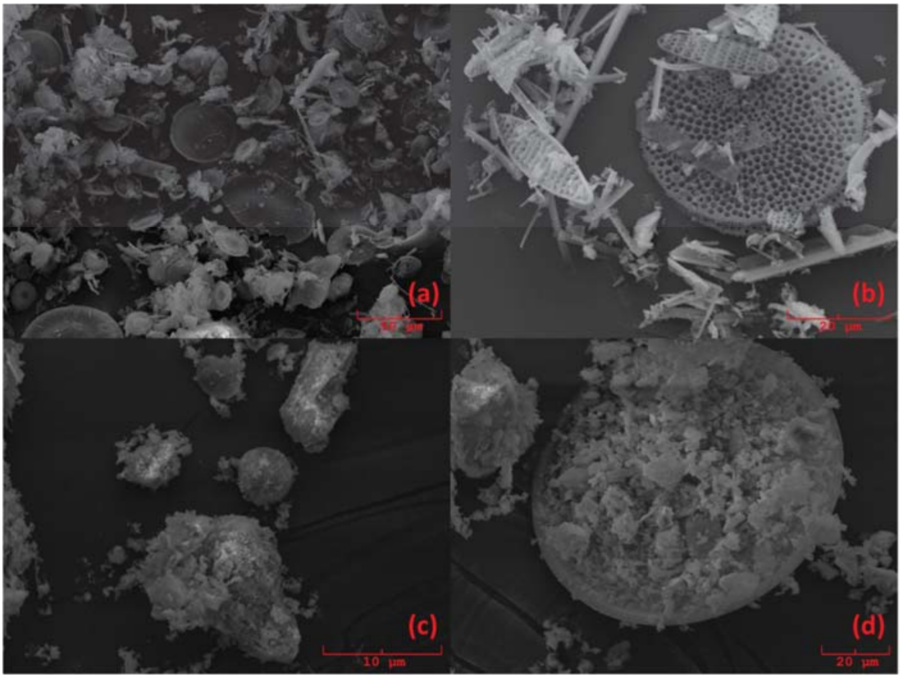


Figure 5. Scanning electron microscope (SEM) images of samples (**a**) pure diatomite; (**b**) 53% standard; (**c**) D1 sample, which is characterized by the presence of both radiolaria and bulks of sedimentary material; and (**d**) D4 sample.

3.3. ATR-FTIR Analysis

Attenuated total reflection spectra were collected using a Bruker ALPHA FT-IR spectrometer (Brucker Optics GmbH, Billerica, MA, USA) equipped with a single-reflection diamond ATR accessory (Bruker Platinum ATR, Billerica, MA, USA) with an approximately 0.6 mm × 0.6 mm active area and a mercury–cadmium–telluride detector. Spectra were collected in the mid-IR range, 400–4000 cm^{-1}, with an optical resolution of 4 cm^{-1}; the registered spectrum is the mean of 64 scans, executed in 3 min. For each sample aliquot, five replicate spectra were recorded to assess precision and ensure the reproducibility of each sample. All measurements were performed at ambient conditions. Before spectra acquisition, a background spectrum (air) was collected with the same operational parameters. Such a background was automatically subtracted to each sample spectrum.

To optimize the analytical reproducibility, some precautions were taken for ATR analysis. Indeed, it is widely reported in the literature how an imprecise sample preparation (especially drying process), sample deposition, and instrumental calibration may cause poor instrumental repeatability and accuracy, fundamental characteristics for quantification purposes [31,41]. For these reasons, a suitable experimental protocol was developed and evaluated (Figure 4).

Before the instrumental analysis, samples were manually ground again for 5 min in an agate mortar, in order to homogenize powder granulometry. Moreover, for each added sample, an aliquot (53 mg) was carefully weighted and lodged in a steel ring of 1 cm in diameter, which was then placed over the spectrophotometer probe. The same amount of material was taken for all the analyzed samples, to maximize reproducibility and reduce scattering and other problems resulting from not optimal (or not constant) contact between the sample and crystal. These problems become relevant in ATR analysis when used for quantification purposes owing to the geometry of ATR irradiation and reflectance, which need an accurate evaluation and the adoption of a suitable experimental protocol [42].

3.4. Chemometrics

Prior to chemometric analysis, the five replicated spectra of each added sample were pre-treated by multiplicative scatter correction (MSC) [43]. MSC allows the reduction of the effects of scattering noise on IR spectra, increasing the reproducibility of sample replicates.

Subsequently, in order to calculate the BSi content in each sediment sample by MSAM, the NAS procedure was applied [35,44]. NAS is a mathematical procedure that allows extracting, from a multivariate signal (in this case, an ATR spectrum), that part of the signal that is only due to the analyte, removing the other signals due to the other interfering species present in the matrix [35]. In this way, the multivariate problem can be reduced to a pseudo-univariate problem, whose results can be obtained by a univariate treatment. NAS computations were performed as follows.

The NAS procedure starts from a partial least square (PLS) regression [45], using the ATR spectra as independent variables (X) and the added concentrations vector as dependent one (y). The best PLS-factor (A) has to be selected and the corresponding PLS-regression coefficient vector (b_A) is used to compute a projection matrix (H) as follows:

$$H = b_A\left(b_A b_A^t\right)^{-1} b_A^t, \tag{1}$$

where t indicates transpose and superscript "−1" indicates matrix inversion. H matrix is then used to compute NAS vectors (x_i^*):

$$x_i^* = Hx_i, \tag{2}$$

where x_i are the rows of matrix X, which means samples of ATR spectra. Each calculated x_i* corresponds to the net signal (devoid of interfering signals) of each replicate of the added samples. The Euclidean norms of such net signals can be then used as pseudo-univariate signals to compute a univariate standard addition linear regression line, from which BSi concentration can be obtained by extrapolation.

The selection of the optimal PLS-factor (A) is a crucial point of the procedure, because, in most of the cases, the final extrapolated concentration varies (also dramatically) while varying A. Therefore, A was chosen (sample by sample) as the PLS-factor that optimizes both the PLS root mean squared error (RMSE), by minimizing it, and the determination coefficient (R^2) of the final pseudo-univariate line, by maximizing it. When these two conditions were not simultaneously achievable for one PLS-factor, A was chosen as the factor giving the best compromise between these two parameters, based on the highest R^2.

The standard deviations of the extrapolated concentration values were computed by the *jackknife* method [46]. Once the optimal PLS-factor is selected, the *jackknife* procedure replicates the NAS computation as many times as the number of objects (x_i), each time keeping out one object. In this way, i different extrapolated values are obtained for each NAS computation and the overall standard deviation is estimated as the standard deviation of the *jackknife*-extrapolated values.

Limits of detection (LoDs) were computed collecting five replicates (the same number of the other samples) of a blank spectrum (empty sample holder) and projecting them onto the NAS space by Equation (2) as if they were real samples [47]. The so obtained NAS-blank signals were mediated to obtain the vector ε, and the LoD was computed as follows [47]:

$$\text{LoD} = 3\frac{||\varepsilon||}{||b_A||}, \tag{3}$$

where $||\cdot||$ indicates the Euclidean norm.

The so far described procedure (*Procedure 1*) was applied to raw spectra, as they were obtained from the spectrophotometer. This procedure gave reasonable results for the majority of the samples, while it failed for three of them; in those cases, for all PLS-factors, the final NAS standard addition line had either a negative slope or intercept, producing a negative extrapolated concentration. The reason behind such behavior is still under evaluation. In order to obtain a result for each sediment sample, another chemometric procedure (*Procedure 2)* has thus been developed. Instead of using raw spectra, a baseline correction was applied directly by the software controlling the instrument, OPUS v.7.2 (Bruker). MSC was applied to baseline-corrected spectra and, before NAS computation, a variable selection was applied. For variable selection, another PLS regression was computed (previously to the one used for NAS). Only factor 1, always retaining more than 95% of the explained variance, was considered, and the variables giving correlation loadings [48] higher than 0.7 (in absolute value) were retained as important. NAS computation was then applied only with these selected variables. The standard deviations on the extrapolated values and LoDs were calculated as before.

MSC and variable selection pre-processing were performed by the software The Unscrambler v.10.3 (CAMO, Olso, Norway), while NAS and *jackknife* procedures were computed by a homemade code in R environment (R Core Team, Vienna, Austria).

4. Conclusions

In this study, we demonstrated the feasibility of a new approach for the quantification of biogenic silica based on IR spectroscopy coupled with chemometrics. Biogenic silica content in marine sediments from Terra Nova Bay in West Antarctica was evaluated with Fourier-transform infrared spectroscopy in attenuated total reflection mode (ATR-FTIR). For quantification, the multivariate standard addition method (MSAM) was applied, and the net analyte signal (NAS) procedure was used to solve the problems deriving from the strong matrix effect affecting such analyses.

Twenty-one subsequent core samples and one reference standard were analyzed. Reliable results were obtained, as observed from the comparison with homologous data from the traditional wet method.

Some drawbacks remain. The chemometric procedure did not give acceptable results for some samples, even if a variable selection was carried out. Moreover, the limits of detection, and perhaps also standard deviations, are in some cases still too high.

However, it has to be taken into account that the quantification of biosilica, in this work, has been carried out with an analytical technique (ATR) that has several intrinsic drawbacks when performing quantitative analysis. In particular, owing to the optical behavior of photons at such low angles as in ATR, extremely careful handling of samples and highly reproducible sample geometry are required when analyzing powdered samples. Moreover, the analyzed samples are powders, which, despite all the precautions taken before and during the analysis, can still have some problems concerning homogeneity and granulometry. The BSi content was also evaluated in natural samples without any chemical pre-treatment, thus its analytical signal may be strongly affected by the presence of lithogenic silica, besides all the other species composing the sediments.

Considering all these aspects, the analytical and chemometric procedure presented in this work, although requiring some more refinements, can be considered a promising alternative to the traditional time-consuming wet method for the quantification of biosilica in marine sediments. In paleolimnological research, the ATR-FTIR technique is seldom used. The results presented here, as well as the fact that this method is fast and cost-effective, requiring only small quantities of sediment sample, should encourage more researchers to use it. Moreover, marine sediments are precious samples, which are difficult to collect; thus, a not-destructive method would be preferable to analyze them, although ATR-FTIR cannot yet entirely replace conventional analytical tools in paleolimnology.

Author Contributions: Conceptualization, D.M., F.G., and L.T.; Methodology, D.M., A.Z., F.G., and F.P.; Software, A.Z. and F.P.; Validation, A.Z. and P.M.; Formal Analysis, F.P., A.Z., and P.M.; Investigation, F.P.; Resources, F.G. and L.T.; Data Curation, D.M., A.Z., and F.P.; Writing—Original Draft Preparation, A.Z. and F.P.; Writing—Review & Editing, D.M., A.Z., P.M., and L.T.; Visualization, F.P. and F.G.; Supervision, D.M.; Project Administration, L.T.

Funding: This study has been conducted with the financial support of the Italian National Research Program for Antarctica (PNRA).

Acknowledgments: We are indebted to the Italian participants to the XIX Expedition to Antarctica, in the year 2003/04, for carrying out the samplings. We are deeply grateful to Prof. Giorgio Gasparotto, of the Department of

Biological, Geological, and Environmental Sciences, University of Bologna for kindly providing the access to his SEM facility and the precious support in producing the diatom images employed herein.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Casalino, C.E.; Malandrino, M.; Giacomino, A.; Abollino, O. Total and fractionation metal contents obtained with sequential extraction procedures in a sediment core from Terra Nova Bay, West Antarctica. *Antarct. Sci.* **2013**, *25*, 83–98. [CrossRef]
2. Toggweiler, J.R.; Russell, J.L.; Carson, S.R. Midlatitude westerlies, atmospheric CO_2, and climate change during the ice ages. *Paleoceanography* **2006**, *21*. [CrossRef]
3. Sprenk, D.; Weber, M.E.; Kuhn, G.; Rosén, P.; Frank, M.; Molina-Kescher, M.; Liebetrau, V.; Röhling, H.-G. Southern Ocean bioproductivity during the last glacial cycle—New detection method and decadal-scale insight from the Scotia Sea. *Geol. Soc. Lond. Spec. Publ.* **2013**, *381*, 245–261. [CrossRef]
4. Kamatani, A.; Oku, O. Measuring biogenic silica in marine sediments. *Mar. Chem.* **2000**, *68*, 219–229. [CrossRef]
5. Kröger, N.; Poulsen, N. Diatoms—From Cell Wall Biogenesis to Nanotechnology. *Annu. Rev. Genet.* **2008**, *42*, 83–107. [CrossRef]
6. Sumper, M. A phase separation model for the nanopatterning of diatom biosilica. *Science* **2002**, *295*, 2430–2433. [CrossRef]
7. DeMaster, D.J. The supply and accumulation of silica in the marine environment. *Geochim. Cosmochim. Acta* **1981**, *45*, 1715–1732. [CrossRef]
8. Broecker, W.S. Tracers in the sea. *Geochim. Cosmochim. Acta* **1983**, *47*, 1336.
9. Sullivan, C.W.; Arrigo, K.R.; McClain, C.R.; Comiso, J.C.; Firestone, J. Distributions of phytoplankton blooms in the Southern Ocean. *Science* **1993**, *262*, 1832–1837. [CrossRef] [PubMed]
10. Smith, W.O.; Gordon, L.I. Hyperproductivity of the Ross Sea (Antarctica) polynya during austral spring. *Geophys. Res. Lett.* **1997**, *24*, 233–236. [CrossRef]
11. Arrigo, K.R.; DiTullio, G.R.; Dunbar, R.B.; Robinson, D.H.; VanWoert, M.; Worthen, D.L.; Lizotte, M.P. Phytoplankton taxonomic variability in nutrient utilization and primary production in the Ross Sea. *J. Geophys. Res. Ocean.* **2000**, *105*, 8827–8846. [CrossRef]
12. Sandrini, S.; Ait-Ameur, N.; Rivaro, P.; Massolo, S.; Touratier, F.; Tositti, L.; Goyet, C. Anthropogenic carbon distribution in the Ross Sea, Antarctica. *Antarct. Sci.* **2007**, *19*, 395–407. [CrossRef]
13. Manno, C.; Sandrini, S.; Tositti, L.; Accornero, A. First stages of degradation of Limacina helicina shells observed above the aragonite chemical lysocline in Terra Nova Bay (Antarctica). *J. Mar. Syst.* **2007**, *68*, 91–102. [CrossRef]
14. Ledford-Hoffman, P.A.; Demaster, D.J.; Nittrouer, C.A. Biogenic-silica accumulation in the Ross Sea and the importance of Antarctic continental-shelf deposits in the marine silica budget. *Geochim. Cosmochim. Acta* **1986**, *50*, 2099–2110. [CrossRef]
15. Frignani, M.; Giglio, F.; Accornero, A.; Langone, L.; Ravaioli, M. Sediment characteristics at selected sites of the Ross Sea continental shelf: Does the sedimentary record reflect water column fluxes? *Antarct. Sci.* **2003**, *15*, 133–139. [CrossRef]
16. Petrovskii, S.K.; Stepanova, O.G.; Vorobyeva, S.S.; Pogodaeva, T.V.; Fedotov, A.P. The use of FTIR methods for rapid determination of contents of mineral and biogenic components in lake bottom sediments, based on studying of East Siberian lakes. *Environ. Earth Sci.* **2016**, *75*, 226. [CrossRef]
17. Maldonado, M.; López-Acosta, M.; Sitjà, C.; García-Puig, M.; Galobart, C.; Ercilla, G.; Leynaert, A. Sponge skeletons as an important sink of silicon in the global oceans. *Nat. Geosci.* **2019**, *12*, 815–822. [CrossRef]
18. Calvert, S.E. Accumulation of diatomaceous silica in the sediment of the Gulf of California. *Geol. Soc. Am. Bull.* **1966**, *77*, 569–572. [CrossRef]
19. Eisma, D.; Van Der Gaast, S.J. Determination of opal in marine sediments by X-ray diffraction. *Netherlands J. Sea Res.* **1971**, *5*, 382–389. [CrossRef]
20. Chester, R.; Elderfield, H. The infrared determination of opal in siliceous deep-sea sediments. *Geochim. Cosmochim. Acta* **1968**, *32*, 1128–1140. [CrossRef]

21. Nancy Ann, B. Chapter 18 The Determination of Biogenic Opal in High Latitude Deep Sea Sediments. *Dev. Sedimentol.* **2008**, *36*, 317–331.
22. Leinen, M. A normative calculation technique for determining opal in deep-sea sediments. *Geochim. Cosmochim. Acta* **1977**, *41*, 671–676. [CrossRef]
23. Leinen, M. Techniques for determining opal in deep-sea sediments: A comparison of radiolarian counts and x-ray diffraction data. *Mar. Micropaleontol.* **1985**, *9*, 375–383. [CrossRef]
24. Pokras, E.M. Preservation of fossil diatoms in Atlantic sediment cores: Control by supply rate. *Deep Sea Res. Part A Oceanogr. Res. Pap.* **1986**, *33*, 893–902. [CrossRef]
25. Donald, W.; Eggimann, F.T.M. Dissolution and Analysis of Amorphous Silica in Marine Sediments. *SEPM J. Sediment. Res.* **2003**, *50*, 215–225.
26. Mortlock, R.A.; Froelich, P.N. A simple method for the rapid determination of biogenic opal in pelagic marine sediments. *Deep Sea Res. Part A Oceanogr. Res. Pap.* **1989**, *36*, 1415–1426. [CrossRef]
27. Rosén, P.; Dåbakk, E.; Renberg, I.; Nilsson, M.; Hall, R. Near-infrared spectrometry (NIRS): A new tool for inferring past climatic changes from lake sediments. *Holocene* **2000**, *10*, 161–166. [CrossRef]
28. Rosén, P. Total organic carbon (TOC) of lake water during the Holocene inferred from lake sediments and near-infrared spectroscopy (NIRS) in eight lakes from northern Sweden. *Biogeochemistry* **2005**, *76*, 503–516. [CrossRef]
29. Vogel, H.; Rosén, P.; Wagner, B.; Melles, M.; Persson, P. Fourier transform infrared spectroscopy, a new cost-effective tool for quantitative analysis of biogeochemical properties in long sediment records. *J. Paleolimnol.* **2008**, *40*, 689–702. [CrossRef]
30. Rosén, P.; Vogel, H.; Cunningham, L.; Hahn, A.; Hausmann, S.; Pienitz, R.; Zolitschka, B.; Wagner, B.; Persson, P. Universally applicable model for the quantitative determination of lake sediment composition using fourier transform infrared spectroscopy. *Environ. Sci. Technol.* **2011**, *45*, 8858–8865. [CrossRef]
31. Mecozzi, M.; Pietrantonio, E.; Amici, M.; Romanelli, G. Determination of carbonate in marine solid samples by FTIR-ATR spectroscopy. *Analyst* **2001**, *126*, 144–146. [CrossRef] [PubMed]
32. Khoshmanesh, A.; Cook, P.L.M.; Wood, B.R. Quantitative determination of polyphosphate in sediments using attenuated total reflectance-fourier transform infrared (ATR-FTIR) spectroscopy and partial least squares regression. *Analyst* **2012**, *137*, 3704–3709. [CrossRef] [PubMed]
33. Melucci, D.; Locatelli, C. Multivariate calibration in differential pulse stripping voltammetry using a home-made carbon-nanotubes paste electrode. *J. Electroanal. Chem.* **2012**, *675*, 25–31. [CrossRef]
34. Lorber, A.; Faber, K.; Kowalski, B.R. Net Analyte Signal Calculation in Multivariate Calibration. *Anal. Chem.* **1997**, *69*, 1620–1626. [CrossRef]
35. Bro, R.; Andersen, C.M. Theory of net analyte signal vectors in inverse regression. *J. Chemom.* **2003**, *17*, 646–652. [CrossRef]
36. Malandrino, M.; Mentasti, E.; Giacomino, A.; Abollino, O.; Dinelli, E.; Sandrini, S.; Tositti, L. Temporal variability and environmental availability of inorganic constituents in an antarctic marine sediment core from a polynya area in the Ross Sea. *Toxicol. Environ. Chem.* **2010**, *92*, 453–475. [CrossRef]
37. Dunbar, R.B.; Anderson, J.B.; Domack, E.W.; Jacobs, S.S. Oceanographic influences on sedimentation along the Antarctic continental shelf. In *Antarctic Research Series*; American Geophysical Union: Washington, DC, USA, 1985; pp. 291–312.
38. Liu, W.; Sun, Z.; Ranheimer, M.; Forsling, W.; Tang, H. A flexible method of carbonate determination using an automatic gas analyzer equipped with an FTIR photoacoustic measurement chamber. *Analyst* **1999**, *124*, 361–365. [CrossRef]
39. Ramer, G.; Lendl, B. Attenuated Total Reflection Fourier Transform Infrared Spectroscopy. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: Hoboken, NJ, USA, 2013; pp. 1–27.
40. Rinnan, Å.; van den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
41. Lorber, A. Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations. *Anal. Chem.* **1986**, *58*, 1167–1172. [CrossRef]
42. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
43. Stute, W. The jackknife estimate of variance of a Kaplan-Meier integral. *Ann. Stat.* **1996**, *24*, 2679–2704. [CrossRef]

44. Hemmateenejad, B.; Yousefinejad, S. Multivariate standard addition method solved by net analyte signal calculation and rank annihilation factor analysis. *Anal. Bioanal. Chem.* **2009**, *394*, 1965–1975. [CrossRef] [PubMed]
45. Lorho, G.; Westad, F.; Bro, R. Generalized correlation loadings. Extending correlation loadings to congruence and to multi-way models. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 119–125. [CrossRef]
46. Gendron-Badou, A.; Coradin, T.; Maquet, J.; Fröhlich, F.; Livage, J. Spectroscopic characterization of biogenic silica. *J. Non Cryst Solids* **2003**, *316*, 331–337. [CrossRef]
47. Meyer-Jacob, C.; Vogel, H.; Boxberg, F.; Rosén, P.; Weber, M.E.; Bindler, R. Independent measurement of biogenic silica in sediments by FTIR spectroscopy and PLS regression. *J. Paleolimnol.* **2014**, *52*, 245–255. [CrossRef]
48. Colthup, N.B.; Daly, L.H.; Wiberley, S.E. *Introduction to Infrared and Raman Spectroscopy*, 3rd ed.; Academic Press: San Diego, CA, USA, 1990; ISBN 978-0-12-182554-6.

Sample Availability: Samples of the compounds are not available.

molecules

MDPI

Article

Raman Spectroscopy and Chemometric Modeling to Predict Physical-Chemical Honey Properties from Campeche, Mexico

F. Anguebes-Franseschi [1], M. Abatal [2], Lucio Pat [3], A. Flores [2], A. V. Córdova Quiroz [1], M. A. Ramírez-Elias [1], L. San Pedro [4], O. May Tzuc [4] and A. Bassam [4,*]

[1] Faculty of Chemistry, Autonomous University of Carmen, Street 56 No. 4 Esq. Av. Concordia, Col. Benito Juárez, Z. C. 24180 Ciudad del Carmen, Campeche, Mexico; fanguebes@pampano.unacar.mx (F.A.-F.); acordova@delfin.unacar.mx (A.V.C.Q.); mramirez@pampano.unacar.mx (M.A.R.-E.)

[2] Faculty of Engineering, Autonomous University of Carmen, Campus III, Avenida Central s/n, Esq. Con Fracc. Mundo Maya, C. P. 24115 Ciudad del Carmen, Campeche, Mexico; mabatal@pampano.unacar.mx (M.A.);aflores@pampano.unacar.mx (A.F.);

[3] South Frontier College, Av. Rancho Polígono 2-A, Ciudad Industrial, 24500 Lerma, Campeche, Mexico; lpat@ecosur.mx

[4] Faculty of Engineering, Autonomous University of Yucatan, Av. Industrias no Contaminantes Periférico Norte, Cordemex, Z.C. 97310 Mérida, Yucatan, Mexico; liliana.cedillo@correo.uady.mx (L.S.P.); maytzuc@gmail.com (O.M.T.)

* Correspondence: baali@correo.uady.mx; Tel.: +52-999-930-0550 (ext. 1053)

Academic Editors: Marcello Locatelli and Angela Tartaglia
Received: 5 September 2019; Accepted: 5 November 2019; Published: 13 November 2019

Abstract: In this work, 10 chemometric models based on Raman spectroscopy were constructed to predict the physicochemical properties of honey produced in the state of Campeche, Mexico. The properties of honey studied were pH, moisture, total soluble solids (TSS), free acidity, lactonic acidity, total acidity, electrical conductivity, Redox potential, hydroxymethylfurfural (HMF), and ash content. These proprieties were obtained according to the methods described by the Association of Official Analytical Chemists, Codex Alimentarius, and the International Honey Commission. For the construction of the chemometric models, 189 honey samples were collected and analyzed in triplicate using Raman spectroscopy to generate the matrix data [X], which were correlated with each of the physicochemical properties [Y]. The predictive capacity of each model was determined by cross validation and external validation, using the statistical parameters: standard error of calibration (SEC), standard error of prediction (SEP), coefficient of determination of cross-validation (R^2_{cal}), coefficient of determination for external validation (R^2_{val}), and Student's *t*-test. The statistical results indicated that the chemometric models satisfactorily predict the humidity, TSS, free acidity, lactonic acidity, total acidity, and Redox potential. However, the models for electric conductivity and pH presented an acceptable prediction capacity but not adequate to supply the conventional processes, while the models for predicting ash content and HMF were not satisfactory. The developed models represent a low-cost tool to analyze the quality of honey, and contribute significantly to increasing the honey distribution and subsequently the economy of the region.

Keywords: quality control; Raman spectroscopy; honey; PLS regression models; physicochemical parameters

1. Introduction

Honey is a natural product, and a complex solution elaborated by honey bees. It is mainly composed of sugars (70–80%) and water (10–20%), and in minor quantities contains flavonoids, phenolic acids, vitamins, proteins, organic acids, lipids, carotenoids, minerals, and enzymes [1]. Honey has been used since ancient times as a food supplement for humans. Additionally, due to its content of phenolic compounds and flavones, it also has several beneficial health effects, which include prebiotic, antimicrobial, anticarcinogenic, antioxidant, antihypertensive, antibacterial, antifungal, anti-inflammatory, and analgesic effects [2–4]. The physical, chemical, and biological properties of honey depend on the type of flowers visited by the honey bees, and the soil where the nectar and pollen are collected. Other influences on its quality are the environmental and storage conditions, as well as the processing for its commercialization [5]. Therefore, quality control of honey represents an important concern for the beekeeping industry, since, on the one hand, it allows tracing of the geographical and botanical origin of the pollen (designation of origin), and, on the other hand, it allows identification of its possible adulteration during processing [6,7].

To classify and determine the honey's quality, standards and methods have been established in the Codex Alimentarious [8], International Honey Commission (IHC) [9], and the Association of Official Analytical Chemists (AOAC) [10]. These standards specify the physical and chemical properties that must be evaluated to determine the honey's quality. The traditional method to perform quality tests on honey involves the analysis of pollen grains contained in its sediments by light microscopy (melissopalynology) [6]. Other methods reported in the literature include chromatography techniques, stable carbon isotope radio analysis, and nuclear magnetic resonance [7,11]. The main drawbacks of all of them are their high cost, time consuming nature, requirement for specialists, and furthermore the fact that many of them are destructive. This has led to the development of analytical methods for the authentication of honey. In this sense, spectroscopy technology combined with chemometric tools represents a good alternative for the fast, reliable, and environmentally friendly quality control of honey samples. The above is due to the development of calibration models that can determine the concentration of a specific chemical species in a mixture of several components [12]. Among the most common chemometric techniques used in honey analysis are Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Linear Discriminant Analysis (LDA), Partial Least Square (PLS), and Principal Component Regression (PCR) [13].

From the spectrometric techniques available, Raman spectrometry has suitable characteristics for food analysis, such as non-interference from water present in the sample with the Raman measurement, ease of sampling and measurement, and minimal fluorescence interference of the sample matrix variation. In recent years, analytic methods based on Raman spectrometry have been explored as an economic and rapid option to determine honey's destination of origin [14–17]. Corvucci et al. [14] contrasted the ability to identify honey's botanic origin using the melissopalynology technique compared to Raman spectroscopy coupled with multivariable analysis (PCA). The study considered honey samples from Italy, Eastern Europe, and Spain. According to the results, the discrimination of honey origin given by the two first principal components was improved from 85% to 99% using the analytical method. Frausto-Reyes et al. [15] determined the floral origin of honey produced by *Apis Mellifera*, applying Raman spectroscopy together with PCA. The study used 66 samples of both monofloral and polifloral honey collected from several regions of Mexico with different climate types. The use of the chemometric approach was adequate to classify the origin of the sample and the purity of the pollen with 90% accuracy. Jandrić et al. [16] presented a method for the authentication of floral origin honey produced in New Zealand. They combined Raman spectrometry, near infrared spectrometry, and Fourier-transform infrared spectroscopy for the analytical study of honey samples in the range between 200 to 12,000 cm^{-1}. This approach was completed with the use of PLS for the development of chemometric models. The results showed a model fit (R^2), a standard error of calibration (SEC), and standard error of prediction (SEP) of 85.0%, 0.219 and 0.315, respectively. Oroian and Ropciuc [17] applied Raman spectra analysis for the botanical authentication of 76 samples of honey from Romania. The use of this analytic method combined with LDA proved to be an excellent authentication tool, achieving 83.33% cross validation accuracy.

Similarly, the literature reports the use of Raman spectra analysis coupled with multi-variable modeling for the detection of external agents that affect the quality of honey [18–22]. Raman spectroscopy and chemometric models have been used to predict the concentration of glucose, fructose, sucrose, and maltose present in honey samples from Turkey and Greece [18]. The correlation between quantified sugar levels and Raman spectra was performed using both PLS and artificial neural networks (ANN). The statistical R^2 for glucose, fructose, sucrose, and maltose were high, with 0.929, 0.930, 0.937, and 0.893 for PLS and 0.930, 0.931, 0.956, and 0.913 for ANN, indicating that both chemometric tools are efficient for the rapid analysis of sugar content. Oroian et al. [19] used Raman spectroscopy to detect honey adulterated with sugars (glucose, fructose, inverter sugar, hydrolyzed inulin syrup, and malt must). The study considered 900 samples with adulteration levels of 5, 10, 20, 30, 40, and 50%. Authentication of honey purity concentration was performed using PLS and PCR. The chemometric models developed showed good fit for both the calibration (R^2_{cal} = 0.983) and validation (R^2_{val} = 0.981) dataset, with low statistical errors (SEC = 0.009 and SEP = 0.103). Anjos et al. [20] evaluated the potential of Raman spectroscopy in the prediction of the physicochemical composition of *Lavandula* spp. monofloral honey. PLS models were used for the quantitative estimation, and the results were correlated with the values obtained using reference methods. Chemometric models were used for pH, sugar reduction, electrical conductivity, apparent sucrose, total phenol content, total flavonoid content, proline, and total acids, achieving R^2_{cal} in the range of 0.973–0.99, R^2_{val} in the range of 0.833–0.99, SEC in the range of 2.03–0.01, and SEP in the range of 1.71–0.01. In the study by Tahir et al. [21], Raman spectroscopy combined with PLS were applied to predict phenolic compounds and antioxidant activity in honey. It was found that the developed models based on Raman were superior to those established using NIR spectra, with R^2_{cal} and R^2_{val} > 90%, SEC < 1.2, and SEP < 1.7. Raman spectroscopy, and PLS-LDA modeling have also been used to determine the adulteration of Chinese honey with corn syrup [22]. The analysis considered adulteration samples in the range of 10, 20, and 40%. An accuracy prediction of 84.4% was obtained, indicating that combining PLS-LDA with Raman spectra is a potential technique for the detection of impure agents in honey.

In this paper, a study is presented to determine the physical-chemical properties of honey from the Mexican region of the Yucatan Peninsula. In this zone, beekeeping is an ancient activity, carried out since the pre-Columbian era by Mesoamerican cultures like the Maya, who already produced honey from apiaries with honey bees (*Melipona beecheii*) long before the arrival of the Spaniards [23]. After their conquest, the species *Apis mellifera* was introduced in Mexico, which proliferated and dispersed throughout the country due to its higher yields of honey. Currently, the Yucatan Peninsula (located in the south of the country and composed of the states of Yucatan, Campeche, and Quintana Roo) is one of the most fruitful regions for the development of beekeeping activity. This region is characterized by ecosystems with great flora diversity, producing nectars and pollen—many of them endemic—that produce honey with unique organoleptic, physical, and chemical properties; these characteristics make honey from this region very appreciated in national and international markets [24]. In this sense, Mayan beekeepers from the Yucatan Peninsula contribute approximately 35% of the national production. In the state of Campeche, there are 4030 honey producers that generate on average 5571 metric tons of honey per year; Campeche is the second honey producer region nationwide, only surpassed by Yucatan. Of the total produced in this region, 95% is exported, producing profits of up to 12 million US dollars and contributing to generating economic welfare for Mayan beekeepers [25–27]. Thus, the introduction of fast and low-cost tools to analyze the quality of the honey produced would contribute significantly to increasing distribution of this natural food, benefiting local beekeepers and the local economy.

Therefore, due to the economic importance of honey production in the state of Campeche, Mexico, the objective of this work was to develop chemometric models based on Raman spectroscopy for the quantification of the following physical and chemical properties: pH, moisture, total soluble solids (TSS), free acidity, lactonic acidity, total acidity, electrical conductivity (EC), Redox potential, hydroxymethylfurfural (HMF), and ash content. These chemometric models represent useful tools for the quality control of honey produced in the state of Campeche, by quickly and economically predicting the main physicochemical indicators.

2. Analysis of Results

2.1. Raman Analysis

Figure 1 shows that the Raman spectra obtained from the honey samples have spectral bands which cover the ranges of 330–404, 404–440, 440–510, 510–595, 595–691, 691–752, 770– 820, 820–1024, 1024–1094, 1094–1191, 1191–1262, 1262–1300, and 1300–1460 cm^{-1}:

- Spectral region between 230–510 cm^{-1} are related to stretching and bending vibrations of the C-O, C-C-O and C-C-C that form the molecular structure of sugars [21].
- The region between 595–691 cm^{-1} is attributed to stretching vibrations of unsaturated rings present in HMF, carotenes, flavones, flavonoids, and polyphenols [22].
- The peak found between 691–752 cm^{-1} is assigned to stretching vibrations of C-O and C-C-O, and bending vibrations of O-C-O. On the other hand, the band between 770–917 cm^{-1} is a product of the stretching vibrations of the C-C and C-H present in glucose [28].
- Regarding the bands between 820–1024 cm^{-1}, these correspond to deformation vibrations of C-H and methylene bonds –CH_2–, as well as the bending vibrations of C-O-H [29].
- The peak present between 1024–1094 cm^{-1} is attributed to bending vibrations of the C-H and C-O-H bonds of sugars, and bending vibrations of the C-N bonds of amino acids and proteins [30].
- The band between 1094–1191 cm^{-1} is assigned to stretching vibrations of the C-O, C-O-C bonds of sugars, and the C-N bonds of proteins and amino acids [18].
- Finally, the spectral region between 1262–1300 cm^{-1} corresponds to vibrations of C-H and O-C-H, while the spectral bands of 1300–1460 cm^{-1} are due to bending and wobble vibrations of the functional groups CH and –OH [30].

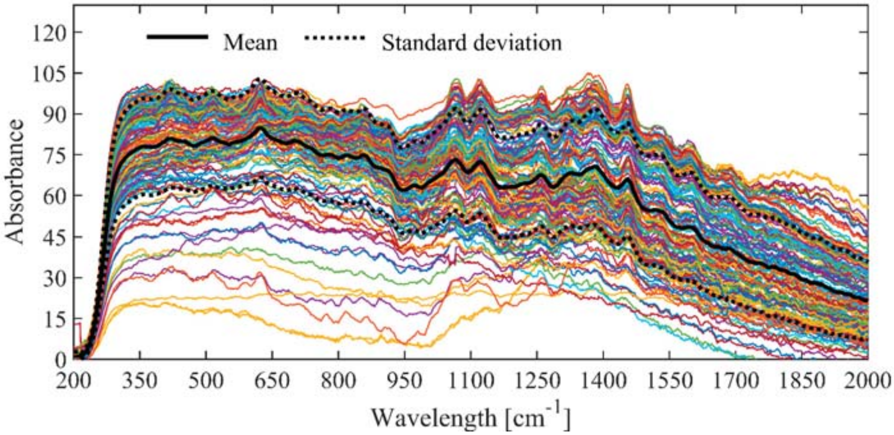


Figure 1. Raman spectral footprints of the honey collected in the various locations of Campeche.

2.2. Chemometric Models

2.2.1. Chemometric Models to Predict pH, Free Acidity, Lactonic Acidity, and Total Acidity

The presence of organic acids, such as gluconic, phenolic, ascorbic, lactic, and metallic ions, causes honey to be slightly acidic by nature. The acidity may be increased due to chemical and biochemical changes that take place in the honey. For example, the glucose oxidase enzyme is capable of transforming glucose into gluconic acid; on the other hand, the ions of the alkaline earth elements can react to form phosphates, sulfates, and chlorides, as well as transform lactone into lactic acid [31]. To measure these chemical changes in honey, in the Codex Alimentarius [8], the pH, free acidity, lactonic acidity, and total acidity were established as quality control criteria. In this sense, free acidity is related to the concentration of organic acids in honey, where a maximum value of 50 meq kg^{-1} is established by the Codex Alimentarius.

Table 1 lists the values of the 10 physicochemical parameters determined for honey samples from the municipalities of the state of Campeche. As reported in the table, the pH of honey samples were in the range of 3.49 to 5.2, within the limit established by the Codex Alimentarius (minimum 3.40 and maximum 6.10). The minimum and maximum values of free acidity were detected between 22.5 and 35.1 meq kg^{-1}, 4.15 y 9.45 meq kg^{-1} for lactonic acidity, and 28.67 a 38.28 meq kg^{-1} for total acidity. According to this, the values of the total acidity present in honey samples agree with the provisions of the Codex Alimentarius, indicating that the honey collected did not show significant degradation.

Table 1. Results obtained for the different physical and chemical parameters of honey from the municipalities of the state of Campeche.

Property	Mean ± σ	Minimum	Maximum	Mean ± σ	Minimum	Maximum
		Calakmul			*Calkini*	
pH	4.01 ± 0.23	3.66	5.11	4.08 ± 0.17	3.80	4.77
Free acidity	21.16 ± 5.03	8.12	32.53	19.79 ± 3.03	15.52	25.51
Lactonic acidity	2.96 ± 1.001	1.23	5.78	2.77 ± 0.84	1.47	4.27
Total acidity	24.17 ± 5.44	11.55	36.78	22.51 ± 3.31	18.25	28.67
Electric conductivity	0.58 ± 0.08	0.35	0.69	0.61 ± 0.05	0.49	0.68
Redox potential	181.94 ± 13.91	133.1	207.2	173.54 ± 8.87	161.6	198.1
Moisture	14.98 ± 1.42	11.81	17.66	12.21 ± 2.27	12.29	16.66
TSS	85.02 ± 1.41	82.37	88.19	85.79 ± 1.09	83.34	87.71
Ash content	0.14 ± 0.06	0.018	0.42	0.143 ± 0.14	0.09	0.21
HMF	2.87 ± 1.33	1.27	5.89	2.31 ± 0.75	1.46	4.35
		Campeche			*Carmen*	
pH	3.95 ± 0.16	3.49	4.18	3.97 ± 0.14	3.64	4.25
Free acidity	17.03 ± 3.52	12.39	26.1	21.22 ± 4.19	8.01	28.53
Lactonic acidity	2.51 ± 0.68	1.47	4.15	3.09 ± 1.08	1.23	5.78
Total acidity	19.53 ± 3.81	14.17	29.65	24.32 ± 4.41	11.45	31.34
Electric conductivity	0.48 ± 0.08	0.28	0.69	0.57 ± 0.08	0.35	0.69
Redox potential	177.49 ± 9.89	151.3	204.2	186.23 ± 8.41	170.1	207.4
Moisture	15.25 ± 3.11	12.76	24.6	15.02 ± 1.53	11.81	17.66
TSS	84.74 ± 3.11	75.42	87.24	84.98 ± 1.53	82.34	88.19
Ash content	0.13 ± 0.018	0.08	0.16	0.14 ± 0.09	0.02	0.88
HMF	2.12 ± 0.46	1.52	3.53	2.98 ± 1.43	1.27	5.89
		Champotón			*Escarcega*	
pH	3.78 ± 0.18	3.55	4.23	3.85 ± 0.17	3.62	4.31
Free acidity	22.81 ± 4.26	11.9	32.5	22.72 ± 5.11	13.5	31.5
Lactonic acidity	3.59 ± 0.78	2.37	5.98	3.51 ± 0.62	1.78	4.37
Total acidity	26.41 ± 4.47	17.01	38.28	26.23 ± 5.13	17.07	35.59
Electric conductivity	0.54 ± 0.11	0.36	0.69	0.58 ± 0.12	0.35	0.755
Redox potential	189.03 ± 11.39	165.4	202.6	172.52 ± 9.38	146.1	185.8
Moisture	16.9 ± 3.11	13.32	25.81	15.16 ± 0.88	13.65	16.89
TSS	83.01 ± 3.11	74.2	86.36	84.83 ± 0.88	83.11	86.35
Ash content	0.14 ± 0.03	0.11	0.17	0.13 ± 0.02	0.068	0.18
HMF	3.34 ± 1.32	1.57	6.39	2.34 ± 1.44	1.57	4.89
		Hecelchacan			*Hopelchén*	
pH	4.09 ± 0.09	3.91	4.21	4.34 ± 0.42	3.51	5.2
Free acidity	17.78 ± 3.06	16.85	22.5	16.64 ± 6.95	6.5	35.1
Lactonic acidity	5.14 ± 2.48	3.19	9.45	3.44 ± 0.91	1.67	5.92
Total acidity	22.93 ± 5.41	21.07	31.95	20.08 ± 6.82	10.41	37.77
Electric conductivity	0.61 ± 0.056	0.51	0.659	0.59 ± 0.08	0.44	0.71
Redox potential	177.49 ± 14.34	167.5	202.1	153.93 ± 22.21	105.6	198.2
Moisture	17.09 ± 3.19	15.17	22.67	14.72 ± 1.23	12.43	17.4
TSS	82.85 ± 3.16	77.33	85.45	85.27 ± 1.23	82.6	87.57
Ash content	0.13 ± 0.015	0.11	0.14	0.14 ± 0.03	0.05	0.21
HMF	2.89 ± 0.265	2.39	3.27	3.18 ± 0.95	1.56	5.78

The variability in the pH, free acidity, lactonic acidity, and total acidity is represented in Table 1 by the standard deviation (σ). In this sense, the honey samples with the highest pH standard deviation were those from the municipalities of Calakmul and Holpechen, with ±0.23 and ±0.42, respectively. This variability is attributed to the diversity of melliferous flora present in the region (Figure 1), which belongs to the Calakmul biosphere reserve and houses more than 150 melliferous flowers, with important differences in their chemical composition [24,25]. On the other hand, honey samples that presented higher pH values (4.18–5.2) correspond to productions from the Tajonal and Mangle Negro plants, characterized by a higher concentration of sodium chloride. The Tajonal is a plant widely distributed in the state of Campeche, which is adapted to alkaline soils and is capable of growing near coastal areas, where a sea breeze is deposited on the flowers. Likewise, Mangle Negro grow in the coastal zone, on the banks of lagoons and estuaries that contain waters with high salinity; this contributes to the fresh honey from these flowers having low acidity due to the presence of sodium chloride.

Regarding total acidity, this presents standard deviations of ±5.44 meq kg^{-1} for honey samples from Calakmul and ±6.82 meq kg^{-1} in honey from Hopelchen. The free acidity for honey from the municipalities of Carmen has standard deviations ±4.41 meq kg^{-1} and ±4.47 meq kg^{-1} for those of Champotón, and ±5.13 meq kg^{-1} for Escarcega. The municipalities of Carmen, Champotón, and Escarcega are geographically are located in the west of the state of Campeche, a region characterized by lagoons, wetlands, rivers and estuaries that are conducive to the growth of melliferous plants such as Arbol de tinto, Pucté, Mangle, Cascarillo, and Ja'abin, among others. The honey of these floral species has a higher moisture content, which favors honey fermentation. On the other hand, the Hecelchacan honey samples showed a standard deviation of ±5.41 meq·kg^{-1}. This variability is attributed to the predominance in this region of melipona honey, which by its nature usually contains water concentrations above 20%, favoring the formation of organic acids by biochemical reactions.

Based on the measurements obtained, chemometric models were created to predict pH, free acidity, lactonic acidity, and total acidity. Figure 2 shows the predictive behavior of the models, while Table 2 contains their statistical performances. The calibration model to predict the pH in honey of the state of Campeche exhibits a standard error of calibration SEC = 0.86 and standard error of prediction SEP = 0.18; likewise, it presents acceptable values for the coefficient correlation of calibration ($R^2{}_{cal}$ = 0.92) and the coefficient correlation of validation ($R^2{}_{val}$ = 0.74). These statistical values show that the chemometric model has an acceptable ability to predict the pH in honey. On the other hand, Student's *t*-test with paired data at 95% confidence obtained t_c = 0.95, within the established confidence interval (t_v = ±1.65). Therefore, the chemometric model based on Near Infrared Spectroscopy (NIRS) has a good reliability but not enough to substitute the standardized method. The statistical values obtained in this work are similar to those reported by Cozzolino et al. [32], who obtained a chemometric model using Vis-NIRS spectroscopy to predict the pH of honey in Uruguay. They also reported values of SEC = 0.13, SEP = 0.21, $R^2{}_{cal}$ = 0.88, and $R^2{}_{val}$ = 0.70. On the other hand, Anjos et al. [20] reported statistical values of SEC = 0.12, SEP = 0.09, $R^2{}_{val}$ = 0.83, and $R^2{}_{cal}$ = 0.98 for a calibration model based on the FT-Raman spectroscopy used to predict the humidity percentage in Portuguese honey.

The chemometric model for predicting free acidity presented a standard error of calibration (SEC = 1.02), a standard error of prediction (SEP = 1.47), coefficient correlation of calibration ($R^2{}_{cal}$ = 0.98, and coefficient correlation of validation ($R^2{}_{val}$ = 0.94). These results indicate that the chemometric model successfully predicts the concentration of honey's free acidity. The Student's *t*-test of paired data (t_c = 0.64) for free acidity is within the confidence interval (t_v = ±1.65), indicating that there are no differences in the prediction capacity of the developed chemometric model with respect to the standard method established in the Codex Alimentarius [8]. In previous studies, such as the one carried out by Ruoff et al. [33], the following statistical values were reported for a chemometric model based on NIRS spectroscopy to predict free acidity in Swiss honey: a standard error of calibration (SEC = 2.01), standard error of prediction (SEP = 2.0), and coefficient correlation of validation ($R^2{}_{val}$ = 0.737).

Table 2. Values of the statistical parameters obtained in cross-validation and external validation to determine the capacity predictability of each chemometric model.

Properties	Units	Calibration LVs	SEC	R^2_{cal}	Validation LVs	SEP	R^2_{val}
pH	-	5	0.86	0.92	4	0.18	0.743
Free acidity	meq kg^{-1}	6	1.02	0.98	6	1.47	0.935
Lactonic acidity	meq kg^{-1}	6	0.37	0.94	7	0.41	0.911
Total acidity	Meq kg^{-1}	6	1.08	0.98	4	1.23	0.897
Electrical conductivity	mS cm^{-1}	6	0.46	0.87	4	0.85	0.79
Redox potential	*mV*	7	1.06	0.99	8	1.48	0.95
Moisture	%	6	0.42	0.98	9	0.52	0.95
TSS	%	6	0.58	0.92	6	1.32	0.87
Ash content	%	6	1.21	0.78	6	2.54	0.21
HMF	mg kg^{-1}	7	0.76	0.82	8	1.73	0.63

With regards to the chemometric model for predicting lactonic acidity in Campechean honey, it showed good predictive capacity, since the values of cross-validation and external validation, along with the standard error of calibration and standard error of prediction, were small (SEC = 0.37; SEP = 0.41), with the following coefficient correlation of calibration and coefficient correlation of validation (R^2_{cal} = 0.94; R^2_{val} = 0.91). For the Student's *t*-test of paired data (t_c = 0.69) at 95% confidence, the value obtained is in the confidence interval (t_v = ±1.65), so there are no differences in the prediction capacity of lactonic acidity between the obtained chemometric model and the standard method [8].

Finally, the chemometric model to predict total acidity in Campeche honey showed a high coefficient correlation in the cross-validation (R^2_{cal} = 0.98) and coefficient correlation in the external validation (R^2_{val} = 0.89), as well as low values of standard error of calibration (SEC = 1.18) and standard error of external validation (SEP = 1.23). Moreover, the Student's *t*-test of paired data (t_c = 0.75) is in the confidence interval (t_v = ± 1.65), which demonstrates that the chemometric model is as reliable as the standardized method. Comparing the obtained results with those reported by Anjos et al. [20] for an FT-Raman spectroscopy calibration model to predict the acidity total in Portuguese honey, similar values were observed (SEC = 0.22; SEP = 0.28; R^2_{cal} = 0.99; R^2_{val} = 0.99).

In Figure 2, it can be seen that the experimental data of the pH, free acidity, lactonic acidity, and total acidity of the honey samples show a certain degree of dispersion compared to the chemometric model predictions. This can be attributed to the following factors: first, in the state of Campeche, several tropical forests are located that give rise to a great diversity of honey blooms; previous works have identified more than 150 blooms in the area of study [24–26]. Thus, the honeys produced in the region are multifloral, giving rise to a wide variety of physical and chemical properties. Second, the geographical origins where the honey samples were collected—specifically in the east of the state of Campeche, in the municipalities of Carmen, Palizada, Escarcega, and Champotón—are characterized by the presence of rivers, lagoons, wetlands, and swamps. These soils are rich in organic matter and have an acidic pH, which contribute to the development of a great diversity of melliferous flora, such as: Tahonal, Ja'abin, Pukte, huano, Xtabentum, Palo Tinto, hulub, Suuk chak lol, Box káatsim, Bohom, Susuk, cascarillo and mangle negro. Flowers from these botanical origins produce nectar with high concentrations of moisture, which is transferred to the honey [26]. The presence of a high percentage of moisture in honey favors biochemical and chemical reactions—for example, the formation of gluconic acid from glucose and the formation of inorganic acids due to the reaction of water with anions and cations present in honey. This means that honey samples collected in these locations show greater variability in pH, free acidity, lactonic acidity, and total acidity [34,35].

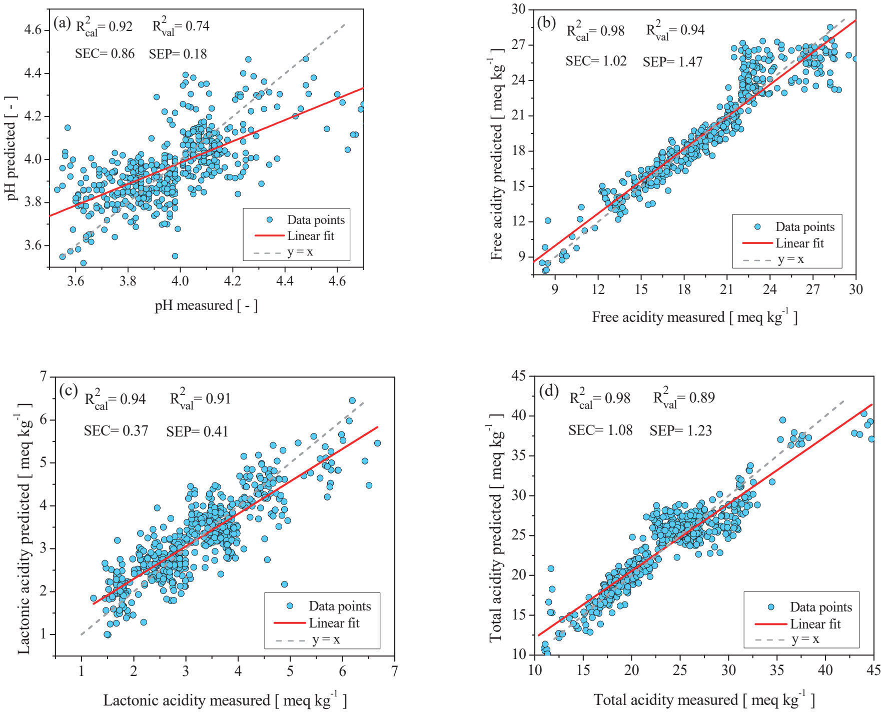


Figure 2. Chemometric models to predict: (**a**) pH; (**b**) free acidity; (**c**) lactonic acidity; (**d**) total acidity.

2.2.2. Chemometric Model to Predict Electrical Conductivity, Redox Potential, Moisture, and TSS

Electrical conductivity is a parameter used to determine the geographical origin of honey. This is related to the content of ashes, organic acids, and dissolved mineral salts; the higher the concentration of these compounds in honey, the greater the value of the electrical conductivity [36]. In this sense, the diverse honey samples from the state of Campeche presented values between 0.28–0.75 mS cm^{-1}, which is below the maximum allowed limit (0.80 mS cm^{-1} [8]). The chemometric model for this physicochemical property had a standard cross-validation error and an external validation error of 0.46 and 0.85, respectively. Moreover, the regression coefficients obtained were R^2_{cal} = 0.87 and R^2_{val} = 0.79. Nevertheless, the R^2_{val} value indicates an acceptable model fit, but are not significant for our propose. In Figure 3a, a noticeable dispersion between the experimental data of the electrical conductivity with respect to the chemometric model is observed. This is attributed to the significant differences in organic matter, salinity content, and carbonates in the soils of the locations where the honey samples were collected. Another cause is the diversity of the honey flora, which contributed to the variation in the content of organic acids in the honeys [24,31].

Comparing the results obtained with previous works, these present slightly lower values than those reported by Anjos et al. [20], who built a chemometric model based on FT-Raman for Portuguese honey. They reported values of calibration errors and external validation of (SEC = 0.01; SEP = 0.01), and coefficients of determination (R^2_{cal} = 0.92.8; R^2_{rval} = 0.938). Nonetheless, the results obtained in our study are similar to those reported by Ruoff et al. [33] for a chemometric model based on NIRS spectroscopy to predict electrical conductivity in Swiss honeys (R^2_{cal} = 0.794 and R^2_{rval} = 0.87); and with the data reported by Cozzolino et al. [32] for a calibration model of Uruguayan honey (R^2_{cal} = 0.83 and R^2_{rval} = 0.80).

On the other hand, honey contains chemical substances dissolved in low concentrations of organic acids, mineral salts, and polyphenols; polyphenols are molecules that contain unsaturated bonds in their chemical structure, and develop a very important function since they are antioxidants; these substances have the property of trapping free radicals generated in biochemical reactions. When honey undergoes cooking processes or remains stored for a long period, the aforementioned substances may undergo oxide–reduction reactions, causing changes in their molecular structure and modifications in the properties of honey. These chemical changes can be monitored using the Redox potential to determine the degree of oxidation. Because the Redox potential can be used as a quality control parameter, it was analyzed in Campeche honeys. The results indicate Redox potential values with a minimum of 133.1 mV and a maximum of 207.2 mV; the difference in these results is attributed to the composition of each bloom. The chemometric model had calibration and validation errors (SEC = 1.06; SEP = 1.48), and high values in the calibration and external validation coefficients (R^2_{cal} = 0.99; R^2_{rval} = 0.95). The reliability of the model was also confirmed by a Student's *t*-test of paired data, with a value of t_c = 0.545 between t_v = ±1.65 at 95% confidence, so the model has a good predictive capacity.

With regards to moisture, a maximum content of 20% was defined in the Codex Alimentarious [8]. This is because an excess of moisture favors the fermentation of sugars, causing the formation of undesirable organic acids that affect the organoleptic properties [37]. The moisture content in honey depends on several factors, such as floral origin, harvest time, climate change, maturity degree of the honey, and improper handling of the honey by beekeepers [38]. The analyzed honey samples presented humidity values between 11.81–25.81%; some samples showed humidity concentrations above 20% because the honey came from tree blooms located in wetlands, near rivers, and near estuaries. In addition, some samples were from melipona honey, that, by nature, contains high concentrations of moisture [39]. The chemometric model to predict moisture in honey presented SEC = 0.42, SEP = 0.52, R^2_{cal} = 0.98, and R^2_{val} = 0.97. Additionally, t_c = 0.41 was obtained in the Student's *t*-test, which indicates that the chemometric model correctly predicts moisture in the honey. The presented results are similar to those reported by Lichtenberg et al. [40] for a predictive model of moisture in German honeys (R^2_{cal} = 0.73 and σ = ±1.22). Likewise, it is consistent with what was reported by García et al. (2000) regarding a chemometric model to predict moisture in honey from the region of Galicia, Spain (SEC = 0.12, SEP = 0.15, and R^2_{cal} = 0.98); and with Cozzolino et al. [32], who reported values of (SEC = 2.7, SEP = 3.1, R^2_{cal} = 0.96, and R^2_{val} = 0.94) for a calibration model focused on predicting moisture content in Uruguayan honey.

The principal component of honey is sugar; honey contains a mixture of sugars, mainly fructose, glucose, sucrose, maltose, and melezitose. Glucose and fructose are the ones that are found in the highest proportion and can represent up to 95% of the sugar content [41]. The honey samples collected in the state of Campeche exhibited total sugar concentrations between 74.19–88.19% w; some samples presented concentrations below 80 ° Brix [8] due to a higher moisture concentration. The chemometric model developed to predict TSS showed the following statistical results: SEC = 0.58; SEP = 1.32; R^2_{cal} = 0.92; R^2_{val} = 0.87. A Student's *t*-test with a value of t_c = 0.28 was in the range t_v = ±1.65 at 95% reliability, which shows that the model for predicting TSS has an acceptable prediction capacity but not adequate to supply the referenced method. The results obtained in this work were similar to those reported by Mignani et al. [42], who built chemometric models based on Raman spectroscopy to predict glucose and fructose concentrations in Italian honeys (SEC = 7.3; SEP = 11; R^2_{cal} = 0.96, R^2_{val} = 0.92) and (SEC = 5.5; SEP = 7.6; R^2_{cal} = 0.89, R^2_{val} = 0.82). Likewise, Özbalci et al. [18] developed calibration models based on Raman spectroscopy to predict glucose and fructose concentrations in Turkish honeys, reporting the following values (SEC = 0.51; SEP = 2.75; R^2_{cal} = 0.98; R^2_{val} = 0.96). Complementing this, Anjos et al. [20] reported statistical results (SEC = 0.34, SEP = 0.39, R^2_{cal} = 0.99, R^2_{val} = 0.99) for a predictive calibration model of reducing sugars in Portuguese honey. The comparisons between values predicted by the chemometric models presented in this section and their respective experimental values are shown in Figure 3.

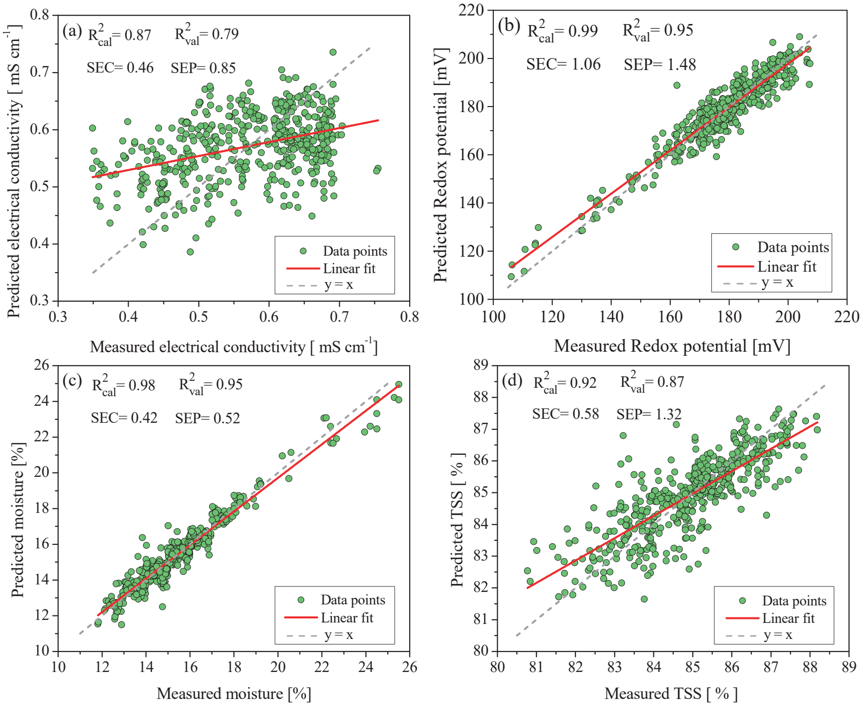


Figure 3. Chemometric models to predict: (**a**) electrical conductivity; (**b**) Redox potential; (**c**) moisture; (**d**) TSS.

2.2.3. Chemometric Model to Predict Content of HMF and Ashes

As presented in Table 2, the chemometric models to predict ash percentage and HMF content presented low coefficients of determination in cross-validation and external validation (R^2_{cal} = 0.78, R^2_{val} = 0.21; and R^2_{cal} = 0.82, R^2_{val} = 0.56). The above indicates that the models are not suitable for the prediction of these chemometric properties.

2.3. Analysis of the PLS loadings

The PLS loading for total acidity, electrical conductivity, Redox potential, humidity and TSS (Figure 4) present six spectral regions (200–600 cm^{-1}, 630–790 cm^{-1}, 870–1000 cm^{-1}, 1080 –1200 cm^{-1}, 1400–1570 cm^{-1}, and 1750–1880 cm^{-1}) that provide useful chemical information for the development of their respective predictive chemometric models. The first spectral band (between 200–400 cm^{-1}) has a positive and negative contribution in the PLS loading. The chemical information provided by this region is related to stretching, bending, and deformation vibrations of C-O, C-C-O, C-C-C and C=O, which form the skeleton of sugar molecules, organic acids, phenolic compounds, and flavonoids. Here, breaks of functional groups and of the sugar backbone can occur due to oxide–reduction reactions; for example, in the transformation of glucose into gluconic acid, fermentation reactions for the production of alcohols and carboxylic acids and the cyclization of fructose produce HMF. These chemical changes in the honey collected would reflect variations in total acidity, pH, Redox potential, and electrical conductivity with respect to time. The band at 630–790 cm^{-1} provides chemical information of the cyclic and alicyclic rings that make up the molecules of HMF, carotenes, flavonols, flavanones, and flavones, among other phenolic compounds. The chemical information related to the band between 870–1000 cm^{-1} is attributed to stretching, bending and deformation vibrations of the C-C, C-H, C-H, –CH_2–, and C-O-H bonds present in the sugars. The Raman region between 1080–1200 cm^{-1} provides information on protein and carbohydrate content in honey, due to stretching vibrations of C-O, C-O-C, C-N carbohydrate, and protein bonds. The region between 1400–1570 cm^{-1} provides chemical information due to bending and wobble vibrations of CH, O-C-H and –OH functional groups present in sugar molecules, and –OH in the water molecules. Finally, the concentrations of moisture, fructose, glucose and moisture in honey are related to stretching vibrations of the unsaturated bonds C=O in fructose and CH=O in glucose, and deformation vibrations –OH of water, which are present in the Raman spectrum between 1750–1880 cm^{-1}.

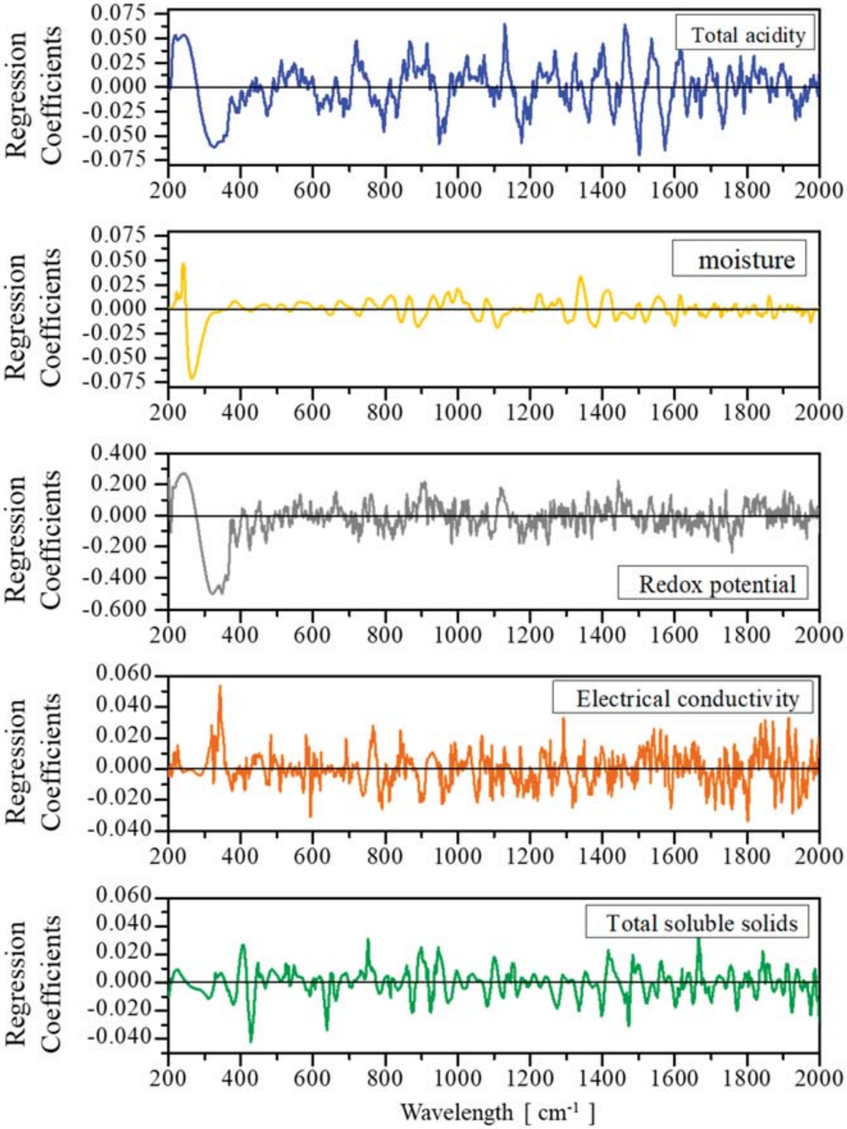


Figure 4. Regression models in the Raman region obtained to predict physical and chemical properties of honey from the state of Campeche.

3. Materials and Methods

3.1. Honey Samples

A total of 189 honey samples were supplied directly from Mayan beekeepers of the state of Campeche, Mexico. The samples were collected between February and June of 2014 and 2015. From the total samples, 175 corresponded to *Apis mellifera* and 14 to *Melipona beecheii*. Figure 5 illustrates the geographical region where the honey samples were collected, which includes the locations of Calakmul (40 samples), Calkiní (14 samples), Campeche (26 samples), Champotón (34 samples), Escárcega (20 samples), Hecelchakán (4 samples), Hopelchén (22 samples), and Sabancuy (29 samples). The predominant floral origin of the honeys was determined according to information provided by the Mayan beekeepers, and included the following: Tahonal (*Viguiera dentata*), Tsíitsilche (*Gymnopodium floribundum*), Ja'abin (*Piscidia piscipula*), Tzalam (*Lysiloma latisiliquum*), Pukte (*Bucida buceras*), Xa'an, huano (*Sabal yapa*), Xtabentum (*Turbina corymbosa*), Palo Tinto (*Haematoxylum campechianum*), Chéechem (*Metopium brownei*), Hulub (*Bravaisia berlandieriana*), Chakàah (*Bursera simaruba*), Suuk, chak lol (*Salvia coccínea*), Box káatsim (*Acacia gaumeri*), Bohom (*Cordia gerascanthus*), Kitim che' (*Caesalpinia gaumeri*), Susuk (*Dyphisa carthagenensis*), Cascarillo (*Erythroxylum confusum*), Machiche (*Lonchocarpus castilloi*) and Mangle negro (*Avicennia germinans*).

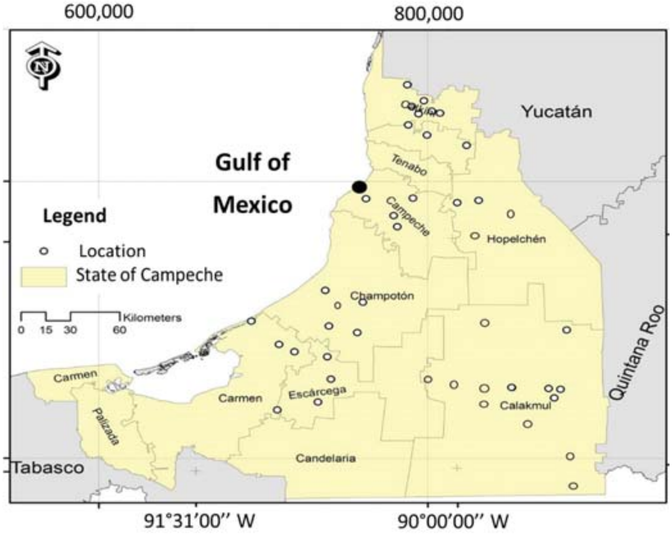


Figure 5. Honey producing communities in the state of Campeche.

3.2. Physicochemical Analysis

The physical-chemical properties of the honey samples were determined according to standards and methods established by Codex Alimentarious [8], International Honey Commission [9], and the Association of Official Analytical Chemists [10]. The honey properties studied were pH, moisture, TSS, free acidity, lactonic acidity, total acidity, EC, Redox potential, HMF, and ash content. The chemical reagents used were standard hydrochloric acid (HCl) solution at 0.05 N, standard sodium hydroxide (NaOH) solution at 0.05 N, deionized water, acetone, buffer solutions, sodium bisulfite (Fermont, Canada), and the reagents Carrez I and Carrez II (Sigma-Aldrich, Saint Louis, MO, USA); all of them of analytical grade. A detailed description of the procedure for obtaining each physical-chemical property is given below.

3.2.1. Moisture and Total Soluble Solid

Moisture and TSS were measured by the refractometric method. One gram of honey was analyzed in an Atago refractometer model PAL-22S (Atago, Tokio, Japan) at 25 °C; TSS was expressed in Brix°, whereas moisture percentage (g/100 g honey) was given according to the method established in [43].

3.2.2. pH, Free Acidity, Lactonic Acidity, and Total Acidity

To determine the pH, 10.0 g of honey was dissolved in 75 mL of deionized water (free CO_2). The solution was analyzed by using a Thermo Scientific brand pH meter (Orion Star A211, Waltham, MA, USA), previously calibrated with standard buffer solutions at pH values of 4–7 and 7–10, respectively. The honey solution was titrated with 0.05 N NaOH until it reached a pH of 8.5 to obtain the free acidity value. Lactonic acidity was determined by adding 10 mL of 0.05 N NaOH to the sample, and then titrating with 0.05 N HCl to return the pH to 8.3. Finally, the total acidity was obtained as the sum of the free acidity and lactonic acidity values, expressed in meq·kg^{-1} [44].

3.2.3. Electrical Conductivity and Redox Potential

The electrical conductivity and Redox potential were measured using a conductivity meter (Thermo Scientific, Waltham, MA, USA), which analyzed a solution composed of 20 g of honey dissolved in 100 mL of deionized water (free CO_2). Measurements were made at 20 °C and the results were expressed in mS·cm^{-1} for electrical conductivity and mV for Redox potential [45,46].

3.2.4. Ash Content and Hydroxymethylfurfural

The determination of ash content was conducted by incineration [47]. Two grams of honey was placed in a crucible and heated in a Lindberg/Blue muffle furnace (Thermo Fisher Scientific, USA) at 650 °C for 6 h. Carbon content results were expressed in g/100 g honey. On the other hand, HMF content was measured based on the standard method [10]. Five grams of honey was dissolved with 25 mL of deionized water (free CO_2) in an (250 mL) Erlenmeyer flask. The solution was clarified by adding 0.5 mL of Carrez I and Carrez II reagents, up to 50 mL. The solution was filtered using Watman paper (No. 42), and subsequently treated with a sodium bisulfite solution. The absorbance was determined on a UV-visible spectrophotometer (DR6000, HACH, Loveland, CO, USA) at wavelengths of 284 and 338 nm. HMF concentration was expressed in mg·kg^{-1}.

3.3. *Raman Analysis*

Honey samples were analyzed in triplicate using a Raman QE65000 spectrometer (Ocean Optics, Edinburgh, UK) equipped with a symmetric crossed Czerny-Turner optical bench, 101 mm focal length, an RPB 785 fiber optic prove, and Hamamatsu S7031-1006 detector with a spectral range between 780–940 nm. The spectrometer was operated with the SPECTRA SUIT software (version 2.0.162, Ocean Optics, Edinburgh, UK) to establish the interface between the computer and the Raman equipment. To perform the analysis of the samples, 30 mL of honey was deposited in an amber glass bottle and subsequently a laser beam was applied at 785 nm with a power of 20 mW for 10 s. All Raman spectra were collected in the range of 0 to 2200 cm^{-1} at 25 °C with a spectral resolution of 1.55 cm^{-1}. The data between 0–200 cm^{-1} and 2001–2200 cm^{-1} were omitted because they had higher spectral noise. Therefore, the spectral data between 201–2000 cm^{-1} was used.

3.4. *Chemometric Model Development*

For the development of the chemometric models, an experimental database was created employing the Raman absorbance (matrix X) and the physical-chemical properties of the analyzed honey samples (vectors Y). Raman analysis results were converted into a data matrix using Microsoft Excel 2013 (Microsoft, Redmond, WA, USA) composed of 900 wavelength values and 567 honey samples (510,300 absorbance samples). The data matrix was transposed and exported to the Pirouette V. 4.5

Software (Infometrix, Bogota, Colombia) to be correlated with each of the physicochemical properties. For the construction of chemometric models, partial least square (PLS) regression was used. In order to minimize spectral noise and errors in the development of the chemometric models, the following mathematical and statistical treatments were applied: auto-scaling or centering and subsequently the treatments baseline correction, smoothing, data normalization, first-order derivation, alignment, Log10 analysis, and Standard Normal Variate (SNV) were performed. To determine the predictability of the models developed, a cross-validation was performed (five out) using 90% of the data. Subsequently, an external validation was carried out with the remaining 10% of the data, which were not used in the construction of the chemometric models. The division of the database for the external calibration and validation processes was carried out by the software Pirouette, implementing the Kennard–Stone selection algorithm [33]. The statistical indicators used during the validation phase were: standard error of calibration (SEC), standard error of prediction (SEP), coefficient correlation of calibration (R^2_{cal}) and coefficient correlation of validation (R^2_{val}), and Student's *t*-test of paired data [33,34]. Figure 6 illustrates the computational procedure for the development of the chemometric models.

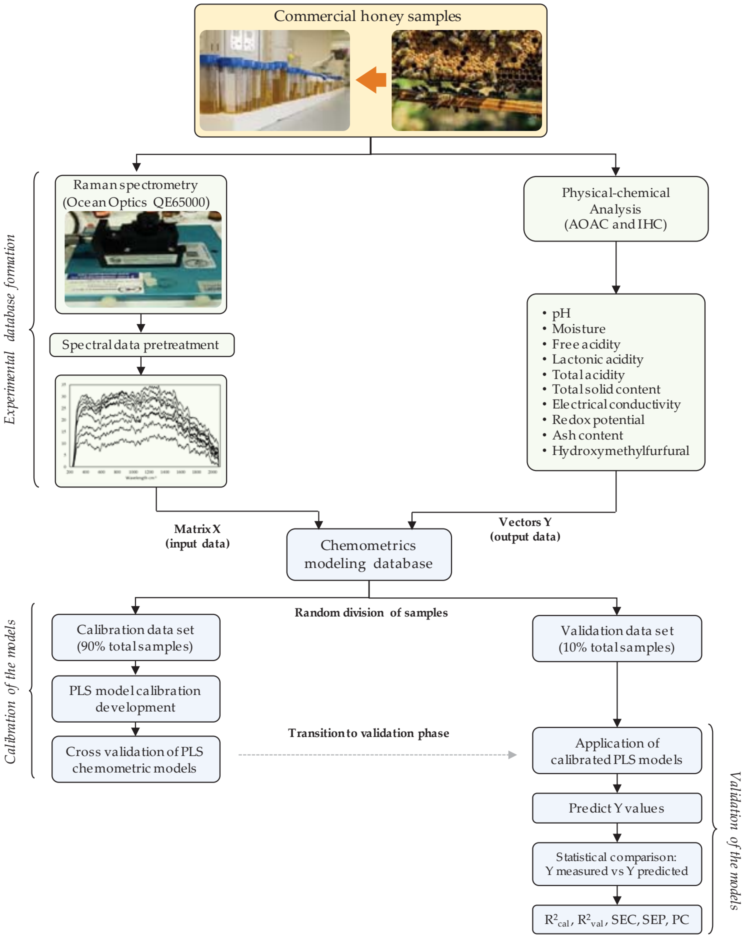

Commercial honey samples
Raman spectrometry (Ocean Optics QE65000)
Physical-chemical Analysis (AOAC and IHC)
Experimental database formation
Spectral data pretreatment
• pH
• Moisture
• Free acidity
• Lactonic acidity
• Total acidity
• Total solid content
• Electrical conductivity
• Redox potential
• Ash content
• Hydroxymethylfurfural
Matrix X (input data)
Chemometrics modeling database
Vectors Y (output data)
Random division of samples
Calibration data set (90% total samples)
Validation data set (10% total samples)
PLS model calibration development
Cross validation of PLS chemometric models
Transition to validation phase
Application of calibrated PLS models
Predict Y values
Statistical comparison: Y measured vs Y predicted
R^2_{cal}, R^2_{val}, SEC, SEP, PC
Calibration of the models
Validation of the models

Figure 6. Schematic diagram of the development and evaluation process of the 10 chemometric models for the estimation of physicochemical properties of honey produced in the region of Campeche, Mexico.

4. Conclusions

In this work, it has been demonstrated that the Raman technique is an analytical tool that has advantages over other conventional techniques for the analysis of honey, since it is friendly to the environment and does not use chemical reagents, obtaining results in less time. Furthermore, it has been demonstrated that chemometric modeling based on Raman technology allows the development of numerical models and good capacity of predicting humidity, free acidity, lactonic acidity, total acidity, and Redox potential for Campechean honeys. The statistical parameters used to evaluate the predictability of each chemometric model show an accuracy similar to the conventional methods established in the standards, with the advantage that they are faster and do not use chemical reagents, so they are more environmentally friendly. Chemometric models to predict the content of HMF and ashes did not achieve good predictive capacity, which can be attributed to the fact that these chemical components are at very low concentrations in honey.

According to the study, the chemometric models that presented adequate prediction results represent an interesting alternative to be used in the development of intelligent portable laboratories that facilitate beekeepers in the region to analyze said chemometric properties at the site. Thus, the models presented represent a low-cost option to contribute significantly to the economic development of the honey industry in the region.

Author Contributions: All authors contributed equally to the design and performance of the experiments, the analysis of the data, and the writing and revision of the manuscript.

Funding: "This research was funded by PROMEP through Project DSA/103.5/14/10634 (UNACAR-EXB076)".

Acknowledgments: The authors thank the Faculty of Chemistry of the Autonomous University of Carmen for the support provided to perform the physical and chemical analysis of honey samples in the laboratories of instrumental analysis and Environmental Engineering. We also thank PROMEP for the financial support granted through Project DSA/103.5/14/10634 (UNACAR-EXB076) and the Mayan beekeepers of the State of Campeche for donating honey samples.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

σ	Standard deviation
R^2_{cal}	Coefficient of determination of cross-validation
R^2_{val}	Coefficient of determination for external validation
t_c	Student's *t*-test confidence value
t_v	Student's *t*-test external validation value
X	Raman spectroscopy matrix data
Y	Vector value of a physicochemical property
ANN	Artificial Neural Networks
AOAC	Association of Official Analytical Chemists
EC	Electrical conductivity
FT	Fourier transform
HCA	Hierarchical Clustering Analysis
HMF	hydroxymethylfurfural
IHC	International Honey Commission
LDA	Linear Discriminant Analysis
LVs	Latent variables
NIRS	Near infrared spectroscopy
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Square
SEC	Standard error of calibration
SEP	Standard error of prediction
SNV	Standard Normal Variate
TSS	Total soluble solids

References

1. Cianciosi, D.; Forbes-Hernández, T.; Afrin, S.; Gasparrini, M.; Reboredo-Rodriguez, P.; Manna, P.; Zhang, J.; Bravo Lamas, L.; Martínez Flórez, S.; Agudo Toyos, P.; et al. Phenolic Compounds in Honey and Their Associated Health Benefits: A Review. *Molecules* **2018**, *23*, 2322. [CrossRef] [PubMed]
2. Ramón-Sierra, J.; Peraza-López, E.; Rodríguez-Borges, R.; Yam-Puc, A.; Madera-Santana, T.; Ortiz-Vázquez, E. Partial characterization of ethanolic extract of Melipona beecheii propolis and in vitro evaluation of its antifungal activity. *Rev. Bras. Farmacogn.* **2019**, *29*, 319–324. [CrossRef]
3. Duca, A.; Sturza, A.; Moacă, E.-A.; Negrea, M.; Lalescu, V.-D.; Lungeanu, D.; Dehelean, C.-A.; Muntean, D.-M.; Alexa, E. Identification of Resveratrol as Bioactive Compound of Propolis from Western Romania and Characterization of Phenolic Profile and Antioxidant Activity of Ethanolic Extracts. *Molecules* **2019**, *24*, 3368. [CrossRef] [PubMed]
4. Przybyłek, I.; Karpiński, T.M. Antibacterial Properties of Propolis. *Molecules* **2019**, *24*, 2047. [CrossRef]
5. Guerrini, A.; Bruni, R.; Maietti, S.; Poli, F.; Rossi, D.; Paganetto, G.; Muzzoli, M.; Scalvenzi, L.; Sacchetti, G. Ecuadorian stingless bee (Meliponinae) honey: A chemical and functional profile of an ancient health product. *Food Chem.* **2009**, *114*, 1413–1420. [CrossRef]
6. Maione, C.; Barbosa, F.; Barbosa, R.M. Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review. *Comput. Electron. Agric.* **2019**, *157*, 436–446. [CrossRef]
7. Se, K.W.; Wahab, R.A.; Syed Yaacob, S.N.; Ghoshal, S.K. Detection techniques for adulterants in honey: Challenges and recent trends. *J. Food Compos. Anal.* **2019**, *80*, 16–32. [CrossRef]
8. Codex Alimentarius Commission Codex Standard for Honey, CODEX STAN 12-1981. Available online: http://www.fao.org/3/w0076e/w0076e30.htm (accessed on 1 November 2019).
9. Bogdanov, S. Harmonised methods of the International Honey Commission. Available online: http://www.ihc-platform.net/ (accessed on 1 November 2019).
10. AOAC. *Official Methods of Analysis of AOAC International*; Association of Official Analysis Chemists International: Rockville, MD, USA, 2005; ISBN 0935584544.
11. Das, C.; Chakraborty, S.; Acharya, K.; Bera, N.K.; Chattopadhyay, D.; Karmakar, A.; Chattopadhyay, S. FT-MIR supported Electrical Impedance Spectroscopy based study of sugar adulterated honeys from different floral origin. *Talanta* **2017**, *171*, 327–334. [CrossRef]
12. Brereton, R.G. Introduction to multivariate calibration in analytical chemistry. *Analyst* **2000**, *125*, 2125–2154. [CrossRef]
13. Aliaño-González, M.J.; Ferreiro-González, M.; Espada-Bellido, E.; Palma, M.; Barbero, G.F. A screening method based on Visible-NIR spectroscopy for the identification and quantification of different adulterants in high-quality honey. *Talanta* **2019**, *203*, 235–241. [CrossRef]
14. Corvucci, F.; Nobili, L.; Melucci, D.; Grillenzoni, F.-V. The discrimination of honey origin using melissopalynology and Raman spectroscopy techniques coupled with multivariate analysis. *Food Chem.* **2015**, *169*, 297–304. [CrossRef] [PubMed]
15. Frausto-Reyes, C.; Casillas-Peñuelas, R.; Quintanar-Stephano, J.; Macías-López, E.; Bujdud-Pérez, J.; Medina-Ramírez, I. Spectroscopic study of honey from Apis mellifera from different regions in Mexico. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2017**, *178*, 212–217. [CrossRef] [PubMed]
16. Jandrić, Z.; Haughey, S.A.; Frew, R.D.; McComb, K.; Galvin-King, P.; Elliott, C.T.; Cannavan, A. Discrimination of honey of different floral origins by a combination of various chemical parameters. *Food Chem.* **2015**, *189*, 52–59. [CrossRef] [PubMed]
17. Oroian, M.; Ropciuc, S. Botanical authentication of honeys based on Raman spectra. *J. Food Meas. Charact.* **2018**, *12*, 545–554. [CrossRef]
18. Özbalci, B.; Boyaci, İ.H.; Topcu, A.; Kadılar, C.; Tamer, U. Rapid analysis of sugars in honey by processing Raman spectrum using chemometric methods and artificial neural networks. *Food Chem.* **2013**, *136*, 1444–1452. [CrossRef]
19. Oroian, M.; Ropciuc, S.; Paduret, S. Honey Adulteration Detection Using Raman Spectroscopy. *Food Anal. Methods* **2018**, *11*, 959–968. [CrossRef]
20. Anjos, O.; Santos, A.J.A.; Paixão, V.; Estevinho, L.M. Physicochemical characterization of Lavandula spp. honey with FT-Raman spectroscopy. *Talanta* **2018**, *178*, 43–48. [CrossRef]

21. Tahir, H.E.; Xiaobo, Z.; Zhihua, L.; Jiyong, S.; Zhai, X.; Wang, S.; Mariod, A.A. Rapid prediction of phenolic compounds and antioxidant activity of Sudanese honey using Raman and Fourier transform infrared (FT-IR) spectroscopy. *Food Chem.* **2017**, *226*, 202–211. [CrossRef]
22. Li, S.; Shan, Y.; Zhu, X.; Zhang, X.; Ling, G. Detection of honey adulteration by high fructose corn syrup and maltose syrup using Raman spectroscopy. *J. Food Compos. Anal.* **2012**, *28*, 69–74. [CrossRef]
23. Yam-Puc, A.; Santana-Hernández, A.A.; Yah-Nahuat, P.N.; Ramón-Sierra, J.M.; Cáceres-Farfán, M.R.; Borges-Argáez, R.L.; Ortiz-Vázquez, E. Pentacyclic triterpenes and other constituents in propolis extract from Melipona beecheii collected in Yucatan, México. *Rev. Bras. Farmacogn.* **2019**, *29*, 358–363. [CrossRef]
24. Porter-Bolland, L.; Ellis, E.A.; Guariguata, M.R.; Ruiz-Mallén, I.; Negrete-Yankelevich, S.; Reyes-García, V. Community managed forests and forest protected areas: An assessment of their conservation effectiveness across the tropics. *For. Ecol. Manage.* **2012**, *268*, 6–17. [CrossRef]
25. Güemes-Ricalde, F.J.; Villanueva-G, R.; Eaton, K.D. Honey production by the Mayans in the Yucatan peninsula. *Bee World* **2003**, *84*, 144–154. [CrossRef]
26. Villanueva-Gutiérrez, R.; Moguel-Ordóñez, Y.B.; Echazarreta-González, C.M.; Arana-López, G. Monofloral honeys in the Yucatán Peninsula, Mexico. *Grana* **2009**, *48*, 214–223. [CrossRef]
27. Mondragón-Cortez, P.; Ulloa, J.A.; Rosas-Ulloa, P.; Rodríguez-Rodríguez, R.; Resendiz Vázquez, J.A. Physicochemical characterization of honey from the West region of México. *CyTA—J. Food* **2013**, *11*, 7–13. [CrossRef]
28. White, J.W. Moisture in Honey Review of Chemical and Physical Methods. *J. Assoc. Off. Anal. Chem.* **1969**, *52*, 729–737.
29. Kek, S.P.; Chin, N.L.; Yusof, Y.A.; Tan, S.W.; Chua, L.S. Classification of entomological origin of honey based on its physicochemical and antioxidant properties. *Int. J. Food Prop.* **2017**, *20*, S2723–S2738. [CrossRef]
30. Belay, A.; Solomon, W.K.; Bultossa, G.; Adgaba, N.; Melaku, S. Physicochemical properties of the Harenna forest honey, Bale, Ethiopia. *Food Chem.* **2013**, *141*, 3386–3392. [CrossRef]
31. Jimenez, M.; Beristain, C.I.; Azuara, E.; Mendoza, M.R.; Pascual, L.A. Physicochemical and antioxidant properties of honey from Scaptotrigona mexicana bee. *J. Apic. Res.* **2016**, *55*, 151–160. [CrossRef]
32. Sereia, M.J.; Marҫo, P.H.; Perdoncini, M.R.G.; Parpinelli, R.S.; de Lima, E.G.; Anjo, F.A. Techniques for the Evaluation of Physicochemical Quality and Bioactive Compounds in Honey. In *Honey Analysis*; InTech: London, UK, 2017.
33. Kizil, R.; Irudayaraj, J.; Seetharaman, K. Characterization of Irradiated Starches by Using FT-Raman and FTIR Spectroscopy. *J. Agric. Food Chem.* **2002**, *50*, 3912–3918. [CrossRef]
34. Zhu, X.; Li, S.; Shan, Y.; Zhang, Z.; Li, G.; Su, D.; Liu, F. Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics. *J. Food Eng.* **2010**, *101*, 92–97. [CrossRef]
35. Salvador, L.; Guijarro, M.; Rubio, D.; Aucatoma, B.; Guillén, T.; Vargas Jentzsch, P.; Ciobotă, V.; Stolker, L.; Ulic, S.; Vásquez, L.; et al. Exploratory Monitoring of the Quality and Authenticity of Commercial Honey in Ecuador. *Foods* **2019**, *8*, 105. [CrossRef] [PubMed]
36. El Sohaimy, S.A.; Masry, S.H.D.; Shehata, M.G. Physicochemical characteristics of honey from different origins. *Ann. Agric. Sci.* **2015**, *60*, 279–287. [CrossRef]
37. Cozzolino, D.; Corbella, E.; Smyth, H.E. Quality Control of Honey Using Infrared Spectroscopy: A Review. *Appl. Spectrosc. Rev.* **2011**, *46*, 523–538. [CrossRef]
38. Ruoff, K.; Luginbühl, W.; Bogdanov, S.; Bosset, J.-O.; Estermann, B.; Ziolko, T.; Kheradmandan, S.; Amad/, R. Quantitative determination of physical and chemical measurands in honey by near-infrared spectrometry. *Eur. Food Res. Technol.* **2007**, *225*, 415–423. [CrossRef]
39. Silva, L.R.; Videira, R.; Monteiro, A.P.; Valentão, P.; Andrade, P.B. Honey from Luso region (Portugal): Physicochemical characteristics and mineral contents. *Microchem. J.* **2009**, *93*, 73–77. [CrossRef]
40. Karabagias, I.K.; Badeka, A.; Kontakos, S.; Karabournioti, S.; Kontominas, M.G. Characterisation and classification of Greek pine honeys according to their geographical origin based on volatiles, physicochemical parameters and chemometrics. *Food Chem.* **2014**, *146*, 548–557. [CrossRef]
41. Khalil, M.I.; Moniruzzaman, M.; Boukraâ, L.; Benhanifia, M.; Islam, M.A.; Islam, M.N.; Sulaiman, S.A.; Gan, S.H. Physicochemical and Antioxidant Properties of Algerian Honey. *Molecules* **2012**, *17*, 11199–11215. [CrossRef]
42. Siddiqui, A.J.; Musharraf, S.G.; Choudhary, M.I.; Rahman, A.-. Application of analytical methods in authentication and adulteration of honey. *Food Chem.* **2017**, *217*, 687–698. [CrossRef]

43. Chen, C. Relationship between Water Activity and Moisture Content in Floral Honey. *Foods* **2019**, *8*, 30. [CrossRef]
44. Lemos, M.S.; Venturieri, G.C.; Dantas Filho, H.A.; Dantas, K.G.F. Evaluation of the physicochemical parameters and inorganic constituents of honeys from the Amazon region. *J. Apic. Res.* **2018**, *57*, 135–144. [CrossRef]
45. Lichtenberg-Kraag, B.; Hedtke, C.; Bienefeld, K. Infrared spectroscopy in routine quality analysis of honey. *Apidologie* **2002**, *33*, 327–337. [CrossRef]
46. Yücel, Y.; Sultanogˇlu, P. Characterization of honeys from Hatay Region by their physicochemical properties combined with chemometrics. *Food Biosci.* **2013**, *1*, 16–25. [CrossRef]
47. Grazia Mignani, A.; Ciaccheri, L.; Mencaglia, A.A.; Di Sanzo, R.; Carabetta, S.; Russo, M. Dispersive Raman Spectroscopy for the Nondestructive and Rapid Assessment of the Quality of Southern Italian Honey Types. *J. Light. Technol.* **2016**, *34*, 4479–4485. [CrossRef]

Sample Availability: The data used to support the findings of this study are available from the first author upon request.

molecules

MDPI

Article

Geographical Authentication of *Macrohyporia cocos* by a Data Fusion Method Combining Ultra-Fast Liquid Chromatography and Fourier Transform Infrared Spectroscopy

Qin-Qin Wang [1,2], Heng-Yu Huang [2,*] and Yuan-Zhong Wang [1,*]

[1] Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, Kunming 650200, China; wqq6501@163.com

[2] College of Traditional Chinese Medicine, Yunnan University of Traditional Chinese Medicine, Kunming 650500, China

[*] Correspondence: hhyhhy96@163.com (H.-Y.H.); boletus@126.com (Y.-Z.W.); Tel.: +86-871-6503-3564 (H.-Y.H.); +86-871-6503-3575 (Y.-Z.W.)

Academic Editor: Marcello Locatelli
Received: 6 March 2019; Accepted: 1 April 2019; Published: 3 April 2019

Abstract: *Macrohyporia cocos* is a medicinal and edible fungi, which is consumed widely. The epidermis and inner part of its sclerotium are used separately. *M. cocos* quality is influenced by geographical origins, so an effective and accurate geographical authentication method is required. Liquid chromatograms at 242 nm and 210 nm (LC_{242} and LC_{210}) and Fourier transform infrared (FTIR) spectra of two parts were applied to authenticate the geographical origin of cultivated *M. cocos* combined with low and mid-level data fusion strategies, and partial least squares discriminant analysis. Data pretreatment involved correlation optimized warping and second derivative. The results showed that the potential of the chromatographic fingerprint was greater than that of five triterpene acids contents. LC_{242}-FTIR low-level fusion took full advantage of information synergy and showed good performance. Further, the predictive ability of the FTIR low-level fusion model of two parts was satisfactory. The performance of the low-level fusion strategy preceded those of the single technique and mid-level fusion strategy. The inner parts were more suitable for origin identification than the epidermis. This study proved the feasibility of the data fusion of chromatograms and spectra, and the data fusion of different parts for the accurate authentication of geographical origin. This method is meaningful for the quality control of food and the protection of geographical indication products.

Keywords: *Macrohyporia cocos*; data fusion; liquid chromatography; fourier transform infrared spectroscopy; partial least squares discriminant analysis; authentication

1. Introduction

The dried sclerotium of *Macrohyporia cocos*, belonging to Polyporaceae, is an herbal medicine (called Poria) that can be used as food, and has been approved by the National Health Commission of the People's Republic of China. It plays an indispensable role in numerous drugs, such as the liquid oral formulation of *Poriacocos* polysaccharides, Sijunzi Tang, Liuwei Dihuang Wan and Chuanbei Pipa Gao. Various kinds of Poria-based foods and skin cosmetics such as sleep-friendly tea, Tuckahoe pie, Guiling jelly (drinks made from turtle shell and medicinal herbs), Guiling jelly soft candy and the Poria facial mask, are pretty popular. Present phytochemical investigation suggests that this fungus contains terpenes and polysaccharides, which present beneficial biological properties, such as a prebiotic effect, through the modulation of gut microbiota composition [1], anti-hyperlipidemic [2], anti-cancer [3] hepatoprotective [4] and affecting adipocyte and osteoblast differentiation effects [5].

Generally, the sclerotium of *M. cocos* is peeled and processed into two products, the epidermis and the inner part. The epidermis is called Poriae Cutis in Chinese, and the inner part is still called Poria. The epidermis and inner part have similar types of compounds and different secondary metabolites contents [6], which are often used and studied separately. Both Poria and Poriae Cutis are officially recorded in the Chinese Pharmacopoeia.

The provenance of *M. cocos* is mainly distributed in the Dabie mountains area and Yunnan Province of China. Yunnan is suggested as the most satisfactory habitat because the quality of Yunnan *M. cocos* is being highly recommended all the time. Due to the large demand for it, and the knowledge of cultivation mastered easily by common people, this fungus is cultivated in large quantities. Although *M. cocos* is cultivated in Yunnan, the chemical profiles influencing biological activities may be uneven owing to various cultivation sites and different management techniques. It was reported in a previous study that the contents of pachymic acid of *M. cocos* in different regions of Yunnan varied significantly [7]. Consequently, customers are increasingly demanding some sort of proof of the geographical origin. For the sake of response to the demand, it is necessary to conduct research with respect to the authentication of geographical origin, which can also provide basic technology for the protection of specific geographical indication products [8].

To date, various analytical technologies that respond to the different chemical information of samples have been implemented for the origin identification of *M. cocos* [9–11]. Although these methods proved promising for the discrimination of provenance, they were separately applied. Nowadays, data fusion has been applied in the fields of food and medicine [12,13]. For example, Ni et al. [14] discovered that, based on high-performance liquid chromatography (HPLC) and Fourier transform infrared spectroscopy (FTIR) data fusion, the type and geographical origin of *Radix Paeoniae* samples could be successfully discriminated, and the fused data matrix showed a prominent result compared with the independent technique.

Data fusion strategies, which fuse the outputs of multiple complementary information to provide rich knowledge about a sample, are hoped to achieve a more accurate characterization than single pieces of information [15]. In addition to the fusion of several datum regarding one sample, the fusion of information regarding different parts was reported. For instance, Casale et al. [16] combined the near-infrared information obtained by the three parts (pileipellis, flesh and hymenium) of each individual to check the authenticity of dried *porcini* mushrooms. Studies mentioned above demonstrated that although time and effort would be taken to collect multiple complementary data, data fusion was suggested as an alternative strategy to show accurate characterization.

Infrared spectroscopy can provide the molecular functional group structure of metabolites. Liquid chromatography can characterize the exist of compounds and determinate the special compounds. Both techniques present different and complementary information, which were used for data fusion in this study. To the best of our knowledge, infrared spectroscopy was widely used for geographical classification because of the features of simplicity and rapidity [17,18]. Liquid chromatography was almost used for determining the contents of compounds [19,20]. Multiple chromatographic data fusion has been merely reported in the authentication of the geographical origin of palm oil [21], predicting antioxidant activity of *Turnera diffusa* [22], authentication of *Valeriana* species [23] as well as a comparison of *Salvia miltiorrhiza* and its variety [24]. Actually, a wealth of information was contained in the chromatographic data, and due to extensive automation, a stable and reliable result could be obtained.

In this study, two data fusion strategies including low and mid-level fusion as well as two data combinations including the fusion of complementary information regarding a single part, and the fusion of information regarding two medicinal parts from one sclerotium were performed for the geographical authentication of *M. cocos*. Liquid chromatograms at two wavelengths (242 nm and 210 nm) and FTIR spectra of two medicinal parts (Poria and Poriae Cutis) of *M. cocos* were analyzed.

Contents of five triterpene acids were measured. Chromatographic data fusion, spectral data fusion as well as chromatography and spectroscopy data fusion were implemented, combined with partial least squares discriminant analysis (PLS-DA).

2. Results and Discussion

2.1. Spectral Analysis

FTIR is an auxiliary method in the structural elucidation of organic compounds, which is also employed to assess the quality attributes of a product and authenticate geographic location [17]. With the characteristics of easy operation and rapid acquisition, it was applied to the identification of cultivation location of *M. cocos*. The second derivative spectra of samples from each geographic origin were given in Figure 1, and absorption peaks were observed in the form of negative peaks. Because a 2600–1750 cm^{-1} signal was caused by ATR crystal material [25], it was discarded and did not present in the Figure.

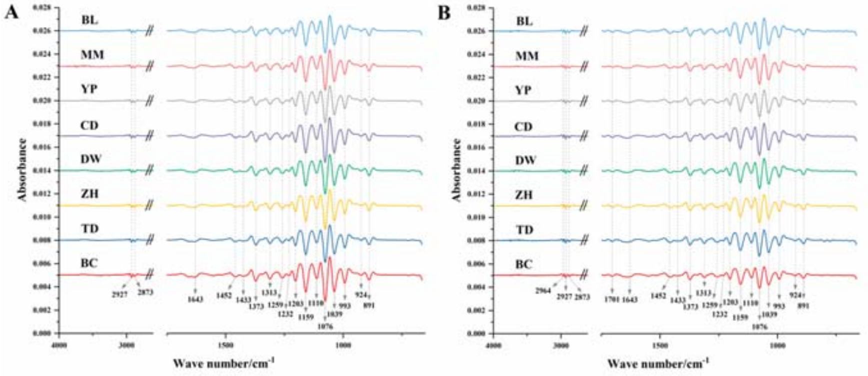


Figure 1. Second derivative spectra of Poria (**A**) and Poriae Cutis (**B**) samples from eight geographic origins.

Absorption bands at 2964 and 1704 cm^{-1} were just observed in Poriae Cutis samples. A disparity of absorption intensity exhibited in samples from different cultivation locations. Relatively high absorbance values were at around 1200–950 cm^{-1}, which were mainly caused by C-O stretching vibration, C-C stretching vibration and C-OH bending vibration of polysaccharides [26,27]. Peaks located at 2964 and 2873 cm^{-1} correspond to C-H antisymmetric and symmetrical stretching vibration of methyl group respectively, while the peak at 2927 cm^{-1} is assigned to C-H antisymmetric stretching vibration of methylene. The absorption at 1452 cm^{-1} and 1373 cm^{-1} belonged to C-H antisymmetric and symmetrical bending vibration of methyl [11]. The peak at 1643 cm^{-1} was assigned to C=O antisymmetric stretching vibration, which was related to triterpenes [28]. The band at 1704 cm^{-1} was associated with C=O group of esters [29,30]. The band at 891 cm^{-1} was assigned to the bending vibration of the C=CH$_2$ functional group [28]. The peak at 1259 cm^{-1} may be related to the amide III band [31]. In total, FTIR spectrum reflected comprehensive structural information of components in *M. cocos* samples, like triterpenes, polysaccharides, and so on.

2.2. Quantitative Analysis of Five Triterpene Acids

The content of each triterpene acid was calculated by their calibration curves and result of the validation of quantitative method was presented in Tables S1 and S2. The calibration curves of five compounds showed good linearity ($R^2 \geq 0.99$). Recovery rates calculated by the standard addition method varied from 96.32% to 106.4%. Values of relative standard deviation (RSD) of intra-day and inter-day precision were less than 1.24% and 5.68%, respectively. RSDs of repeatability did not exceed 5.95% after analyzing six solutions from the same sample in parallel. RSDs of stability were less than 0.71% after detecting a sample solution at 0, 6, 12, 17, 21 and 24 h, respectively. The method validation above indicated that the quantitative analysis was feasible. In particular, due to the obvious difference in the contents of poricoic acid A in Poria and Poriae Cutis samples, the calibration curves in two concentration ranges were prepared separately.

Contents of five triterpene acids were displayed as box-plot given in Figure 2. One-way analysis of variance was computed by SPSS 21.0 software (IBM Corporation, Armonk, NY, USA) to display the difference among eight cultivated locations. A value of $p < 0.05$ was considered significant. Poricoic acid A contents of Mengmeng were significantly different from those of Beicheng, Tuodian and Zhanhe in inner parts, and Yongping in cutis samples. Contents of dehydropachymic acid and pachymic acid in inner parts from Caodian were higher than those of other geographical origins except for Baliu. Inner parts from Baliu showed higher contents of dehydropachymic acid than those from Beicheng, Dawen and Mengmeng, and higher contents of pachymic acid than those from Tuodian, Yongping, Beicheng and Mengmeng. Inner parts from Dawen contained fairly low contents of dehydrotrametenolic acid compared with those from others with the exception of Baliu. Compared with epidermis samples from Dawen, Beicheng and Yongping showed higher contents of dehydrotumulosic acid, and Caodian and Baliu presented higher amount of pachymic acid. From the results, it was found that it was difficult to distinguish *M. cocos* samples from eight cultivation origins just in terms of contents of several target compounds. Therefore, it was necessary to take full advantage of the chromatographic fingerprint, namely, the intensity data for each retention time, to extract more information related to cultivation location.

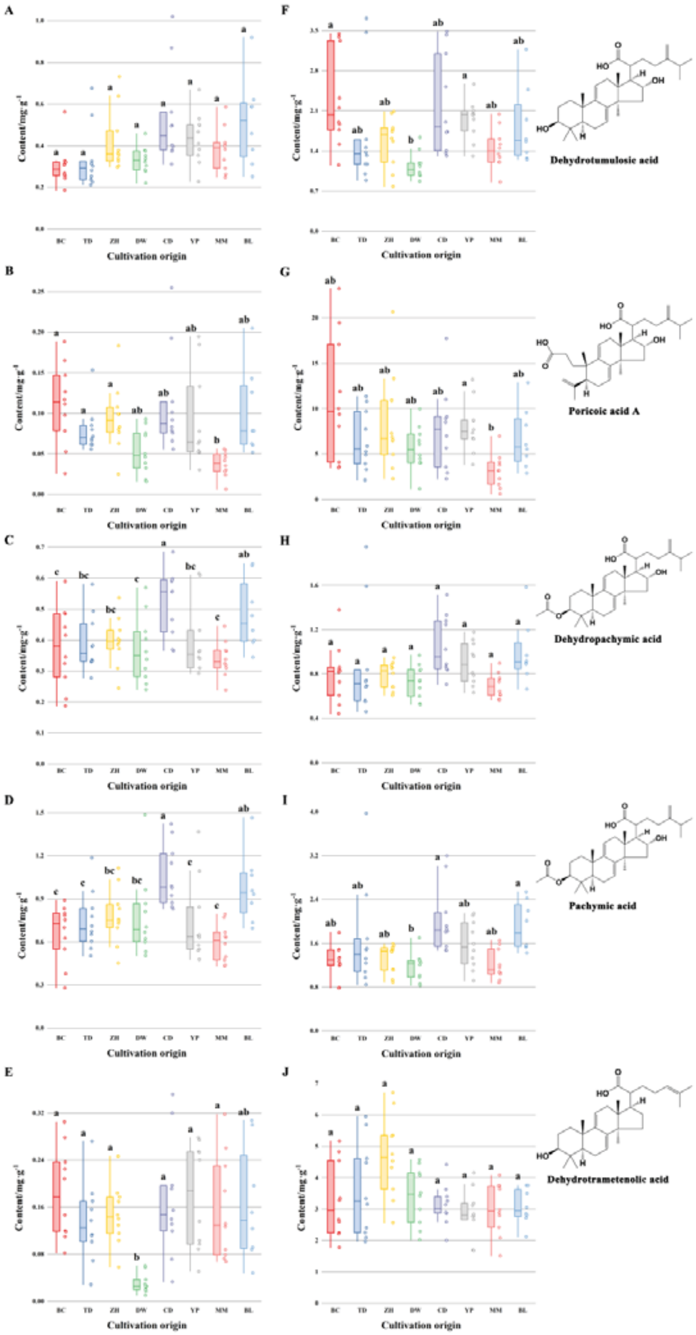


Figure 2. Box-plots of contents of five triterpene acids of Poria (**A**–**E**) and Poriae Cutis (**F**–**J**) samples from eight geographical origins. Note: Different letters show significant difference ($p < 0.05$).

2.3. Chromatographic Data Preprocessing

The chromatograms recorded at 242 nm in Figure S1 were obtained by analyzing the solution from the same sample five times successively within a day and on two consecutive days. Obviously, the retention time of each peak shifted in two days, which was inconvenient for the qualitative results of chemometric analyses. Hence, all of the chromatographic data should be aligned prior to further analysis.

The correlation optimized warping algorithm proposed by Skov et al. [32] was used to correct the retention time shifts among samples. The chromatogram that was most similar to all others was selected to be the reference chromatogram for alignment. The global search space was set to a combination of segment length from 10 to 200 and a slack size from 1 to 20 according to the observed peak widths and shifts on the chromatograms. Then the optimal combination of segment length slack size was automatically selected according to the criterion of well alignment while at the same time considering the preservation in peak shape and area. The theory for the algorithms with respect to the automated alignment of chromatographic data can be consulted in [32].

As a result, suitable combinations of segment length and slack size were achieved for chromatographic data at 242 nm of Poria (197 and 11), 210 nm of Poria (105 and 16), 242 nm of Poriae Cutis (105 and 11) and 210 nm of Poriae Cutis (198 and 16), respectively. Figure 3 presented the aligned *M. cocos* chromatographic fingerprints using these warping parameters, which displayed that the retention time shifts were properly corrected. What's more, it was observed that chromatograms of the same medicinal part recorded at 242 nm and 210 nm showed complementary information, i.e., some peaks obviously presented in liquid chromatograms at 242 nm (LC_{242}) and some compounds just displayed in liquid chromatograms at 210 nm (LC_{210}). Further, chromatograms of two parts were appreciably different. In other words, multiple chromatographic profiles presented abundant chemical information of *M. cocos* that probably facilitated to confirm cultivation areas.

The chromatographic data of one Poria sample and one Poriae Cutis sample could be represented as 7201 and 7801 data points, respectively. In order to save the time for calculation, the number of data points in the retention time dimension of the matrix was reduced by taking one in every three points without affecting the chromatographic features. Therefore, 2401 and 2601 data points were obtained after reducing data, respectively. It was proved that this method was feasible by comparing the PLS-DA results since reducing data had little influence on identifying different groups (Table S3). Additionally, the first 11 min data in the chromatogram that mainly comprised unseparated peaks and baseline shift (Figure 3), which were discarded to obtain representative fingerprints and accurate results. In this way, the final data points were 1960 and 2160, respectively.

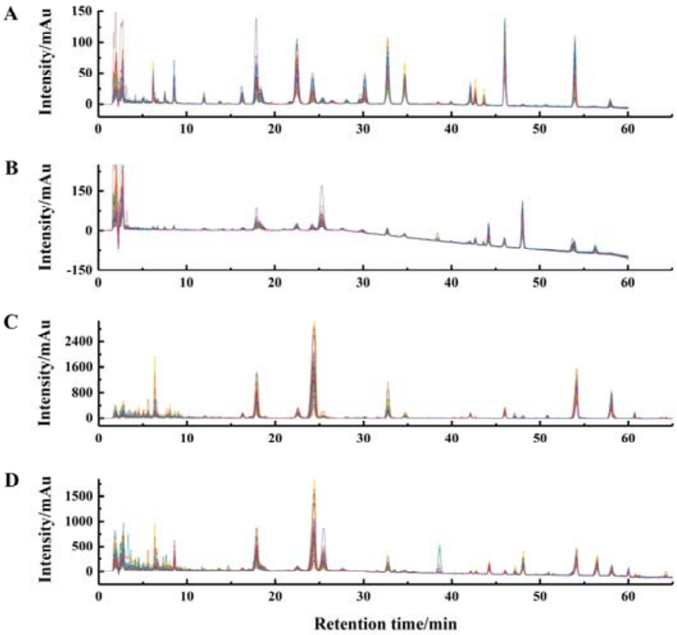


Figure 3. Chromatograms of Poria (**A**,**B**) and Poriae Cutis (**C**,**D**) recorded at 242 (**A**,**C**) and 210 nm (**B**,**D**) after the transformation of correlation optimized warping.

2.4. PLS-DA Using Chromatograms and FTIR Spectra

Partial least squares discriminant analysis is a widely-used linear classification method [33–36]. The selection of the optimal number of latent variables was an essential question for PLS-DA model, which was implemented on the basis of 7-fold cross validation procedure in present study. Unit variance scaling, which could give all variables of the same or different measurements equal importance, was performed by default when developing each PLS-DA model. The parameters of classification models were shown in Table 1 and Tables S4–S6 in detail.

Based on the preprocessing of chromatograms and FTIR spectra, a model of PLS-DA was established using the single dataset (Table 1 and Table S4). The LC_{210} dataset of Poriae Cutis samples did not build model successfully, so results of classification were not listed. FTIR and LC_{242} datasets showed better performance with higher accuracy not only in calibration set but in validation set than LC_{210} dataset. The sensitivity values of class 2 and class 8 in the validation set were 1 for Poria LC_{242} model and were smaller values for the Poria FTIR model, which indicated that LC_{242} model had stronger ability to correctly recognizing samples of class 2 and class 8. While the sensitivity of class 1 and 7 in calibration set was 0.8571 for Poria LC_{242} model smaller than that of Poria FTIR model, indicating that FTIR model had stronger ability to correctly recognizing samples of class 1 and class 7. Moreover, LC models of Poriae Cutis samples presented poorer results than those of Poria samples, which reflected the difference of two medicinal parts of *M. cocos*.

Table 1. The major parameters of PLS-DA model.

Fusion Approach	Data Matrix		Calibration Set			Validation Set
			R^2(cum)	Q^2(cum)	Accuracy	Accuracy
single technique	Poria	FTIR	0.8883	0.7268	100%	92.31%
		LC_{242}	0.6634	0.5277	96.15%	100%
		LC_{210}	0.5174	0.4012	90.38%	76.92%
	Poria Cutis	FTIR	0.9292	0.6981	100%	96.15%
		LC_{242}	0.2874	0.2204	65.38%	34.62%
low-level data fusion	Poria	FTIR-LC_{242}	0.9599	0.7917	100%	100%
		FTIR-LC_{210}	0.9468	0.7663	100%	100%
		$LC_{242\text{-}210}$	0.8097	0.6547	98.08%	92.31%
		FTIR-$LC_{242\text{-}210}$	0.8823	0.7566	100%	100%
	Poria Cutis	FTIR-LC_{242}	0.9016	0.7032	100%	100%
		FTIR-$LC_{242\text{-}210}$	0.905	0.698	100%	100%
	combination data of two medicinal parts	FTIR	0.9548	0.8064	100%	100%
		LC_{242}	0.8147	0.6495	100%	100%
		LC_{210}	0.6489	0.4806	94.23%	88.46%
mid-level data fusion	Poria	FTIR-LC_{242}	0.8266	0.5745	100%	100%
		FTIR-LC_{210}	0.7453	0.5053	96.15%	96.15%
		FTIR-$LC_{242\text{-}210}$	0.8286	0.5882	100%	100%
	Poria Cutis	FTIR-LC_{242}	0.7386	0.5493	100%	92.31%
		FTIR-LC_{210}	0.7518	0.4991	100%	96.15%
		$LC_{242\text{-}210}$	0.4617	0.228	76.92%	73.08%
		FTIR-$LC_{242\text{-}210}$	0.7607	0.5558	100%	96.15%
	combination data of two medicinal parts	FTIR	0.7564	0.5982	98.08%	88.46%
		LC_{242}	0.7761	0.4973	98.08%	100%
		LC_{210}	0.676	0.3756	96.15%	88.46%

Variable importance for the projection (VIP) plot [37] was used for assessing the significance of variable, and that the VIP score of retention time was greater than one means the compound separated at the time was important on distinguishing different cultivation origins. As an example of the Poria LC_{242} model, there were lots of variables whose VIP were higher than one including the corresponding retention time of poricoic acid A and dehydrotrametenolic acid (Figure 4). It indicated that the potential of the chromatographic fingerprint from the aspect of origin identification was greater than that of the contents of several compounds. However, all single technique models did not achieve a perfect performance, so it was necessary to carry out the data fusion strategy that was expected to enhance the classification and prediction ability of the model.

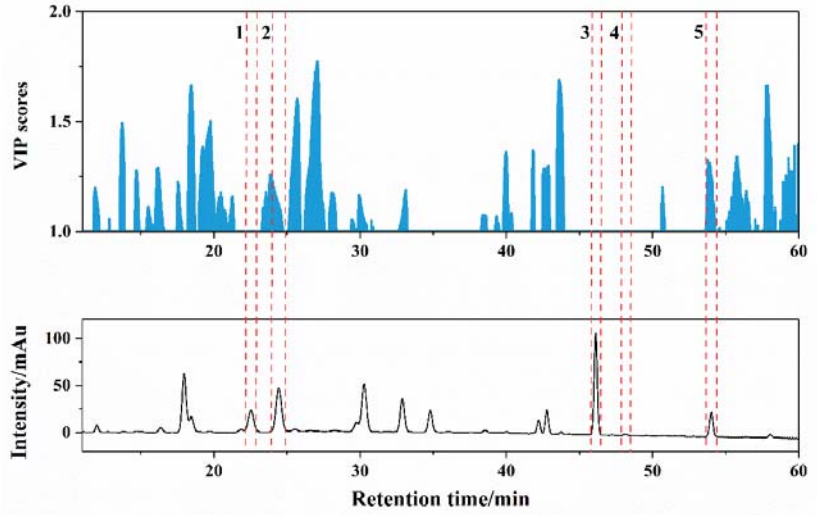


Figure 4. VIP scores of PLS-DA using LC_{242} chromatogram data of Poria samples. Note: 1, dehydrotumulosic acid; 2, poricoic acid A; 3, dehydropachymic acid; 4, pachymic acid; 5, dehydrotrametenolic acid.

2.5. Low-Level Data Fusion

2.5.1. PLS-DA of Poria

Figure 5 was the workflow of geographical authentication using data fusion, which was helpful to understand how data was combined. As shown in Table 1, accuracy rates of low-level data fusion datasets about Poria samples were 100% and higher than those of single technique models except for the model using $LC_{242\text{-}210}$ data, which implied that these models had strong classification performance. The highest R^2(cum) (0.9599) and Q^2(cum) (0.7917) were observed in FTIR-LC_{242} model, indicating a high goodness of fit for the established model in the data and good predictive ability. Therefore, the combination of FTIR and LC_{242} datasets was deemed a suitable strategy, and the fusion of three datasets was unnecessary and verbose. Furthermore, compared with the $LC_{242\text{-}210}$ model, the accuracy of FTIR-LC_{210} model was higher both in calibration and validation sets. It could be interpreted that FTIR dataset provided more helpful information to identify eight geographical origins than LC_{242} dataset in data fusion model of Poria samples. By analogy, it was found that FTIR data showed more contribution for origin discrimination than LC_{210} data when compared $LC_{242\text{-}210}$ model with FTIR-LC_{242} model.

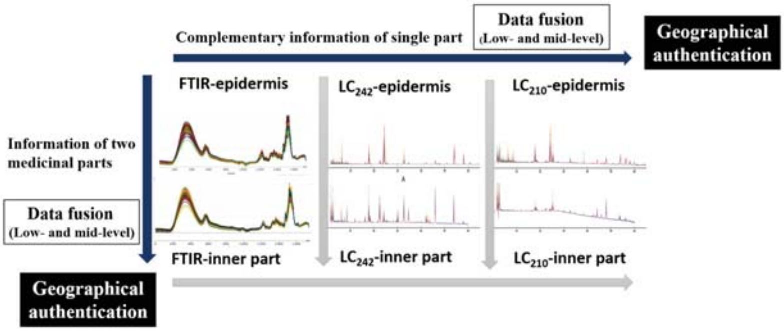


Figure 5. The workflow of geographical authentication using data fusion.

2.5.2. PLS-DA of Poriae Cutis

The accuracy of FTIR-LC$_{242}$ and FTIR-LC$_{242\text{-}210}$ models was 100%, which was greater than that of the models using the independent technique. It indicated the effectiveness of low-level data fusion. The similar Q^2(cum) of FTIR-LC$_{242}$ and FTIR-LC$_{242\text{-}210}$ models was observed. Accordingly, FTIR-LC$_{242}$ was considered as a preferred combination, and the fusion of three datasets was superfluous. Furthermore, the Q^2(cum) values of low-level fusion models about Poriae Cutis samples ($\leq$ 0.7032) were less than those of corresponding models about Poria samples (> 0.75), indicating that Poria samples were more suitable for origins identification than Poriae Cutis species. In the developing LC$_{242\text{-}210}$ and FTIR-LC$_{210}$ low-level models, latent variables could not be calculated so the models were not successfully built. It was in consistent with the state that epidermis LC$_{210}$ dataset did not built PLS-DA model, which was probably attributed by a lot of irrelevant classification information included in LC$_{210}$ dataset of epidermis.

2.5.3. PLS-DA of Combination Data of Two Medicinal Parts

Both FTIR and LC$_{242}$ datasets of two parts samples showed better performance than LC$_{210}$ dataset, which was in accordance with the results of single technique mentioned above. Compared with the single spectrum or chromatogram, data fusion of two medicinal parts proved more advantageous with greater sensitivity, specificity and efficiency. Therein, the FTIR fusion model of two part samples presented the best prediction performance from the Q^2(cum) point of view. What's more, compared with FTIR-LC$_{242}$ model of Poria samples, the Q^2(cum) of LC$_{242}$ fusion model of two parts was smaller. It could be interpreted that Poria FTIR dataset provided more helpful information to predict different geographical origins than Poriae Cutis LC$_{242}$ dataset in data fusion model. By analogy, it was found that the contribution of FTIR dataset was always more than that of LC$_{242}$ and LC$_{210}$ datasets in low-level data fusion. The low-level data fusion strategy has achieved a good classification result, but the mid-level data fusion could spend less computation time compared to the low level. Therefore, mid-level fusion was performed.

2.6. Mid-Level Data Fusion

2.6.1. The Extraction of Feature Variables

Mid-level fusion needed to first extract relevant features from each dataset independently and then concatenated them into a new matrix employed for origins identification. Principal component analysis (PCA) is a dimension reduction technique that creates a small number of new variables called principal components (PCs) from a large number of original variables, which would be applied to extract features. These PCs almost retain the same information as the original variables [38]. The optimal number of PCs was determined by 7-fold cross-validation procedure. The results of feature extraction were shown in Table S7. As an example of LC_{210} dataset of Poria samples, the first thirteen PCs were extracted, which account for 90.92% of the information concerning the original variables. Then the scores of the thirteen PCs were used for data fusion.

2.6.2. PLS-DA of Poria

In agreement with the results of low-level data fusion, the accuracy rates of FTIR-LC_{242} and FTIR-$LC_{242\text{-}210}$ of Poria samples were 100% not only in calibration set but in validation set. And they had stronger recognition performance with higher sensitivity, specificity, efficiency than corresponding single dataset. Nonetheless, all Q^2(cum) values of mid-level data fusion models of Poria samples were less than those of low-level data fusion models, indicating that low-level fusion presented stronger prediction ability than mid-level fusion according to cross validation.

As always, The $LC_{242\text{-}210}$ fusion model did not build successfully. The fusion of LC_{242} and LC_{210} could not gain satisfactory discrimination and even could not construct the model, and it was likely caused by the similar chemical information provided by both chromatograms. Although they presented different peak shapes, there were many common chromatographic peaks that did not provide complementary and useful information.

2.6.3. PLS-DA of Poriae Cutis

$LC_{242\text{-}210}$ model that was not built successfully in low-level fusion finished construction in mid-level fusion. The fact indicated the significance of mid-level data fusion and might be due to the feature extraction. The accuracy rates of FTIR-LC_{210} and FTIR-$LC_{242\text{-}210}$ models were equal, but the detail of incorrect identification was different from sensitivity and specificity points of view. Further analysis showed that one sample belonging to Tuodian was judged as the sample from Baliu in FTIR-LC_{210} model and Mengmeng in FTIR-$LC_{242\text{-}210}$ model by mistake, respectively. FTIR-LC_{242} and FTIR-$LC_{242\text{-}210}$ mid-level fusion models of Poriae Cutis samples presented poorer results than those of Poria samples as well as low-level data fusion models and FTIR model of epidermis samples.

2.6.4. PLS-DA of Combination Data of Two Medicinal Parts

Both FTIR data fusion and LC_{242} data fusion of two medicinal parts had stronger recognition ability when compared to the LC_{210} combination. Both LC_{242} and LC_{210} of two medicinal parts improved performance of single LC_{242} and LC_{210} models. However, the result of FTIR was the opposite. Compared to low-level data fusion, the identification ability of mid-level data fusion did not show any obvious advantage. This might be due to the limitation of our method of feature extraction. In terms of FTIR datasets, only more than 73.29% original information (Table S7) was extracted.

To validate the performance of the PLS-DA model, a 30-iteration permutation test was performed. As shown in Figure S2 that one of permutations plots for Poria $LC_{242\text{-}210}$ model, all permutated Q^2 and R^2 values (bottom left) were lower than the corresponding original values (top right). It indicated that the PLS-DA model was considered as an appropriate model without randomness and overfitting. The results showed that all the PLS-DA models were not overfitting.

3. Materials and Methods

3.1. Reagents, Solvents and Standard References

Dehydrotumulosic acid (purity ≥ 96%) was supplied by ANPEL Laboratory Technologies Inc. (Shanghai, China). Dehydropachymic acid, pachymic acid, poricoic acid A and dehydrotrametenolic acid (purity ≥ 98%) were purchased from Beijing Keliang Technology Co., Ltd. (Beijing, China). HPLC grade acetonitrile and formic acid were purchased from Thermo Fisher Scientific (Fair Lawn, NJ, USA) and Dikma Technologies (Lake Forest, CA, USA), respectively. Purified water was purchased from Guangzhou Watsons Food & Beverage Co., Ltd. (Guangzhou, China). Other chemicals and reagents were analytical grade.

3.2. Samples

Seventy-eight intact cultivated *M. cocos* sclerotia (Figure 6) from eight geographical origins of Yunnan Province, China were collected and identified by Prof. Yuanzhong Wang (Institute of Medicinal Plant, Yunnan Academy of Agricultural Sciences, Kunming, China). Voucher specimens (FL20160217) were deposited in the herbarium of Institute of Medicinal Plant, Yunnan Academy of Agricultural Sciences. After digging sclerotium up, the soil was brushed away. Fresh *M. cocos* sclerotium was air-dried in the shade and then peeled. Both the epidermis and inner part of the dried sclerotium, i.e., Poria and Poriae Cutis, were powdered to a homogeneous size using pulverizer and sieved through No. 60 mesh sieve. The powder was stored in the airproof, dry and dark condition prior to analysis. Detailed information of samples was summarized in Table 2.

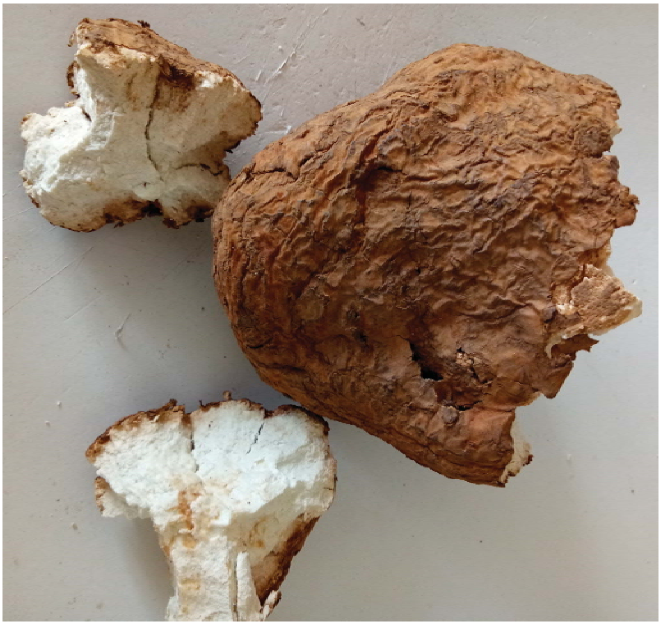


Figure 6. Dried sclerotium of *M. cocos*.

Table 2. The information of *M. cocos* samples.

Class	Location	Abbreviation	Elevation (m)	Latitude (°N)	Longitude (°E)	Parts	Sample Size
1	Beicheng Town, Hongta, Yuxi	BC	1720	24.4319	102.5182	inner part epidermis	10 10
2	Tuodian Town, Shuangbai, Chuxiong	TD	2062	24.6912	101.6493	inner part epidermis	10 10
3	Zhanhe Town, Ninglang, Lijiang	ZH	2560	26.8832	100.9275	inner part epidermis	10 10
4	Dawen Town, Shuangjiang, Lincang	DW	1438	23.3487	100.0047	inner part epidermis	10 10
5	Caodian Town, Yunlong, Dali	CD	2066	25.6360	99.1320	inner part epidermis	10 10
6	Yongping Town, Jinggu, Pu'er	YP	1077	23.4204	100.4044	inner part epidermis	10 10
7	Mengmeng Town, Shuangjiang, Lincang	MM	1052	23.4779	99.8378	inner part epidermis	10 10
8	Baliu Town, Mojiang, Pu'er	BL	1979	23.0676	101.9765	inner part epidermis	8 8

3.3. FTIR Spectra Acquisition

A Fourier transform infrared spectrometer from Perkin Elmer equipped with an attenuated total reflectance (ATR) sampling accessory with a diamond focusing element was employed for FTIR spectroscopy measurement. The sample powder was pressed under a consistent pressure with a pressure tower when collecting spectral. FTIR spectrum of each sample was scanned 16 times successively with a resolution of 4 cm^{-1} in the range of 4000–650 cm^{-1}. After the measurement of one sample was finished, the surface of ATR crystal and the apex of pressure tower were cleaned for the next sample detection. All spectra were background corrected utilizing air spectrum. The laboratory environment was maintained a constant temperature (25 °C) and humidity (30%).

3.4. Chromatographic Analysis

Sample powder was weighed accurately to 0.5 g and extracted with 2.0 mL of methanol by an ultrasound-assisted method for 40 min at ambient temperature. The extract solution was filtered using a 0.22 μm membrane filter. The filtrate was loaded into the auto-sampler vial and stored at 4 °C before injecting into the chromatographic system for analysis.

Analyses of all 156 samples (including Poria and Poriae Cutis) were implemented using a Shimadzu ultra-fast liquid chromatography system equipped with a UV detector, binary gradient pumps, a degasser, an auto sampler and a column oven. The chromatographic separation was achieved using an Inertsil ODS-HL HP column (3.0 × 150 mm, 3 μm particle size) operated at 40 °C. The mobile phase consisted of acetonitrile (A) and 0.05% formic acid (B). Before use, the mobile phase constituents were degassed and filtered through a 0.2 μm filter. The gradient elution sequence was conducted as follows: 0–25 min, 40% A; 25–52 min, 40–69% A; 52–56 min, 69–72% A; 56–58 min, 72–78% A; 58–58.01 min, 78–90% A; and 58.01–60 min, remaining at 90% A (eluting to 65 min for Poriae Cutis samples). Each run was followed by an equilibration period of 3 min with initial conditions (40% A and 60% B). The flow rate was kept at 0.4 $mL \cdot min^{-1}$ and the injection volume was 7 μL. Detective wavelengths were set at 242 nm and 210 nm.

3.5. Method Validation

The developed UFLC method was validated in terms of precision, stability, repeatability, accuracy and linearity under the above chromatographic condition.

A mixed standard solution was determined six times successively within a day and on three consecutive days for evaluating intra- and inter-day precision. For the stability test, the extract of a sample was analyzed at 0, 6, 12, 17, 21 and 24 h, respectively. Six sample solutions prepared individually from the same sample were analyzed in parallel for evaluating the repeatability. The recovery test was performed to evaluate the accuracy by adding reference compounds of three different amounts (low, middle, and high) to the sample with known concentration accurately. The following equation was used to calculate recovery rate: Recovery rate (%) = [(measured amount − original amount)/spiked amount] × 100%.

The standard solutions of five compounds for constructing calibration curves were prepared by diluting the stock solutions with methanol individually. The ranges of concentration in the linearity study were 5.00–999 $\mu g \cdot mL^{-1}$ (dehydrotumulosic acid), 0.22–6730 $\mu g \cdot mL^{-1}$ (poricoic acid A), 2.4–480 $\mu g \cdot mL^{-1}$ (dehydropachymic acid), 10.3–1240 $\mu g \cdot mL^{-1}$ (pachymic acid) and 0.49–2450 $\mu g \cdot mL^{-1}$ (dehydrotrametenolic acid). Due to the obvious difference in contents of poricoic acid A of Poria and Poriae Cutis samples, two concentration ranges of 0.22–1121.95 $\mu g \cdot mL^{-1}$ (Poria) and 0.22–6730 $\mu g \cdot mL^{-1}$ (Poriae Cutis) were prepared. More than seven levels (in arithmetic progression) of every concentration range were guaranteed. The limit of detection (LOD) and limit of quantification (LOQ) were determined by diluting continuously standard solution until the signal-to-noise ratios (S/N) reached about 3 and 10, respectively.

3.6. Preprocessing of Chromatograms and Spectra

The correlation optimized warping algorithm was applied to correct the retention time shifts of chromatogram using MATLAB software (MathWorks, R2017a, Natick, MA, USA). Then the corrected chromatographic data was reduced by taking one in every three points without affecting the chromatographic features to save computation time, which was inspired by the 'data binning' of Lucio-Gutiérrez et al. [22,23]. The first 11 minutes of data that mainly comprised unseparated peaks and baseline shift were discarded.

Raw FTIR spectra were subjected to advanced ATR correction to reduce the impact of skewing of band intensity using OMNIC 9.7.7 software (Thermo Fisher Scientific). Due to the fact that spectra contained hidden and overlapped absorption peaks, second derivative was used for highlighting slight differences employing SIMCA-P$^+$ 13.0 software (Umetrics, Umeå, Sweden). Derivative spectra were calculated with a Savitzky–Golay filter using a second-order polynomial and a 15-point window. The band of 2600–1750 cm^{-1} was associated to diamond crystal in ATR accessory, of which data were excluded prior to chemometrics analysis. These pre-processed data were used to data fusion and PLS-DA.

3.7. Multiple Chromatograms and Spectra Data Fusion

According to the source of data, there were two kinds of data fusion techniques, including the fusion of multiple complementary pieces of information about a single part and the fusion of information about two parts from one sclerotium. For instance, data matrices of LC-Poria and FTIR-Poria could be fused into a new dataset, and data matrices of FTIR-Poria and FTIR-epidermis could be fused into a dataset. It was important to note that information must correspond in the process of data fusion, namely, the LC and FTIR data of the same Poria sample must correspond, or the FTIR data of inner parts and epidermis from the same sclerotium should correspond.

The data fusion could be classified into three levels in light of the combination of data: low level, mid-level and high level. Low and mid-level fusion has been widely used, and was applied to the identification of geographical origin of *M. cocos*. The scheme of low and mid-level data fusion approaches is shown in Figure 7. In the low-level fusion, pre-processed different datasets were straightforward concatenated into a matrix, and the number of variables was equal to the sum of number of original variables. For the mid-level fusion, the scores obtained independently from different data by PCA were concatenated into a dataset applied for provenance traceability, and the number of variables of the dataset was significantly less than that of original variables. Compared with low level, mid-level data fusion could save more time on the operation. Specific types of the data fusion in this study were shown in Table 1.

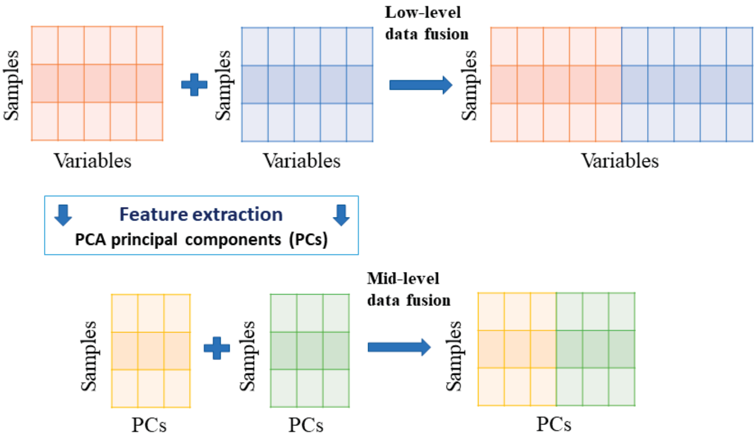


Figure 7. The scheme of the data fusion approaches.

3.8. Evaluation of Model Performance

The calibration and validation sets were selected for assessing the quality of model. The calibration set was used to construct a model that was performed 7-fold cross validation for internal validation, and the validation set was used to externally estimate the practicability of model. To avoid the influence of randomness caused by random sampling, and to obtain a representative calibration set from a pool of samples, the Kennard-Stone algorithm [39] was performed to systematically divide dataset of 78 samples into calibration (52) and validation (26) sets using MATLAB R2017a software (MathWorks).

The performance of discrimination model could be evaluated by sensitivity, specificity and efficiency [40]. The three parameters are dependent on these values: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). TP and TN represent the correctly identified samples in target positive and negative classes, respectively. By analogy, FP and FN represent the incorrectly identified samples in positive and negative classes, respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

$$\text{Efficiency} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{3}$$

Therein, sensitivity shows the ability to correctly recognize samples belonging to the target class while specificity reflects the model ability to reject samples belonging to all other classes. The measure combining the sensitivity and specificity value is called efficiency.

In addition, the accuracy rate of calibration set, the accuracy rate of validation set, R^2(cum) and Q^2(cum) were also employed for assessing the classification performance. Accuracy was obtained by calculating the proportion of correctly classified samples in the total amount of calibration set (or validation set) samples. R^2 is calculated by following equation: $R^2 = 1 - \text{RSS}/\text{SSX}$, where RSS is the residual sum of squares of calculated and measured values, and SSX is the total sum of squares after mean centralization [41]. R^2(cum) represents the percentage of explained variance for a defined number of latent variables, indicating how well the model fits the data. Q^2(cum) represents the cross-validated cumulative R2, suggesting how well the model predicts new data. The higher values of these parameters (close to 1 or 100%), the better performance of model.

4. Conclusions

In order to establish an effective method for geographical authentication of *M. cocos*, two data fusion strategies, including low and mid-level fusion, as well as two data combinations, including the fusion of complementary information regarding a single part and the fusion of information about two parts from one sclerotium were compared. FTIR, LC_{242} and LC_{210} were used to characterize the epidermis and inner part of *M. cocos* sclerotium from different places individually and jointly. The results showed that, chromatographic fingerprint was more suitable than content data of five triterpene acids for origin identification. In the fusion of complementary information about single part, good classification performance was achieved obtained by merging LC_{242} chromatograms and FTIR spectra in low-level fusion way. In the fusion of information about two parts from one sclerotium, the predictive ability of the FTIR low-level fusion model of two parts was the most satisfactory, and all analyzed samples were classified correctly.

In most cases, FTIR proved to be more efficient than LC_{242} and LC_{210}, not only in a single data source but in data fusion. Mid-level data fusion was slightly worse than low-level data fusion. The performance of low-level data fusion models was superior to single technique models. Moreover, Poria samples were more suitable for origin identification than Poriae Cutis samples. On the basis of effective and comprehensive fingerprint information, the low-level data fusion strategy could be used for the discrimination of *M. cocos* samples from different origins with the aid of appropriate mathematical algorithms.

Supplementary Materials: The following are available online.

Author Contributions: Y.-Z.W. and H.-Y.H. designed the project and revised the manuscript. Q.-Q.W. performed the experiments, analyzed the data and wrote the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 31860584 and grant number 31460538.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khan, I.; Huang, G.; Li, X.; Leong, W.; Xia, W.; Hsiao, W.L.W. Mushroom polysaccharides from *Ganoderma lucidum* and *Poria cocos* reveal prebiotic functions. *J. Funct. Foods* **2018**, *41*, 191–201. [CrossRef]
2. Miao, H.; Zhao, Y.; Vaziri, N.D.; Tang, D.; Chen, H.; Chen, H.; Khazaeli, M.; Tarbiat-Boldaji, M.; Hatami, L.; Zhao, Y. Lipidomics biomarkers of diet-induced hyperlipidemia and its treatment with *Poria cocos*. *J. Agric. Food Chem.* **2016**, *64*, 969–979. [CrossRef]
3. Lee, S.; Lee, S.; Roh, H.; Song, S.; Ryoo, R.; Pang, C.; Baek, K.; Kim, K. Cytotoxic constituents from the sclerotia of *Poria cocos* against human lung adenocarcinoma cells by inducing mitochondrial apoptosis. *Cells* **2018**, *7*, 116. [CrossRef]
4. Wu, K.; Fan, J.; Huang, X.; Wu, X.; Guo, C. Hepatoprotective effects exerted by *Poria cocos* polysaccharides against acetaminophen-induced liver injury in mice. *Int. J. Biol. Macromol.* **2018**, *114*, 137–142. [CrossRef] [PubMed]
5. Lee, S.; Choi, E.; Yang, S.; Ryoo, R.; Moon, E.; Kim, S.; Kim, K.H. Bioactive compounds from sclerotia extract of *Poria cocos* that control adipocyte and osteoblast differentiation. *Bioorg. Chem.* **2018**, *81*, 27–34. [CrossRef] [PubMed]
6. Zhu, L.; Xu, J.; Zhang, S.; Wang, R.; Huang, Q.; Chen, H.; Dong, X.; Zhao, Z. Qualitatively and quantitatively comparing secondary metabolites in three medicinal parts derived from *Poria cocos* (Schw.) Wolf using UHPLC-QTOF-MS/MS-based chemical profiling. *J. Pharm. Biomed.* **2018**, *150*, 278–286. [CrossRef]
7. Li, Y.; Zhang, J.; Jin, H.; Liu, H.; Wang, Y. Ultraviolet spectroscopy combined with ultra-fast liquid chromatography and multivariate statistical analysis for quality assessment of wild *Wolfiporia extensa* from different geographical origins. *Spectrochim. Acta Part A* **2016**, *165*, 61–68. [CrossRef]
8. Biancolillo, A.; Marini, F. Chapter four—Chemometrics applied to plant spectral analysis. In *Vibrational Spectroscopy for Plant Varieties and Cultivars Characterization, Comprehensive Analytical Chemistry*, 1st ed.; Lopes, J., Sousa, C., Eds.; Elsevier: Amsterdam, The Netherlands, 2018; Volume 80, pp. 69–104.

9. Yuan, T.; Zhao, Y.; Zhang, J.; Wang, Y. Application of variable selection in the origin discrimination of *Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb. based on near infrared spectroscopy. *Sci. Rep.* **2018**, *8*, 89.
10. Zhu, L.; Xu, J.; Wang, R.; Li, H.; Tan, Y.; Chen, H.; Dong, X.; Zhao, Z. Correlation between quality and geographical origins of *Poria cocos* revealed by qualitative fingerprint profiling and quantitative determination of triterpenoid acids. *Molecules* **2018**, *23*, 2200. [CrossRef]
11. Chen, J.; Sun, S.; Ma, F.; Zhou, Q. Vibrational microspectroscopic identification of powdered traditional medicines: Chemical micromorphology of *Poria* observed by infrared and Raman microspectroscopy. *Spectrochim. Acta Part A* **2014**, *128*, 629–637. [CrossRef]
12. Orlandi, G.; Calvini, R.; Foca, G.; Pigani, L.; Vasile Simone, G.; Ulrici, A. Data fusion of electronic eye and electronic tongue signals to monitor grape ripening. *Talanta* **2019**, *195*, 181–189. [CrossRef]
13. Wu, X.; Zhang, Q.; Wang, Y. Traceability of wild *Paris polyphylla* Smith var. yunnanensis based on data fusion strategy of FT-MIR and UV-Vis combined with SVM and random forest. *Spectrochim. Acta Part A* **2018**, *205*, 479–488. [CrossRef]
14. Ni, Y.; Li, B.; Kokot, S. Discrimination of *Radix Paeoniae* varieties on the basis of their geographical origin by a novel method combining high-performance liquid chromatography and Fourier transform infrared spectroscopy measurements. *Anal. Methods-UK* **2012**, *4*, 4326. [CrossRef]
15. Borràs, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O. Data fusion methodologies for food and beverage authentication and quality assessment—A review. *Anal. Chim. Acta* **2015**, *891*, 1–14. [CrossRef]
16. Casale, M.; Bagnasco, L.; Zotti, M.; Di Piazza, S.; Sitta, N.; Oliveri, P. A NIR spectroscopy-based efficient approach to detect fraudulent additions within mixtures of dried *porcini* mushrooms. *Talanta* **2016**, *160*, 729–734. [CrossRef]
17. Bureau, S.; Cozzolino, D.; Clark, C.J. Contributions of Fourier-transform mid infrared (FT-MIR) spectroscopy to the study of fruit and vegetables: A review. *Postharvest. Biol. Technol.* **2019**, *148*, 1–14. [CrossRef]
18. Li, Y.; Zhang, J.; Wang, Y. FT-MIR and NIR spectral data fusion: A synergetic strategy for the geographical traceability of *Panax notoginseng*. *Anal. Bioanal. Chem.* **2018**, *410*, 91–103. [CrossRef]
19. Wu, Z.; Zhao, Y.; Zhang, J.; Wang, Y. Quality assessment of *Gentiana rigescens* from different geographical origins using FT-IR spectroscopy combined with HPLC. *Molecules* **2017**, *22*, 1238. [CrossRef]
20. Wang, Y.; Shen, T.; Zhang, J.; Huang, H.; Wang, Y. Geographical authentication of *Gentiana rigescens* by high-performance liquid chromatography and infrared spectroscopy. *Anal. Lett.* **2018**, *51*, 2173–2191. [CrossRef]
21. Obisesan, K.A.; Jiménez-Carvelo, A.M.; Cuadros-Rodriguez, L.; Ruisánchez, I.; Callao, M.P. HPLC-UV and HPLC-CAD chromatographic data fusion for the authentication of the geographical origin of palm oil. *Talanta* **2017**, *170*, 413–418. [CrossRef]
22. Lucio-Gutiérrez, J.R.; Garza-Juárez, A.; Coello, J.; Maspoch, S.; Salazar-Cavazos, M.L.; Salazar-Aranda, R.; Waksman De Torres, N. Multi-wavelength high-performance liquid chromatographic fingerprints and chemometrics to predict the antioxidant activity of *Turnera diffusa* as part of its quality control. *J. Chromatogr. A* **2012**, *1235*, 68–76. [CrossRef] [PubMed]
23. Lucio-Gutiérrez, J.R.; Coello, J.; Maspoch, S. Enhanced chromatographic fingerprinting of herb materials by multi-wavelength selection and chemometrics. *Anal. Chim. Acta* **2012**, *710*, 40–49. [CrossRef] [PubMed]
24. Zhang, L.; Liu, Y.; Liu, Z.; Wang, C.; Song, Z.; Liu, Y.; Dong, Y.; Ning, Z.; Lu, A. Comparison of the roots of *Salvia miltiorrhiza* Bunge (Danshen) and its variety *S. miltiorrhiza* Bge f. Alba (Baihua Danshen) based on multi-wavelength HPLC-fingerprinting and contents of nine active components. *Anal. Methods-UK* **2016**, *8*, 3171–3182. [CrossRef]
25. Horn, B.; Esslinger, S.; Pfister, M.; Fauhl-Hassek, C.; Riedl, J. Non-targeted detection of paprika adulteration using mid-infrared spectroscopy and one-class classification–Is it data preprocessing that makes the performance? *Food Chem.* **2018**, *257*, 112–119. [CrossRef]
26. Cael, J.J.; Koenig, J.L.; Blackwell, J. Infrared and Raman spectroscopy of carbohydrates. Part VI: Normal coordinate analysis of V-amylose. *Biopolymers* **1975**, *14*, 1885–1903. [CrossRef]
27. Li, S.; Wang, L.; Song, C.; Hu, X.; Sun, H.; Yang, Y.; Lei, Z.; Zhang, Z. Utilization of soybean curd residue for polysaccharides by *Wolfiporia extensa* (Peck) Ginns and the antioxidant activities in vitro. *J. Taiwan Inst. Chem. E* **2014**, *45*, 6–11. [CrossRef]

28. Akihisa, T.; Uchiyama, E.; Kikuchi, T.; Tokuda, H.; Suzuki, T.; Kimura, Y. Anti-tumor-promoting effects of 25-methoxyporicoic acid A and other triterpene acids from *Poria cocos*. *J. Nat. Prod.* **2009**, *72*, 1786–1792. [CrossRef]
29. Lee, S.; Lee, D.; Lee, S.O.; Ryu, J.; Choi, S.; Kang, K.S.; Kim, K.H. Anti-inflammatory activity of the sclerotia of edible fungus, *Poria cocos* Wolf and their active lanostane triterpenoids. *J. Funct. Foods* **2017**, *32*, 27–36. [CrossRef]
30. Ying, Y.; Shan, W.; Zhang, L.; Zhan, Z. Lanostane triterpenes from *Ceriporia lacerate* HS-ZJUT-C13A, a fungal endophyte of *Huperzia serrata*. *Helv. Chim. Acta* **2013**, *95*, 2092–2097. [CrossRef]
31. Maquelin, K.; Kirschner, C.; Choo-Smith, L.P.; van den Braak, N.; Endtz, H.P.; Naumann, D.; Puppels, G.J. Identification of medically relevant microorganisms by vibrational spectroscopy. *J. Microbiol. Methods* **2002**, *51*, 255–271. [CrossRef]
32. Skov, T.; van den Berg, F.; Tomasi, G.; Bro, R. Automated alignment of chromatographic data. *J. Chemom.* **2006**, *20*, 484–497. [CrossRef]
33. Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal. Methods-UK* **2013**, *5*, 3790. [CrossRef]
34. Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196. [CrossRef]
35. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. [CrossRef]
36. Sjöström, M.; Wold, S.; Söderström, B. PLS discriminant plots. In *Pattern Recognition in Practice*; Gelsema, E.S., Kanal, L.N., Eds.; Elsevier: Amsterdam, The Netherlands, 1986; pp. 461–470.
37. Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial least squares projections to latent structures. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; KLUWER ESCOM Science Publisher: Leiden, The Netherlands, 1993; pp. 523–550.
38. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab.* **1987**, *2*, 37–52. [CrossRef]
39. Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148. [CrossRef]
40. Oliveri, P.; Downey, G. Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends Anal. Chem.* **2012**, *35*, 74–86. [CrossRef]
41. Aa, J.Y. Analysis of metabolomic data: Principal component analysis. *Chin. J. Clin. Pharmacol. Ther.* **2010**, *15*, 481–489.

Sample Availability: Samples of the compounds are not available from the authors.

molecules

MDPI

Review

Approaching Authenticity Issues in Fish and Seafood Products by Qualitative Spectroscopy and Chemometrics

Sergio Ghidini, Maria Olga Varrà * and Emanuela Zanardi

Department of Food and Drug, University of Parma, Strada del Taglio 10, 43126 Parma, Italy; sergio.ghidini@unipr.it (S.G.); emanuela.zanardi@unipr.it (E.Z.)
* Correspondence: mariaolga.varra@studenti.unipr.it; Tel.: +39-0521-902-761

Received: 9 April 2019; Accepted: 8 May 2019; Published: 10 May 2019

Abstract: The intrinsically complex nature of fish and seafood, as well as the complicated organisation of the international fish supply and market, make struggle against counterfeiting and falsification of fish and seafood products very difficult. The development of fast and reliable omics strategies based on spectroscopy in conjunction with multivariate data analysis has been attracting great interest from food scientists, so that the studies linked to fish and seafood authenticity have increased considerably in recent years. The present work has been designed to review the most promising studies dealing with the use of qualitative spectroscopy and chemometrics for the resolution of the key authenticity issues of fish and seafood products, with a focus on species substitution, geographical origin falsification, production method or farming system misrepresentation, and fresh for frozen/thawed product substitution. Within this framework, the potential of fluorescence, vibrational, nuclear magnetic resonance, and hyperspectral imaging spectroscopies, combined with both unsupervised and supervised chemometric techniques, has been highlighted, each time pointing out the trends in using one or another analytical approach and the performances achieved.

Keywords: fish and seafood; food authentication; chemometrics; fingerprinting; wild and farmed; geographical origin; vibrational spectroscopy; absorption/fluorescence spectroscopy; nuclear magnetic resonance; hyperspectral imaging

1. Introduction

The demand for fish and seafood products has increased notably during the last years, mostly as a consequence of the new special attention paid by consumers towards healthier food. The technological development that has invested the whole fisheries sector has additionally contributed to overcome the well-known obstacles to export fish and seafood worldwide, deriving from the high vulnerability of the products, to the point that today more than 35% of all caught and cultured fish is traded across national boundaries [1]. The growing competitiveness of the sector and diversification in fish supply chain have, in turn, led to the presence of a huge variety of look-alike products on the international market, whose global quality features are, however, quite different. More than 700 different species of fish, 100 of molluscan, and 100 of crustacean are, in fact, used as food for humans [2].

In this scenario, what is remarkable is that consumers demand not only for more fish, but for even safer and higher-quality fish, whilst the deliberate or accidental lack of transparency about the identity of products and fraudulent or negligent activities continue to grow. Based on what has been recently reported by the Food and Agriculture Organization, fish and related products have become among the most vulnerable to fraud category of food. Nevertheless, the effective monitoring of illicit practices in the fisheries sector is hampered by the increasing spread of highly processed fish products, in which the presence of different types of fraud can be hidden with ease [3].

The voluntary substitution of commercially valuable fish species with lower quality ones, represents the most recurrent form of fish fraud, although substitution can also take place accidentally when species look so similar that they are mistaken for each other. The geographical provenance and the production process are other current authenticity topics concerning fish and seafood products, whose falsification which is hard to bring to light, has a negative economic impact. Despite being economically motivated, mislabelling concerning these issues may occasionally represent a risk to public health. The illegal commercialisation of poisonous fish species (*Tetraodontidae, Molidae, Diodontidae*, and *Canthigasteridae* families) or the replacement of certain kinds of raw fish fillets with gastro-intestinal toxic fish (i.e., those belonging to the *Gempylidae* family) are just some of many examples. Likewise, occurrence of some harmful marine biotoxins may be linked to the geographical distribution of the producing organisms [4], while the presence of higher levels of heavy metals or residues of antibiotic and pesticides are more likely to be found in farmed products than in wild ones [5–7].

Ensuring a clear discrimination of the authenticity of fish and seafood is of special concern today not only for consumers, but also for producers, traders, and industries. Traceability throughout the whole production chain and at all stages of the market, covered by Regulations 178/2002/EC [8], 1005/2008/EC [9], and 1224/2009/EC [10], is considered to be the starting point for the assurance of a high level of safety and quality of food and ingredients, as it represents the basic instrument not only for preventing illegal activities, but also for protecting consumers through the opportunity to access information about the exact nature and characteristics of fish. Specific regulations for the provision of information to consumers [11], and the requirement to uniquely identify fish and seafood on the label [12], play also an essential role in providing more transparency regarding the nature of the products, as they allow consumers to make informed choices and further contribute to the implementation of seafood traceability. As a matter of fact, labels of all unprocessed and some processed fishery and aquaculture products must include information on both the commercial and scientific names of the species, whether the fish has been caught or farmed, the catch or the production area, the fishing gear used, whether the product has been defrosted, and the date of minimum durability (where appropriate). Many other voluntary claims can also be reported on the label, including the date of catch/harvest for wild/aquaculture products, information about the production techniques and practices, and environmental and ethical information [12].

All the claimed declarations appearing on the label must always be checked to verify whether they are truthful. Therefore, in spite of the utility of the traceability system, the fisheries sector needs effective methods to address the problem of fish authenticity and ensure product quality. Innovative analytical approaches based on the evaluation of total spectral properties, are rapidly gaining ground at all levels of current food authenticity research, thanks to their ability to simultaneously provide lots of information related to physical and chemical characteristics of the food matrix. Recent advances in chemometrics, moreover, have represented a major turning point in the dissemination of 'fingerprinting strategies', as they allow for the study of all the genetic, environmental, and other external factors influencing food identity, and to bypass many obstacles related to the application of conventional techniques [13]. This way, chemometrics can be now considered an essential tool for differentiation of similar samples according to the authentication issues of interest.

Until now, several spectroscopic techniques in conjunction with chemometrics have been used as rapid, simple, and cheap tools for fish quality and authenticity testing. Among these, vibrational (near-infrared (NIR), mid-infrared (MIR), Raman), fluorescence or absorption ultraviolet-visible (UV–Vis), and nuclear magnetic resonance (NMR) spectroscopies, together with hyperspectral imaging (HSI) spectroscopy, represent the most used techniques, even if they are still being developed.

Based on this background, the present review article has been designed to highlight the uses and developments of fast and reliable omics strategies based on UV–Vis, NIR, MIR, Raman, NMR, and HSI spectroscopies, with the attempt to address the key authenticity challenges within the fish and seafood sector. To this end, a brief discussion concerning basilar concepts underlying these techniques has been provided, and has been accompanied by a short overview about the implementation of several chemometric tools, in order to highlight the potential benefits in extracting relevant information from spectral data.

The main body of this review focuses specifically on the application, over the years, of spectroscopy and chemometrics to distinguish products in accordance with the species, production method (wild or farmed), farming system (conventional or organic; intensive, semi-intensive, or extensive), geographical provenance (different FAO areas and countries of origin), and the processing technique (fresh or fresh/thawed) that at present, correspond to the key authenticity concerns for which there must be ongoing and effective monitoring.

2. A Conceptual Framework of Spectroscopy and Chemometrics

Spectroscopy is the study of electromagnetic radiation interacting with matter, which can be absorbed, transmitted, or scattered on the basis of both the specific frequency of the radiation and the physical/chemical nature of the matter. When absorbed, radiation leads to a change in the energy states of atoms, nuclei, molecules, or crystals that make up matter, inducing an electronic, vibrational, or rotational transition, depending on the energy of the incident radiation [14]. When the radiation, at a specific frequency, is scattered by molecules (as in Raman spectroscopy), some changes can occur in the energy of the incident photon, which transfers parts of its energy to the matter. In any case, the result of these interactions is a spectrum enclosing many features of the matter analysed, which, when properly interpreted with the help of chemometrics, can be used in a great number of different applications. In choosing the most appropriate spectroscopic method to be used, consideration should be given to some factors, which go beyond the purely analytical purposes: the physical state and chemical composition of the sample, sensitivity, specificity, and overall accuracy of the technique, scale of operation, time of analysis, and cost/availability of the instrumentation [15].

For the sake of conciseness, the main features related to spectroscopic techniques used mostly in the food authentication field are summarised in Table 1.

Table 1. Comparison of different spectroscopic techniques used for food authentication purposes: summary of the main characteristics.

Spectroscopic Technique		Wavelength Range (nm)	Interaction Light-matter	Basic Principle	Sensitive Compounds	Information Obtained	Applications	Possible Limitations
UV–Vis	UV Vis	2×10^2–4×10^2 4×10^2–7.5×10^2	Absorption/emission	Electronic transitions	Double-conjugated bonds; isolated double, triple, peptide bonds; aromatic and carbonyl groups	Molecular structure	Qualitative/ quantitative	Need of sample preparation pH and temperature interferences
IR[1]:	NIR MIR	7.5×10^2–2.5×10^3 2.5×10^3–2.5×10^4	Absorption	Vibrations/rotations of molecular bonds (changes in dipole moments)	Polar bonds (N–H, C–H, O–H, S–H, C–O)	Chemical bonds and physical structure	Qualitative/ quantitative	Water interferences Overlapping of spectral peaks
Raman		2.5×10^3–1.0×10^6	Scattering	Vibrations of molecular bonds (changes in polarizability)	Non-polar double or triple bonds (C = C, C ≡ C)	Chemical bonds and physical structure	Qualitative/ quantitative	Fluorescence and photodecomposition interferences Low-intensity Peaks
HSI		Varying by spectroscopic modules	Absorption/emission/scattering	Varying by vibrational spectroscopic modules	Varying by vibrational spectroscopic modules	Varying by vibrational spectroscopic modules	Qualitative/ quantitative/ spatial	Varying by vibrational spectroscopic modules
NMR		5.0×10^8–7.5×10^9	Absorption	Nuclear spin changes	Nuclei having a proper magnetic field (spin quantum number ≠ 0 [2]	Regio/stereo chemistry of molecules	Qualitative/ quantitative/ structural	Cost of the equipment

[1] Infrared (IR) electromagnetic regions taken into consideration do not include far-infrared (FIR) range (2.5×10^4–1.0×10^5 nm) since it is not commonly used in food authentication studies. [2] H-1, C-13, and P-31 are the most frequently investigated nuclei in food science-related nuclear magnetic resonance (NMR) applications.

2.1. UV–Vis Absorption and Fluorescence Emission Spectroscopy

UV–Vis spectroscopy involves the electronic excitation of molecules containing specific chromophore groups, which results from the absorption of photons at two wavelength regions of the electromagnetic spectrum. In the absorption mode, the amount of light retained by the sample is measured, while in the fluorescence mode the amount of light emitted after absorption is taken into consideration [15]. Typically, the UV–Vis spectrum is characterised by broad absorption or emission peaks which reflect the molecular composition of the matrix: by exploiting the unicity absorption or emission patterns of the entire spectrum, or by measuring the absorbance or fluorescence intensity of the analyte at one wavelength, this spectrum can be used for many food analytical qualitative and quantitative applications, respectively [16,17].

2.2. IR Spectroscopy

Infrared spectroscopy involves three different sub-regions of the electromagnetic spectrum, namely NIR, MIR, and FIR, whose absorption by samples results in vibrations of atoms in molecular bonds [18]. These vibrations give out a great amount of information related not only to chemical bonding, but also to the general molecular conformation, structure, and intermolecular interactions within the sample [19]. This way, IR spectra enclose the total sample composition, whose pattern of peaks distribution represents a unique signature profile and whose intensity of bands is linked to the concentration of specific compounds [20,21].

The NIR spectrum of food samples results from absorption by molecular bonds containing prevalently light atoms and it is characterised by the presence of broad and overlapping overtone and combination bands [22,23]. By contrast, spectral signature in the MIR region is characterised by the presence of more intense and delineated bands, whose position and intensity are more informative of molecule's concentration in the sample [24,25]. Here too, the spectral profile is complex and data mining is very difficult without the use of multivariate data analysis. Finally, with reference to FIR spectroscopy, it is noted that no applications to food authentication are currently available since it relates to molecules containing halogen atoms, organometallic compounds, and inorganic compounds, whose interest is more limited within the context of food research [26].

2.3. Raman Spectroscopy

Raman spectroscopy is a molecular vibration technique based on the inelastic Raman scattering, a physical effect that comes with molecular vibrations and triggers a change in the polarizability of the molecule [27]. In particular, this kind of spectroscopy focuses on the measurement of those small fractions of the radiation which is scattered by specific categories of compounds at higher or lower frequencies than incident photons. The typical Raman spectrum, showing intensities of the scattered light versus the wavelengths of the Raman shift, is characterised by sharp and well-resolved bands, which provide information about molecular structure and composition of the matter analysed.

For a long time after its discovery, Raman spectroscopy has been poorly exploited in food applications, by reason of several analytical disadvantages and interference (see Table 1). These drawbacks have now been overcome thanks to the overall technological improvement of Raman equipment: by way of example, surface-enhanced Raman spectroscopy (SERS) has recently made it possible to surmount hurdles related to faint scattering signals [28].

2.4. Hyperspectral Imaging

HSI is a technique cobbling together spectroscopy and computer vision to give useful information concerning the physicochemical characteristics of samples in relation to their specific spatial distribution. Briefly, HSI systems provide several hyperspectral images of the tested sample, corresponding to three-dimensional data containers, of which each sub-image is a map showing spatial distribution of the sample constituents in relation to each single wavelength [29,30].

Over the recent years, the steady usage growth of HIS technology in the field of food research has been mainly driven by the availability of different instrumental configurations that exploit fluorescence, absorbance, or light scattering phenomena. On the other side, application of spectral imaging technologies is not at all widespread in the food industry, due to a variety of factors ranging from high costs and low availability of instrumentations, to the computation speed and necessity of expertise by users [31].

2.5. NMR Spectroscopy

NMR spectroscopy is a very versatile technique for food analysis and its untargeted applications have become very popular. The first reason for NMR popularity is that the composition of the matter under study can be perfectly mapped out by the overall NMR spectral profiles, thus giving a comprehensive view for the identification of all major and minor food components [32]. At the same time, the area of the NMR spectral bands is directly proportional to the number of nuclei that produce the signal, so the technique is also well-suited for quantitative purposes. Additionally, despite relatively high NMR equipment costs and spectra interpretation difficulties, NMR spectroscopy is one of the only techniques available that can provide information about the regio/stereo chemistry of molecules [33].

On the basis of the physical state of the matter and on the intended aim of NMR application, different methodologies involving the use of NMR have been optimized. Among these, high-resolution NMR, low-field NMR, solid-state NMR, liquid-state NMR, and NMR imaging are the most used ones, any of which requires specific instrumentation and different approaches to sample preparation, data acquisition, and processing [34].

2.6. Qualitative Chemometric Methods

Raw spectra resulting from spectroscopic analyses are usually characterised by broad and unresolved bands containing too much information, some of which are certainly useful and need to be retained, but some of which hamper the correct data interpretation and need to be removed. Recent advances in chemometrics have marked an important milestone in spectra analysis, since they have simplified the identification of hidden interrelations between variables providing the key for discrimination and classification of samples [20,35]. In other words, qualitative chemometrics methods help to recognise similarities and dissimilarities within spectral data, which can be used to confirm the authenticity or detect adulteration of food samples [36].

Based on the explorative or predictive nature of the methodology, qualitative chemometric techniques are usually classified into unsupervised and supervised techniques. While unsupervised techniques are independent of prior knowledge of class membership of samples to perform classification, supervised techniques call for such knowledge. Brief descriptions of the principles behind the chemometric techniques which are being used to a greater extent are provided below.

2.6.1. Spectral Pre-Treatments

Pre-treatment of spectral data is recognized as being fully integrated into the chemometric set-up itself. Prior to the development of chemometric models, raw spectroscopic data are suggested to be pre-processed by applying some corrections, aimed to enhance spectral properties and minimize the fraction of systematic variation which does not contain relevant information to the discrimination of samples. One such systematic variation is the sum of different physical effects which arise during instrumental acquisition of spectra (e.g., light scattering or background fluorescence phenomena), which are responsible for the appearance, especially in solids samples, of multiplicative, additive, and non-linearity effects (e.g., overlapping bands, baseline shifts/drifts, random noise) [37].

Thus, pre-processing algorithms are usually classified into signal correction methods (e.g., multiplicative scatter correction, MSC; standard normal variate, SNV), differentiation methods (first, second, or third order derivation), and filtering-based methods (e.g., orthogonal signal correction, OSC;

orthogonal wavelet correction, OWAVEC) [38]. While signal correction and filtering-based methods are conceived to retain only the spectral information mainly by suppressing the light-scattering effects, derivative-based methods also help to reduce the spectral complexity through the separation of the broad overlapping bands.

A more detailed description of spectral pre-processing techniques can be widely found in the literature [37,39,40]. Either way, it is essential to point out that spectral filters are most often concatenated to exploit the effects of each one, but this concatenation might increase model complexity and background noise, resulting in an inaccurate chemometric modelling of data and, thus, wrong predictions. For this reason, it is recommended to customize the selection of the pre-treatments prior to performing chemometric analysis according to the spectroscopic technique used and the sample characteristics, trying to restrict, whenever possible, their number.

2.6.2. Unsupervised Methods

Unsupervised methods look at the study of variability among samples for the purpose of identifying their natural characteristics and possible similarities among them, without the need to provide any information about the class to which samples belong.

Between the various available techniques, principal component analysis (PCA) is the most used one. PCA is a quite basic projection method able to reduce the original correlated variables into a smaller number of new uncorrelated latent variables (known as principal components), containing as much systematic variation as possible of the original data [41]. Score plot outputs deriving from PCA applications show in a simple and intuitive graphical way the hidden structures among samples, the interrelations among variables and between samples and variables, the probable presence of any outliers, and possible groupings or dispersion of sample according to specific class membership.

Hierarchical cluster analysis (HCA) is another frequently employed unsupervised method, based on the splitting of samples into different clusters. This splitting is based on the degree of analogy among samples and it is generally performed by evaluating the Mahalanobis or Euclidean distance between the same samples. The hierarchical approach followed is thus aimed at constructing a ladder, in which the most closely related samples are first classified into small groups, and then progressively assembled into bigger groups including less similar samples [35]. Results of HCA are graphically expressed by tree diagrams (dendrograms) showing relationships among clusters; nevertheless, despite being easily computable, dendrograms are often misunderstood, since the number of clusters to be considered is arbitrary, making the interpretation of results more subjective than objective.

2.6.3. Supervised Methods

Supervised techniques require the previous knowledge of the class membership of the samples tested, which can be used to develop predictive models able to discriminate and classify future unidentified samples. There are several different chemometric techniques belonging to the category of the supervised methods, most of which require a training set (to find classification rules for the sample), and a test set (to assess the predictability of the model developed) [42].

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are variance-based methods which use Euclidean distance to find those combinations of the original variables determining maximum separation among the different groups of samples [20]. Both techniques presume that the measurements within each class are normally distributed, but while LDA supposes that dispersion (covariance) is identical for all the classes, QDA, on the contrary, allows the possibility of different dispersion to be present within different classes [35]. Although QDA is considered an extension of LDA, there are some common limitations, for instance the risks of overfitting and failing in classification, especially when the samples size for each class in unbalanced.

K-nearest neighbors (k-NN) clustering is one of the simplest method to discriminate samples on the basis of the distance among them. After choosing the adequate number of k-neighbor samples, the algorithm identifies the k-nearest samples of known class membership to select the classification of unknown samples. This method, unlike LDA and QDA, does not require any prior assumption and its success is independent of the homogeneity of sample numbers in each tested class [43].

Among supervised machine learning approaches, support vector machines (SVM) are particularly advantageous when samples classification is complicated by non-linearity and high dimensional space. The core of the method is the use of specific functions for pattern analysis (kernel algorithms), through which the margin of separation between classes is maximised and complex classification problems that are not linear in the initial dimension (but may be at high dimensional spaces) are resolved [20].

Similarly, artificial neural networks (ANN) is a machine learning method characterised by the ability to adapt to the data, providing classification also in the presence of non-linearity input–output relationships. Structured and organized in a less complex way than SVM, ANN usually generate a more rapid response at a lower computational cost; these efforts, however, are counterbalanced by a reduction in accuracy [20,44]. Nevertheless, ANN suffers from poor data generalisation and, by consequence, it is inclined to return model's overfitting errors. This tendency to overfitting is the main reason why accurate ANN computation analyses call for a very high number of samples to be considered, and at the same time, require strict internal and external validations to be performed, where the training set and the test set should enclose as much similar variability as possible [45].

Soft independent modelling of class analogy (SIMCA) is an alternative pattern recognition method which first performs individual PCA on the samples for each class they must be assigned to, in order to compress original variables into a smaller number of new principal components. Principal components and critical distances computed are then used to delineate a confidence limit for each class. Unknown samples are then assigned to the class to which they get close by projection into the resulting multidimensional space [36]. SIMCA is particularly useful when samples belong to several different classes; since maximum class-separation is not covered by the method, the interpretation of the outcomes may be difficult, if not impossible [20].

Regression-based supervised discriminant analyses exploit specific classification algorithms to model the interrelations existing between measured variables (i.e., spectra) and qualitative parameters (i.e., class membership), such that maximum separation between the different groups of samples is achieved. Partial least square-discriminant analysis (PLS-DA) and orthogonal partial least square-discriminant analysis (OPLS-DA) belong to this category of techniques. PLS-DA involves a standard PLS regression to find interrelations between the X-matrix (containing measured variables) and Y-matrix (containing categorical variables) by building new variables (latent variables). These interrelations allow not only to classify new samples into one of the Y-groups based on measured spectrum, but also to identify variables that mostly contribute to the classification. Although PLS-DA has the advantage of modelling noisy and highly collinear data efficiently, the technique is often unsuccessful when the non-related (orthogonal) variability in the X-matrix is substantial, since it hinders the correct interpretation of the results [20]. This drawback can be overcome by the application of OPLS-DA, through which the orthogonal variability within the X-matrix is separated from the related (predicted) variability and then modelled apart. Consequently, if samples cannot be discriminated along the predictive direction, the orthogonal variability may be handled to increase the effectiveness of discrimination among classes [46].

3. Authenticating Fish and Seafood through the Application of Qualitative Spectroscopy and Chemometrics

Spectroscopic and chemometric analyses have been used over the years for many applications in fishery research, those in the authentication field being among the most promising ones. Some of the works concerning the flexibility of spectroscopy in fish and seafood analysis have already been reviewed by different authors [24,25,47–49], but they have mainly centred on illustration of the advances of the available techniques for quality attributes assessment, as well as on the advantages and limitations of the single type of technique over traditional methods.

Therefore, in the following section, more attention has been paid to the resolution, on a case-by-case basis, of the weightiest authentication issues in the fish and seafood sector, namely species substitution, geographical origin falsification, production method or farming system misrepresentation, and fresh for frozen/thawed product substitution, each time pointing out the trends in using one or another method as well as the discrimination performances achieved, which are considered to be the most intuitive parameters used for chemometric models diagnostics. An overview of the most frequently investigated authentication issues in the fishery sector and the trend of using each spectroscopic technique over the years by the scientific community are plotted in Figures 1 and 2, respectively.

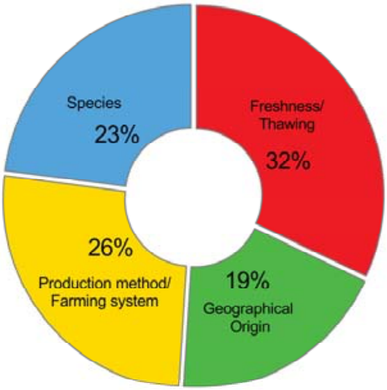


Figure 1. Percentage distribution of the authenticity issues covered by the scientific literature reviewed in the present work. Data were collected in February 2019 from the web search engine Google Scholar (search criteria: time period: "any time", and keywords: "fish and/or seafood"; "authenticity"; "spectroscopy"; "chemometrics".

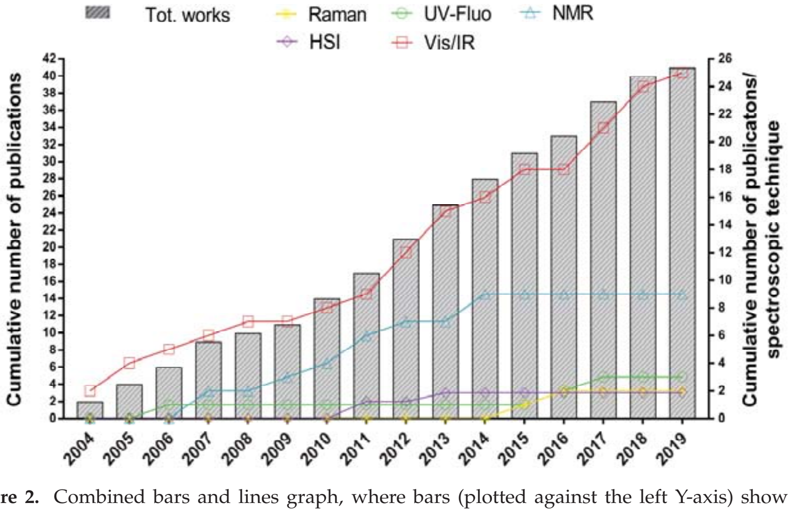


Figure 2. Combined bars and lines graph, where bars (plotted against the left Y-axis) show the cumulative number of scientific works concerning the use of spectroscopy and chemometrics for fish authentication purposes, and lines (plotted against the right Y-axis) show the cumulative number of works using each spectroscopic technique. Data were collected in February 2019 from the web search engine Google Scholar (search criteria: time period: "any time", and keywords: "fish and/or seafood"; "authenticity"; "spectroscopy"; "chemometrics".

3.1. Species Substitution

Substitution or counterfeit of high-value fish species with low-value ones has many quality and safety implications. Therefore, the confirmation of scientific and commercial names declared on the label through the use of rapid and low-cost methods is increasingly popular in food research.

3.1.1. Application of Vibrational Spectroscopy

An early study explored Vis-NIR spectroscopy as a tool to detect the counterfeit of Atlantic blue crabmeat (*Callinectes sapidus*) with blue swimmer crabmeat (*Portunus pelagicus*) in 10% increments, taking into consideration their different commercial values [50]. Qualitative chemometric analysis was performed on 400–2498 nm Vis-NIR spectra (previously subjected to different pretreatments to evaluate the effects on model performance), by means of a full-spectrum PCA and a sequential-spectrum PCA. As a result, both the first derivative-pretreated full spectra and second derivative-pretreated sequential spectra, highlighted a trend of samples towards moving from the left part to the right part of the PCA score plot with increased adulteration levels, but authors identified the sequential approach, using 400–1700 nm second derivative spectra, as being the most informative and, thus, the most suitable approach [50].

Based on the fact that the past several years have seen a sharp rise in the interest towards the portability of instruments, which may provide greater flexibility especially in on-line, in-line, and at-line routine quality control, a study performed by O'Brien et al. (2013), explored the ability of a hand-held NIR spectrometer to give positive results of discrimination between high-value and low-value whole fish and fish fillet species [51]. In particular, the objective was to discriminate between two different species of mullet (red mullet from mullet), cod (winter cod from cod), and trout (samlet from salmon trout). NIR spectra (906–1648 nm) obtained from skin (whole fish) and meat (fish fillets), were first pre-processed and then elaborated by PCA and SIMCA analysis. Successful PCA results were achieved only in separating the whole mullet samples, but the discrimination performances improved significantly also for mullet fillets after the application of the SIMCA analysis. PCA failed to discriminate both whole cod and cod fillets, but here too, SIMCA predictions provided a correct assignment of the tested fish samples. Similar outcomes for samlet from salmon trout were achieved [51]. Thus, although PCA investigation failed, SIMCA supervised analysis clearly outlined the possibility to authenticate high quality fish species which are potentially substitutable with lower-quality alternatives. Still in the context of the use of hand-held and compact NIR devices, a broader attempt to distinguish fillets and patties of Atlantic cod (*Gadus morhua*) from those of haddock (*Melanogrammus aeglefinus*) was recently made [52]. Raw fillets and patties of the two fish species were scanned at 950–1650 nm (by the portable instrument) or at 800–2222 nm (by a benchtop instrument) and after being pre-treated with SNV, MSC, or Savitzky–Golay smoothing (SG) coupled with first or second derivative, they were elaborated by means of supervised LDA and SIMCA analysis. Regardless of instrumentation used, the best LDA models were computed on the MSC spectra of both fillets and patties, since the correct classification rate in the external validation step reached 100% [52]. SIMCA class-modelling strategy obtained 100% correctly classified SNV, SG-first derivative, or SG-second derivative fillets spectra acquired by benchtop NIR, and 100% correctly classified MSC fillets spectra acquired with a portable NIR [52]. As for patties, samples acquired by benchtop NIR and portable NIR were 100% correctly classified when spectra were subjected to SG-first derivative or SG-second derivative, and SNV or MSC, respectively. The worst SIMCA outcomes in prediction for patties and fillets were obtained for SG-second derivative spectra acquired with the portable instrument. Despite these results, no significant differences in the performances of the two instruments tested were found, thus confirming equivalent discrimination powers also in processed product.

Different species of freshwater fish of the Cyprinidae family, namely black carp (*Mylopharyngodon piceus*), grass carp (*Ctenopharyngodon idellus*), silver carp (*Hypophthalmichthys molitrix*), bighead carp (*Aristichthys nobilis*), common carp (*Cyprinus carpio*), crucian (*Carassius auratus*), and bream (*Parabramis pekinensis*), were also investigated by NIR spectroscopy [53]. Fish samples were scanned in the

1000–1799 nm region, MSC pre-treated, and pre-reduced in dimensionality by different methods, including PCA, PLS, and fast Fourier transform (FFT). In this case, LDA models were built by using only nine pre-selected spectra wavelengths from the entire spectrum and results obtained showed a good prediction ability of the adopted strategy: PCA-LDA and FFT-LDA models, in fact, showed 100% accuracy, specificity, sensitivity, and precision, even if most of the information was not taken into account by calculation [53].

Zhang et al. (2017) attempted to classify marine fish surimi by 1100–2500 nm NIR spectroscopy, according to the species by which products were composed, namely white croaker (*Argyrosomus argentatus*), hairtail (*Trichiurus haumela*), and red coat (*Nemipterus virgatus*) [54]. According to results obtained from PCA of the pre-processed spectra, the presence of a well-defined and separated cluster associated with red coat surimi species was observed, but the separation of the other two species of surimi samples was not clear [54]. However, as regards LDA results, 100% correct classification rate for external validation datasets after MSC pre-treatment was achieved, demonstrating once again the greater effectiveness of supervised analyses compared to unsupervised ones.

Species authenticity was also studied by comparing FT-NIR and FT-MIR spectra of red mullet and plaice fillets (higher-value species) to those of Atlantic mullet and flounder fillets (lower-value species) [55]. LDA and SIMCA analysis applied to differently pre-treated NIR and MIR spectra (800–2500 nm and 2500–14,300 nm spectral ranges, respectively), clearly discriminated Atlantic mullet fillets from those of the more valuable red mullet. While LDA gave a 100% correct classification percentage in prediction (irrespective of the spectroscopic technique considered), sensitivity and specificity higher than 70% and 100%, respectively, were calculated for FT-NIR spectra subjected to SIMCA analysis [55]. Poorer, but acceptable, results were obtained for flounder and plaice fillets discrimination: in this case, FT-IR spectroscopy showed the best discrimination power, with a prediction ability higher than 83% and a specificity of 100%.

The usefulness of NIR spectroscopy was explored to identify different fish species used to make fishmeal under industrial conditions. The 1100–2500 nm raw or second derivative NIR spectra of samples containing salmon, blue whiting, and other (i.e., mackerel or herring) fish species were elaborated by PCA, LDA, and DPLS (PLS-DA). Models developed correctly classify, on average, more than 80% of the fish meal samples into the three groups assigned according to the fish species [56].

In contrast to the multiple applications of NIR spectroscopy, only one study explored the discrimination abilities of MIR spectroscopy [57]. This study coupled SG- and SNV-pre-treated MIR spectra (2500–20,000 nm) with chemometrics (PCA) to specifically detect adulteration of Atlantic salmon (*Salmo salar*) mini-burgers with different percentage (from 0 to 100%, in steps of 10%) of Rainbow trout (*Onconrhynchus mykiss*). The resulting 11 formulations of salmon burgers were grouped into 11 distinct clusters, even when the samples were stored for different periods of time before acquisition [57].

Only two applications of Raman spectroscopy concerning fish species authentication are available. The aim of the first study was to discriminate 12 different fish fillets of different species by using pre-treated Raman spectra in the range 300–3400 cm^{-1} (about 3940–33,333 nm) recorded by a Raman spectrometer equipped with a 532 nm laser exciting source [58]. HCA analysis applied to the Raman spectra revealed the presence of three major clusters, one corresponding to fish from the Salmonidae family (rainbow trout and Chum salmon), one corresponding to various freshwater fish (zander, Nile perch, pangasius, and European seabass), and one corresponding to various saltwater fish (Atlantic herring, Atlantic pollock, Alaska pollock, Atlantic cod, blue grenadier, and yellowfin tuna). Within these large clusters, spectra were also grouped according to their species in sub-clusters, with a high degree of accuracy of the spectral classification on species level (95.8%) [58]. Similarly, PCA analysis performed on 5000–50,000 nm Raman spectra (acquired by using a 785 nm laser exciting source) discriminated among horse mackerel (*Trachurus trachurus*), European anchovy (*Engraulis encrasicolus*), Bluefish (*Pomatamus saltatrix*), Atlantic salmon (*Salmo salar*), and flying gurnard (*Trigla lucerna*) samples. In this case, however, the study was less rapid and more elaborate since the spectral acquisition was performed on the previously extracted lipid fraction of fish [59].

3.1.2. Application of NMR Spectroscopy

Muscle lipids of four different species of fish belonging to the Gadoid family, namely cod (*Gadus morhua*), haddock (*Melanogrammus aeglifinus*), saithe (*Pollachius virens*), and pollack (*P. pollachius*), were subjected to ^{13}C-NMR spectroscopic analysis of phospholipid profiles, in order to authenticate samples according to the species [60]. As a result, supervised LDA and Bayesian belief network (BBN) performed on the resulting ^{13}C-NMR spectral peaks provided 78% and 100% of the correctly classified samples, respectively [60]. Other applications of NMR and chemometrics concerning fish species discrimination were not reported in literature until now. In our opinion, the method should be further explored in view of the several potentials and benefits provided, despite disadvantages deriving from the need of sample preparation prior to analysis.

3.2. Production Method and Farming System Misrepresentation

The differentiation of the production method of fish and seafood is another relevant aspect in certifying authenticity and traceability. During the last few years, the wild fish catches have been decreasing compared to the aquaculture production, thus supply of the market in farmed products has been growing very fast. From a compositional and organoleptic point of view, a wild fish is quite different from an aquaculture one, and this diversity is inevitably reflected on the different economic value of the two types of products [61–63]. By way of example, wild fish is usually characterised by higher levels of muscle protein, saturated, and polyunsaturated fatty acids, while farmed fish by a higher content of total lipid and monounsaturated fatty acids [64,65]. Consequently, the illegal substitution of higher-value wild fish with lower-value farmed fish is not an uncommon occurrence. Additionally, aquaculture fish consist of a number of high-variable products (i.e., extensively, semi-intensively, or intensively farmed fish, as well as organic or conventional farmed fish), whose final characteristics, since influenced by the husbandry environment and, above all, by the diet, are slight and very difficult to identify. This the reason is why the authentication of the production method (wild or farmed, organic or conventional), but also of the farming system of the aquaculture products is of extreme importance from the standpoint of fraud prevention and transparency towards consumers.

3.2.1. Application of Vibrational Spectroscopy

Among various vibrational spectroscopic methods applied to differentiate production processes and farming systems of fish, NIR is once again the most widely used. No application of UV or Raman spectroscopy, to the best of our knowledge, are currently available.

Ottavian et al. (2012) proposed a comparison between the classification performances of wild and farmed European sea bass obtained by three different NIR spectroscopic/chemometric approaches, and the classification performances obtained using only chemical and morphometric features [66]. The use of 1100–2500 nm raw spectra, WPTER-pre-treated spectra (wavelet packet transform for efficient pattern recognition), or of some parameters predicted by building a regression-based model, were found to be equivalent in terms of predictability assessed by PLS-DA and no differences between classification obtained by these models and classification obtained by using only chemical and morphometric data was observed. Moreover, authors identified (by using the variable influence of projection indexes, VIP) the wavelengths related to the absorbance of fat, fatty acids, and water as most influential in differentiating the production process of the fish tested.

More recently, the systems behind the production of European sea bass, was also investigated by applying unsupervised PCA and supervised OPLS-DA to 1100–2500 nm NIR spectra [67]. PCA built to SNV-SG-second derivative spectral data did not return a clear separation of groups, mainly as a consequence of the fact that the intraclass variability among samples was higher than the among-class variability between samples. A correct classification rate of 100% for both wild and farmed sea bass was instead achieved by OPLS-DA, and, in this case, authors found VIP indexes related to proteins exerting a greater contribution to the variance between the two types of fish. A deeper insight into the different farming systems of aquaculture samples, moreover, showed the ability of NIR and OPLS-DA to authenticate 67%, 80%, 100% of extensively, semi-intensively, and intensively-reared subjects, respectively, thanks above all to the spectral bands associated with protein absorption [67]. Concrete tank-cultured sea bass were also successfully discriminated from sea cage-cultured sea bass during storage, by means of Vis-NIR spectroscopy coupled with PLS-DA [68]. The best performances (87% of correct classification), were observed for spectral measurements performed at 48 h post mortem [68]. However, the greater contributions of the wavelengths to the PLS discrimination of samples analysed at 48 h post mortem were different from those of samples analysed at 96 h post, thus classification by farming system may have been affected also by other unrelated factors, such as the well-known compositional changes occurring during shelf life.

Authentication by NIR and SIMCA analysis of European sea bass raised in extensive ponds, semi-intensive ponds, intensive tanks, and intensive sea-cages, was also performed both on fresh fillets and freeze-dried fillets [69]. Authors found that freeze-drying the samples gave the best classification outcomes. The same results were obtained when classifying fresh minced fillets and freeze-dried fillets of farmed European sea bass according to the semi-intensive conventional or the organic production system [70]. SIMCA classification based on second-derivative spectra (1100–2500 nm) of samples, in fact, generated good results when fitted on the freeze-dried fillets (65–75% of correct classification), and worse results when performed on fresh fillets (20–25% of correct classification) [70]. All these results are particularly informative about problems posed by water when analysing high-moisture foods like fish. One of the main drawbacks of NIR spectroscopy is, in fact, the difficulty in separating relevant from useless information from spectra, in which peaks of water are predominant. These peaks, when included in chemometric calculations may hinder reliable features related to functional groups of molecules of interest and, thus, produce misleading results, especially when samples only slightly differ, such as in the case of fish reared under different conditions.

Following these principles, NIR spectroscopy was also used to directly authenticate freeze-dried rainbow trout fillets by rearing farm and, at the same time, to check whether NIR discriminating capability changed between raw and cooked freeze-dried fillets [71]. Rainbow trout samples came from three different aquaculture systems, varying in average well water temperatures, of which one consisted in indoor rearing at 11–14 °C, one in outdoor rearing at 9–11 °C, and one in outdoor rearing at 3–14 °C. Results for classification by farm (using SNV and second derivative 1100–2500 nm spectra of raw samples) showed approximately 97–100% of accuracy, with k-NN analysis giving the best overall statistical performances and PLS-DA the worst ones. As for cooked freeze-dried samples discrimination, the accuracy was approximately the same as those obtained for raw samples (90–100% for LDA, QDA, k-NN and 80% for PLS-DA), highlighting that the cooking process did not alter the capabilities of the technique to discriminate the sample by rearing farm [71].

3.2.2. Application of NMR Spectroscopy

Several applications of NMR spectroscopy aimed at authenticating the production process or the farming system were found in literature. In particular, proton (^{1}H) NMR spectroscopy can be used to analyse lipid mixtures such as fish oil, requiring simple preparation of samples and short time of spectra acquisition and providing a great deal of useful information [72]. Thus, considering that fish flesh lipids are the main compounds changing on the basis of the feeding regime, many attempts to use ^{1}H-NMR to identify the production process or the farming system were made. One of the earliest studies used SVM to elaborate ^{1}H-NMR spectra, and it was highly effective in predicting the wild or the farmed origin of salmon from different European countries [72]. Similarly, encouraging results were achieved through the combination of ^{1}H-NMR fingerprinting of lipids from gilthead sea bream with more complex chemometric data analyses [73]. The only unsupervised PCA applied on raw or processed ^{1}H-NMR spectral profiles returned, in fact, a clear separation between wild and farmed samples, which was found to be linked to methyl and methylene protons, together with methylene and methyne protons in unsaturated fatty acids [73]. Moreover, LDA variables selection allowed classification of 100% of the tested wild and farmed samples, and results from probabilistic neural network (PNN) analyses further reinforced the findings that such class discriminations were readily feasible.

If the previous studies were performed on fresh raw fish, other studies were intended to evaluate any differences in classification outcomes deriving from various degrees of fish processing. Lipids extracted from different types of processed Atlantic salmon products (frozen, smoked, and canned) were subjected to ^{1}H-NMR fingerprinting to develop models for determining labelling authenticity (wild/farmed) of these products [74]. SIMCA analysis applied to 138 pre-selected spectral peaks of NMR data, correctly classified as 100% of wild and 100% farmed samples, thanks mostly to the influence of a higher content of linoleic and oleic acid in farmed salmon compared to wild salmon [74]. A higher content of unsaturated fatty acids (and especially $n-3$ polyunsaturated fatty acids) was also found to play a special role in the discrimination between wild and farmed specimens of gilthead sea breams [75]. The influence exercised by these compounds was studied though the application of a supervised OPLS-DA to the whole lipid fingerprinting data obtained by ^{1}H-NMR spectroscopy. Just like SIMCA classification did in the previous study, OPLS-DA also led to a perfect separation of samples, but with the great advantage of being able to highlight the most effective variables in discrimination in the simplest of ways.

The ^{1}H-NMR molecular profiles of gilthead sea bream fish specimens produced according to different farming systems, have also been investigated, to seek out differences among three different kinds of aquaculture practices (cage, tank, and lagoon), but also any variations in the molecular patterns after a 16-day storage time under ice [76]. PCA-score plot of the pre-treated spectra showed a clear separation of fresh samples from ice-stored samples. At the same time, three distinct sub-clusters for each of the storage times, corresponding to the three farming systems investigated, highlighted the ability of the proposed methods to detect those molecular changes taking place during fish storage and exploited them for authentication purposes.

Another different NMR approach retrieved from the published literature concerned the use of carbon-13 (^{13}C) NMR instead of ^{1}H-NMR. Authors combined ^{13}C-NMR spectra of muscle lipids of Atlantic salmon with PNN and SVM chemometric elaborations, to discriminate between farmed and wild samples and obtained excellent discrimination performances (98.5% and 100.0% of correctly classified samples, respectively) [77]. Despite ^{13}C-NMR signals being generally much weaker than those provided by ^{1}H-NMR (as well as time of analysis is often longer), useful and complementary information can be obtained by this technique.

3.3. Geographical Origin Falsification

Proving the geographical origin authenticity of fish and seafood often involves the use of multi-disciplinary and cross-disciplinary approaches which take account of the environmental and genetic backgrounds affecting fish final characteristics [78]. Several published scientific researches concerning the use of spectroscopic methods pointed out the usefulness in classification of fish and seafood according to country or FAO area of origin.

3.3.1. Application of Vibrational Spectroscopy

Unlike the other authentication issues discussed above, NIR spectroscopy has been less explored for fish geographical origin identification. The reason, probably, is the great difficulty experienced in modelling total variability of NIR spectra and uniquely steering it to provenance, since provenance is the sum of a huge amount of different intrinsic or extrinsic factors (genetic, growth pattern, feeding regime, muscular activity, water temperature and salinity, etc.).

A traceability model able to predict the geographical origin of Chinese tilapia fillets coming from four different Chinese provinces, was developed by NIR spectroscopy [79]. SIMCA analysis, performed on 1000–2500 nm spectra of the minced samples, allowed more than 80% of fillets from Guangdong, Hainan, and Fujian provinces and 75% of fillets from the Fujian province to be correctly and exclusively assigned to the corresponding area of origin. Several locations in the Northern China Sea and East China Sea, from which sea cucumber (*Apostichopus japonicus*) come from, were also identified by using NIR spectroscopy [80]. In this case, authors found pre-treated (SNV or MSC, and second derivative) 1000–1800 nm spectra to give the best performance in PCA, since 100% correct classification rate was obtained both in the internal calibration model and in the external validation model. Similarly, 100% of sea cucumber analysed by means of diffuse reflectance MIR spectroscopy (fingerprint 5800–16,600 nm region) combined with SIMCA, were discriminated by the Chinese geographical region of provenance [81].

The last available application of NIR spectroscopy concerned the authentication of European sea bass according to Western, Central, or Eastern Mediterranean Sea provenances, by using OPLS-DA as a classification technique [67]. Results showed an overall discrimination performance of 89% according to these geographical origins, with 100% of Eastern, 88% of Central, and 85% of Western Mediterranean Sea samples being correctly classified. The VIP index analysis, moreover, identified lipid-associated bands as the most influential variables on the samples geographic discrimination.

3.3.2. Application of NMR Spectroscopy

Masoum et al. (2007) proposed a method for the origin authentication of Atlantic salmon based on ^{1}H-NMR and SVM of spectra extracted from samples coming from Canada, Alaska, Faroes, Ireland, Iceland, Norway, Scotland, and Tasmania. SVM returned a low degree of misclassification (4.6%) and, thus, an excellent correct classification rate for all the salmon samples [72]. Likewise, Aursand et al. (2009), used NMR combined with pattern recognition techniques to assess the geographical origin of Atlantic salmon and to verify the origin of market samples [77]. Here too, muscle lipids were extracted from tissues of fish coming from the same origins as those previously listed, but on the contrary, lipid composition was studied by ^{13}C-NMR coupled with PNN or SVM. This time, although the PNN- and SVM-based approaches used returned different correct classification rates (93.8% and 99.3%, respectively), a comparable classification accuracy between the two methodologies approaches was observed [77]. The ^{1}H-NMR lipid fingerprint, elaborated by LDA or PNN, allowed also to differentiate 76.2–100% of wild and farmed gilthead sea bream samples coming from Italy, Greece, Croatia, Turkey, and the Mediterranean Sea (for wild specimens), with better classification rates when PNN was applied [73]. Farmed gilthead sea bream specimens coming from five geographically distinct sites of Sardinia (Italy) and Greece were also discriminated by means of ^{1}H-NMR lipid fingerprint [75]. In this case, the fraction of unwanted variability related to the different production system of samples (off-shore sea cages and lagoon) was successfully overlooked thanks to the application of the OPLS-DA and, although authors did not provide statistical outcomes from internal or external classification, the significance of the clusters observed in the score plot was confirmed by bootstrap statistical analysis. The highest bootstrap values (indicating a well-defined class separation) were obtained for discrimination between Greek and Sardinian fish (100%), while lower but meaningful bootstrap values were obtained for discrimination among samples coming from different Sardinian offshore sea cage farms (68–57%) [75].

One last interesting application of ^{1}H-NMR dealt with the geographical authentication of bottarga, a fish-derived product consisting of salted and dried mullet (*Mugil Cephalus*) roe [82]. Low-molecular weight metabolites of aqueous extracts of samples, were analysed by PCA in order to identify clusters corresponding to one of the specific geographical provenances studied, namely FAO 37.1.3, FAO 34, FAO 41, FAO 31, and one unknown provenance. Results from PCA confirmed the possibility to characterise bottarga samples having different geographical origins, since samples with the same known geographical origin were closely clustered in the same region of the PCA scores plot, and those of different origin were far away from each other.

3.4. Discrimination between Fresh and Frozen/Thawed Fish and Seafood

Fish is commonly processed by freezing in order to be preserved from deterioration. Frozen fish, however, is usually characterised by much lower quality and commercial value compared to fresh fish. Therefore, fraudulent practices consisting in the substitution of fresh with frozen/thawed products are not uncommon events [83]. Considering that labelling of fish must state if the fish is fresh, frozen, or previously frozen (or refreshed), discriminating fresh from frozen/thawed products is one of the most important authenticity issues. The differentiation between fresh and frozen/thawed products is hampered by difficulties in detecting those tiny physical and chemical variations occurring during freeze storage, which, moreover, do not cause any perceptible organoleptic change [83,84]. Therefore, the rapid confirmation of fish freshness by spectroscopy has been widely studied during the last few years and several published researches are currently available.

3.4.1. Application of Fluorescence and Vibrational Spectroscopy

Front-face fluorescence spectroscopy is one of the earliest spectroscopic techniques historically applied to differentiate fresh from frozen/thawed fish. It has been demonstrated that typical changes in fluorescence spectra of aromatic amino acids, nucleic acids, and nicotinamide adenine dinucleotide (NADH) occur during storage, as a consequence of several reactions involving free amino acids and carbonyl compounds of reducing sugars, formaldehyde (produced from trimethylamine oxide), and malondialdehyde (produced from oxidation of fish lipids during storage). Therefore, changes in fluorescence of fish samples may be considered as fingerprints for fresh and aged fish fillet identification [85]. The fluorescence emission spectra of tryptophan (305–400 nm) recorded directly on whiting fillets and elaborated by factorial discriminant analysis (FDA) led to correct classification rates of 62.5% and 70.8% in the calibration and validation set, respectively. NADH fluorescence spectra (360–570 nm), indeed, were found to have a higher potential to differentiate fresh from frozen/thawed products as they allowed to achieve 100% of correct discrimination for both calibration and validation set [85]. More recently, the same authors confirmed the success of a similar methodology in authenticating freshness of sea bass samples. Fluorescence emission spectra at 340 and 380 nm, elaborated by FDA, led to 94.87% of total correct classification rate [86]. Additionally, the elaboration of NADH fluorescence spectra by Fisher's linear discriminant analysis, was stated as a reliable method to rapidly discriminate fresh and frozen/thawed large yellow croaker fillets, since 100% of total correct classification rate was achieved [87].

More applications of IR spectroscopy are reported in the published literature. Uddin and Okazaki (2004) used NIR reflectance spectroscopy on dry extract of horse mackerel specimens to evaluate freshness [88]. Both PCA (using 1100–2500 nm spectra) and SIMCA analysis (using only three selected wavelengths which were strongly related to protein content) successfully discriminated 100% of fresh and frozen/thawed samples. Thereafter, the same authors performed further investigations on fresh and frozen/thawed red sea bream by using Vis-NIR spectroscopy in the 400–1100 nm region [89]. In this case, raw spectra were used to build an LDA model, by which 100% classification accuracy in prediction was reached. PLS-DA of SG-smoothed spectra (670–1100 nm) of shrimps subjected to different treatments (including ice, water, and brine at various salt concentrations), also led to 100% of fresh and frozen/thawed samples to be authenticated [90].

Another study was directed to compare classification ability of Vis-NIR (380–1080 nm) and NIR (1100–2500) spectroscopy in authenticating fresh and frozen/thawed swordfish and, through the application of PLS-DA, it was found that in this case, Vis-NIR spectra gave better results in the external validation (≥96.7% of correctly classified samples) [91]. Although worse outcomes were obtained by only using the NIR region, the technique, combined with SVMs, also authenticated 93% of fresh and 83% of frozen/thawed sole (*Solea vulgaris*) samples [92]. Again, high accuracy (90%) and sensitivity (80%) in prediction were observed for the discrimination of fresh and frozen/thawed tuna sample by Vis-NIR spectral analysis (350–2500 nm) combined with PLS-DA [93], while better and more homogenous SIMCA prediction results were obtained when using MIR (2500–14,300 nm) instead of

NIR (800–2500 nm) regions for the discrimination between fresh and previously frozen Atlantic mullet fillets [94].

Ottavian et al. (2013) proposed an interesting three-step approach based only on NIR spectra and latent variable modelling techniques to develop a species-independent classifier able to simultaneously discriminate between fresh and frozen/thawed fish and, remarkably, overall classification accuracy of the method ranged between 80% and 91%, based on the strategy adopted and the instrument used [94]. By contrast, the only MIR region was found to be useful for determining whether whiting fish fillets have been frozen/thawed: when FDA was applied to the 3300–3570 nm MIR subregion (usually related to fatty acids absorption), 87.5% of sample spectra in the validation set was correctly identified [95].

Finally, one single application of Raman spectroscopy to the authentication of fresh fish is now available [59]. Lipid fraction of fish from several species (horse mackerel, European anchovy, bluefish, Atlantic salmon, red mullet, and flying gurnard) was extracted from three samples batches (fresh samples, once frozen/towed samples, and twice frozen/thawed samples), and then collected by a Raman spectrometer along the 5000–50,000 nm spectral range and using a 785 nm laser exciting source. Chemometric analysis, performed by PCA, identified three different clusters in the score plot, each corresponding to one of the three batches of fish investigated [59].

3.4.2. Application of Hyperspectral Imaging Spectroscopy

Discrimination between fresh and frozen/thawed cod fillet was studied by Vis-NIR/HSI, using both a handheld interactance probe and an imaging spectrometer (for automatic online analysis at typical industrial speeds) [96]. Spectra resulting from the two instruments were pre-treated (SNV and second derivative) and statistically analysed by applying the Rosenblatt's perceptron linear classifier to the first and third principal component of the imaging data. Results showed that fresh cod fillets can be completely separated from fresh/thawed cod fillets using only a few wavelengths in the Vis region, mainly related to the oxidation of haemoglobin and myoglobin which occur during freezing/thawing [97]. Similarly, hyperspectral data from Vis-NIR/HSI (380–1030 nm) combined with least square-SVMs, returned an average correct classification rate of 91.67% for fresh and frozen/thawed halibut fillets [97].

3.4.3. Application of NMR Spectroscopy

NMR spectroscopy is considered to be a useful and suitable tool for the discrimination of fresh from frozen/thawed fish, since NMR signals are sensitive enough to changes in water mobility and its interaction with other molecules [98]. NMR spectroscopy has been already widely exploited to identify the various modifications in fish tissues occurring during freezing and thawing of fish [99–102]; however, as far as we know, no application of this technique for fish freshness authentication is currently available.

4. Critical Aspects and Limitations to Overcome

The food scientists' interest towards the development of reliable methods for the resolution of several food authenticity issues is well documented by the increasing number of scientific works which, albeit through different methodologies, have attempted to address the same problems. It is clear from the analysis of the latest literature that spectroscopy combined with chemometrics is just one of the many untargeted strategies adopted: chromatographic, MS-based, as well as bio-molecular and sensory techniques have been already widely exploited and have demonstrated their exceptional multipurpose qualities for fish authenticity testing [78,103–108].

These techniques are known to share certain common disadvantages, such as the long time needed for analysis, high costs of the equipment, the need of sample preparation prior to analysis, destructiveness, and the demand for qualified personnel. On the other hand, as they become more consolidated within the research community, these techniques excel by their higher accuracy, specificity, and sensitivity compared to spectroscopic ones, to the point that many of them are used in food official controls. Despite this, the attractiveness of spectroscopy and chemometrics is evidenced by not only by the large literature provided in the present review, but also by several other applications covering a wide range of food and foodstuffs: fruits and vegetables, honey, wine, edible oils and fats, cereal and cereal-based products, milk, and dairy products [109–114] have been successfully investigated and authenticated by means of spectroscopy.

Having said that, some critical reflections should be made about the problems related to the use of spectroscopy and chemometrics, which still have not been overcome. In accordance to what has been already reported and to our opinion, the research papers analysed were found to be highly variable to each other in terms of analytical set-up (e.g., sample pre-processing, spectral ranges, spectra pre-treatments, resolutions, number of samples tested, and statistical elaboration). This variability, as easily understood from Section 3, is further worsened by the fact that only a few of the works analysed reported in-depth statistical outputs and, where present, they were not comparable to each other.

A critical and objective evaluation of these works is also severely hampered by a lack, in certain cases, of comprehensive data with regards to the validation of the results. Alongside the internal cross-validation, the external validation of the qualitative chemometric model is, in our opinion, a crucial point in assessing the overall goodness of the classifiers and avoiding misleading interpretations. The last aspect which should be emphasised is that a detailed description of the characteristics of the sample dataset was not often reported and the lack of standardisation of external factors (e.g., storage times and conditions), may have interfered with spectral analysis, possibly affecting the robustness of the model. In this scenario, a recommendation for future works is to consider the intrinsically natural variability of the fish products (as well as those of all other foodstuffs), and to organise the sampling in such a way that as much of the expected variability of samples is collected during the calibration stage. That way, the robustness of the models can make their way to the spread of applications also in the industrial sector.

As a final remark, no technique should be universally regarded as the optimal solution. However, the possibility of using UV, IR, Raman, and NMR spectroscopies with no distinction for food authentication purposes is still an obstacle to overcome, and therefore, in accordance to our experience, untargeted NIR spectroscopy represents the most versatile option thanks to its high sensitivity to organic molecules of food, cost-effectiveness, and ease of use. Additionally, the use of NIR spectroscopy with supervised chemometric method, able to separate relevant from non-relevant spectral variation like OPLS-DA, should be encouraged since the interpretability of results is enhanced.

5. Conclusions and Prospects for the Future

Recent increases in the complexity and competitiveness of the fishery and seafood sectors, have resulted in the presence, on the international market, of a huge variety of fresh and processed products, but at the same time, have meant that the risk of fraud deriving from substitution among look-alike products is now exponentially higher than it was even a few years ago. Thus, ensuring the truthfulness of fish and seafood claims concerning their quality and origin, has become an exceptionally important topic, firstly with a view to enable consumers to make informed decisions.

The overview presented in this review clearly highlights the effective support provided by analytical approaches based on spectroscopy and multivariate data analysis for the evaluation and monitoring of fish and seafood products authenticity. Fluorescence, vibrational, NMR, and HSI spectroscopic applications have been discussed, with an accent on the trends toward their use for several authentication purposes. In this connection, IR spectroscopy has been the most exploited technique, especially in studies concerning species and fresh for frozen/thawed products substitutions. NMR, instead, has shown many applications in the field on the production method, farming system, and geographical origin identification. By contrast, Raman and HIS have provided very encouraging results in some fish authentication fields, but their overall potential has so far been largely ignored.

Rapidity, non-destructive nature, ease of use, and high-throughput measurements make the spectroscopic non-targeted approach an ideal tool for quality control operations, especially in the context of daily routine and screening analysis in the food industry, and as a possible substitute of traditional analytical techniques. Thanks to the technological development of the spectroscopic instrumentation, the availability of miniaturized and portable devices on the market is rapidly growing, and this will contribute to an additional growth of applications in the food sector. On the other hand, these analytical strategies in the official control of foodstuffs are still far from being effectively applied, largely due to the need of a strict validation to assure further reliability and robustness of results before implementation as standalone tools. For these reasons, standardisation of the working conditions, optimisation of the chemometric software, and creation of large databases for data-sharing and for encouraging greater cooperation between food scientists, represent important current research fields and future challenges to be faced.

Author Contributions: All authors contributed equally to this work.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ANN	artificial neural networks;
BBN	Bayesian belief network;
FIR	far-infrared;
FDA	factorial discriminant analysis;
FFT	fast Fourier transform;
FT	Fourier transform;
HCA	hierarchical cluster analysis;
HSI	hyperspectral imaging;
IR	infrared;
k-NN	k-nearest neighbors;
LDA	linear discriminant analysis;
LW-NIR	long-wave near infrared;
MIR	mid-infrared;
NMR	nuclear magnetic resonance;
MSC	multiplicative scatter correction;
NIR	near-infrared;
OPLS-DA	orthogonal partial least square-discriminant analysis;
PCA	principal component analysis;
PLS-DA	partial least square-discriminant analysis;
PNN	probabilistic neural network;
QDA	quadratic factorial analysis;
SERS	surface-enhanced Raman spectroscopy;
SG	Savitzky–Golay smoothing;
SIMCA	soft independent modelling of class analogy;
SNV	standard normal variate;
SVM	support vector machine;
SW-NIR	short-wave near infrared;
UV	ultraviolet;
Vis	visible.

References

1. FAO. *The State of World Fisheries and Aquaculture 2018–Meeting the Sustainable Development Goals*; FAO: Rome, Italy, 2018; Volume 35, pp. 52–62. ISBN 978-92-5106-029-2.
2. Rehbein, H.; Oehlenschläger, J. Basic Facts and Figures. In *Fishery products: Quality, Safety and Authenticity*; Rehbein, H., Oehlenschläger, J., Eds.; John Wiley & Sons: Chichester, West Sussex, UK, 2009; Volume 1, pp. 1–18. ISBN 978-1-4051-4162-8.
3. FAO. *Overview of Food Fraud in the Fisheries Sector, by Alan Reilly; Fisheries and Aquaculture Circular No. 1165*; FAO: Rome, Italy, 2018; pp. 4–6. ISBN 978-92-5-130402-0.
4. Van Dolah, F.M. Marine Algal Toxins: Origins, Health Effects, and Their Increased Occurrence. *Environ. Health Perspect.* **2000**, *108*, 133–141. [CrossRef]
5. Fallah, A.A.; Saei-Dehkordi, S.S.; Nematollahi, A.; Jafari, T. Comparative study of heavy metal and trace element accumulation in edible tissues of farmed and wild rainbow trout (Oncorhynchus mykiss) using ICP-OES technique. *Microchem. J.* **2011**, *98*, 275–279. [CrossRef]
6. Okocha, R.C.; Olatoye, I.O.; Adedeji, O.B. Food safety impacts of antimicrobial use and their residues in aquaculture. *Public Health Rev.* **2018**, *39*, 1–22. [CrossRef]
7. Kelly, B.C.; Ikonomou, M.G.; Higgs, D.A.; Oakes, J.; Dubetz, C. Flesh residue concentrations of organochlorine pesticides in farmed and wild salmon from British Columbia, Canada. *Environ. Toxicol. Chem.* **2011**, *30*, 2456–2464. [CrossRef]
8. Council regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. *Off. J. Eur. Commun.* **2002**, *31*, 1–24.
9. Council regulation (EC) No 1005/2008 of 29 September 2008 establishing a Community system to prevent, deter and eliminate illegal, unreported and unregulated fishing, amending Regulations (EEC) No 2847/93, (EC) No 1936/2001 and (EC) No 601/2004 and repealing Regulations (EC) No 1093/94 and (EC) No 1447/1999. *Off. J. Eur. Union* **2008**, *286*, 1–32.
10. Council regulation (EC) No 1224/2009 of 20 November 2009 establishing a Community control system for ensuring compliance with the rules of the common fisheries policy, amending Regulations (EC) No 847/96, (EC) No 2371/2002, (EC) No 811/2004, (EC) No 768/2005, (EC) No 2115/2005, (EC) No 2166/2005, (EC) No 388/2006, (EC) No 509/2007, (EC) No 676/2007, (EC) No 1098/2007, (EC) No 1300/2008, (EC) No 1342/2008 and repealing Regulations (EEC) No 2847/93, (EC) No 1627/94 and (EC) No 1966/2006. *Off. J. Eur. Union* **2009**, *343*, 1–50.
11. Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers, amending Regulations (EC) No 1924/2006 and (EC) No 1925/2006 of the European Parliament and of the Council, and repealing Commission Directive 87/250/EEC, Council Directive 90/496/EEC, Commission Directive 1999/10/EC, Directive 2000/13/EC of the European Parliament and of the Council, Commission Directives 2002/67/EC and 2008/5/EC. *Off. J. Eur. Union* **2011**, *304*, 18–63.
12. Regulation (EU) No 1379/2013 of the European Parliament and of the Council of 11 December 2013 on the common organisation of th emarkets in fishery and aquaculture products, amending Council Regulations (EC) No 1184/2006 and (EC) No 1224/2009 and repealing Council Regulation (EC) No 104/2000. *Off. J. Eur. Union* **2013**, *354*, 12–14.
13. Esslinger, S.; Riedl, J.; Fauhl-Hassek, C. Potential and limitations of non-targeted fingerprinting for authentication of food in official control. *Food Res. Int.* **2014**, *60*, 189–204. [CrossRef]
14. Picò, Y. Near-Infrared, Mid-Infrared, and Raman Spectroscopy. In *Chemical Analysis of Food: Techniques and Applications*; Picò, Y., Ed.; Academic Press: San Francisco, CA, USA, 2012; Volume 1, pp. 59–91. ISBN 978-0-1238-4862-8.
15. Schrieber, A. Introduction to Food Authentication. In *Modern Techniques for Food Authentication*; Sun, D.W., Ed.; Academic Press: San Francisco, CA, USA, 2008; pp. 1–21.
16. Penner, M.H. Basic Principles of Spectroscopy. In *Food Analysis, Food Science Text Series*, 2nd ed.; Nielsen, S.S., Ed.; Springer: Cham, Switzerland; New York, NY, USA, 2017; pp. 79–88. ISBN 978-3-319-45776-5.
17. Strasburg, G.M.; Ludescher, R.D. Theory and applications of fluorescence spectroscopy in food research. *Trends Food Sci. Technol.* **1995**, *6*, 69–75. [CrossRef]

18. Xu, J.L.; Riccioli, C.; Sun, D.W. An Overview on Nondestructive Spectroscopic Techniques for Lipid and Lipid Oxidation Analysis in Fish and Fish Products. *Compr. Rev. Food Sci. Food Saf.* **2015**, *4*, 466–477. [CrossRef]
19. Rodriguez-Saona, L.E.; Allendorf, M.E. Use of FTIR for rapid authentication and detection of adulteration of food. *Ann. Rev. Food Sci. Technol.* **2011**, *2*, 467–483. [CrossRef] [PubMed]
20. Rodriguez-Saona, L.E.; Giusti, M.M.; Shotts, M. *Advances in Infrared Spectroscopy for Food Authenticity Testing*; Downey, G., Ed.; Woodhead Publishing: Duxford, UK, 2016; ISBN 978-0-08-100220-9.
21. Lohumi, S.; Lee, S.; Lee, H.; Cho, B.K. A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. *Trends Food Sci. Technol.* **2015**, *1*, 85–98. [CrossRef]
22. Blanco, M.; Villarroya, I.N.I.R. NIR spectroscopy: A rapid-response analytical tool. *TrAC Trends Anal. Chem.* **2002**, *21*, 240–250. [CrossRef]
23. Cen, H.; Yong, H. Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci. Technol.* **2007**, *18*, 72–83. [CrossRef]
24. Cheng, J.H.; Dai, Q.; Sun, D.W.; Zeng, X.A.; Liu, D.; Pu, H.B. Applications of non-destructive spectroscopic techniques for fish quality and safety evaluation and inspection. *Trends Food Sci. Technol.* **2013**, *34*, 18–31. [CrossRef]
25. Cozzolino, D.; Murray, I. A review on the application of infrared technologies to determine and monitor composition and other quality characteristics in raw fish, fish products, and seafood. *Appl. Spectrosc. Rev.* **2012**, *47*, 207–218. [CrossRef]
26. Stuart, B.H. Spectral Analysis. In *Infrared Spectroscopy: Fundamentals and Applications*; Stuart, B.H., Ed.; Jhon Wiley & Sons Ltd.: Chichester, WS, UK, 2004; pp. 47–48. ISBN 0-470-85427-8.
27. Boyaci, I.H.; Temiz, H.T.; Geniş, H.E.; Soykut, E.A.; Yazgan, N.N.; Güven, B.; Uysal, R.S.; Bozkurt, A.G.; Ilaslan, K.; Torun, O.; et al. Dispersive and FT-Raman spectroscopic methods in food analysis. *RSC Adv.* **2015**, *5*, 56606–56624. [CrossRef]
28. Zheng, J.; He, L. Surface-Enhanced Raman Spectroscopy for the Chemical Analysis of Food. *Compr. Rev. Food Sci. Food Saf.* **2014**, *13*, 317–328. [CrossRef]
29. Feng, Y.Z.; Sun, D.W. Application of hyperspectral imaging in food safety inspection and control: A review. *Crit. Rev. Food Sci. Nutr.* **2012**, *52*, 1039–1058. [CrossRef]
30. Wu, D.; Sun, D.W. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review—Part I: Fundamentals. *Innov. Food Sci. Emerg. Technol.* **2013**, *19*, 1–14. [CrossRef]
31. Roberts, J.; Power, A.; Chapman, J.; Chandra, S.; Cozzolino, D. A short update on the advantages, applications and limitations of hyperspectral and chemical imaging in food authentication. *Appl. Sci.* **2018**, *8*, 505. [CrossRef]
32. Sacchi, R.; Paolillo, L. NMR for Food Quality and Traceability. In *Advances in Food Diagnostics*; Nollet, L.M.L., Toldrà, F., Eds.; Blackwell Publishing: Oxford, UK, 2007; Volume 1, pp. 101–107. ISBN 978-0-4702-7780-5.
33. Hatzakis, E. Nuclear Magnetic Resonance (NMR) Spectroscopy in Food Science: A Comprehensive Review. *Compre. Rev. Food Sci. Food Saf.* **2019**, *18*, 189–220. [CrossRef]
34. Sobolev, A.P.; Circi, S.; Mannina, L. Advances in Nuclear Magnetic Resonance Spectroscopy for Food Authenticity Testing. In *Advances in Food Authenticity Testing*; Downey, G., Ed.; Woodhead Publishing: Duxford, UK, 2016; Volume 1, pp. 147–170. [CrossRef]
35. Oliveri, P.; Simonetti, R. Chemometrics for food authenticity applications. In *Advances in Food Authenticity Testing*; Downey, G., Ed.; Woodhead Publishing: Duxford, UK, 2016; pp. 701–728. [CrossRef]
36. Manley, M.; Baeten, V. Spectroscopic technique: Near infrared (NIR) spectroscopy. In *Modern Techniques for Food Authentication*; Sun, D.W., Ed.; Academic Press: San Francisco, CA, USA, 2008; Volume 1, pp. 51–102.
37. Rinnan, Å.; Van Den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal. Chem.* **2009**, *28*, 1201–1222. [CrossRef]
38. Pereira, A.C.; Reis, M.S.; Saraiva, P.M.; Marques, J.C. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD. *Chemom. Intell. Lab. Syst.* **2011**, *105*, 43–55. [CrossRef]
39. Rinnan, Å. Pre-processing in vibrational spectroscopy–when, why and how. *Anal. Methods* **2014**, *6*, 7124–7129. [CrossRef]
40. Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Downey, G.; Blanchet, L.; Buydens, L.M.C. Breaking with trends in pre-processing? *TrAC Trends Anal. Chem.* **2013**, *50*, 96–106. [CrossRef]

41. Ziegel, E.R. A User-Friendly Guide to Multivariate Calibration and Classification. *Technometrics* **2004**, *46*, 108–111. [CrossRef]
42. Voncina, D.B. Chemometrics in analytical chemistry. *Nova Biotechnol.* **2009**, *9*, 211–216. [CrossRef]
43. Berrueta, L.A.; Alonso-Salces, L.M.; Heberger, K. Supervised pattern recognition in food analysis. *J. Chromatog. A* **2007**, *1158*, 196–214. [CrossRef] [PubMed]
44. Ahmad, A.R.; Khalid, M.; Yusof, R. Machine Learning Using Support Vector Machines. In Proceedings of the International Conference on Artificial Intelligence in Science and Technology, Hobart, Australia, 19–21 September 2002.
45. Tetko, I.V.; Livingstone, D.J.; Luik, A.I. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826–833. [CrossRef]
46. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J.K.; Holmes, E.; Trygg, J. OPLS Discriminant Analysis, Combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2007**, *20*, 341–351. [CrossRef]
47. He, H.J.; Wu, D.; Sun, D.W. Nondestructive Spectroscopic and Imaging Techniques for Quality Evaluation and Assessment of Fish and Fish Products. *Crit. Rev. Food Sci. Nutr.* **2015**, *55*, 864–886. [CrossRef] [PubMed]
48. Uddin, M.; Okazaki, E. Applications of Vibrational Spectroscopy to the Analysis of Fish and Other Aquatic Food Products. *Handb. Vib. Spectro.* **2006**, 439–459. [CrossRef]
49. Fiorino, G.M.; Garino, C.; Arlorio, M.; Logrieco, A.F.; Losito, I.; Monaci, L. Overview on Untargeted Methods to Combat Food Frauds: A Focus on Fishery Products. *J. Food Qual.* **2018**, *3*, 1–13. [CrossRef]
50. Gayo, J.; Hale, S.A.; Blanchard, S.M. Quantitative analysis and detection of adulteration in crab meat using visible and near-infrared spectroscopy. *J. Agric. Food Chem.* **2006**, *54*, 1130–1136. [CrossRef]
51. O'Brien, N.; Hulse, C.A.; Pfeifer, F.; Siesler, H.W. Near infrared spectroscopic authentication of seafood. *J. Near Infrared Spectrosc.* **2013**, *21*, 299–305. [CrossRef]
52. Grassi, S.; Casiraghi, E.; Alamprese, C. Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties. *Food Chem.* **2018**, *243*, 382–388. [CrossRef]
53. Lv, H.; Xu, W.; You, J.; Xiong, S. Classification of freshwater fish species by linear discriminant analysis based on near infrared reflectance spectroscopy. *J. Near Infrared Spectrosc.* **2017**, *25*, 54–62. [CrossRef]
54. Zhang, X.Y.; Hu, W.; Teng, J.; Peng, H.H.; Gan, J.H.; Wang, X.C.; Sun, S.Q.; Xu, C.H.; Liu, Y. Rapid recognition of marine fish surimi by one-step discriminant analysis based on near-infrared diffuse reflectance spectroscopy. *Int. J. Food Prop.* **2017**, *20*, 2932–2943. [CrossRef]
55. Alamprese, C.; Casiraghi, E. Application of FT-NIR and FT-IR spectroscopy to fish fillet authentication. *LWT–Food Sci. Technol.* **2015**, *63*, 720–725. [CrossRef]
56. Cozzolino, D.; Chree, A.; Scaife, J.R.; Murray, I. Usefulness of Near-infrared reflectance (NIR) spectroscopy and chemometrics to discriminate fishmeal batches made with different fish species. *J. Agric. Food Chem.* **2005**, *53*, 4459–4463. [CrossRef] [PubMed]
57. Sousa, N.; Moreira, M.; Saraiva, C.; de Almeida, J. Applying Fourier Transform Mid Infrared Spectroscopy to Detect the Adulteration of Salmo salar with Oncorhynchus mykiss. *Foods* **2018**, *7*, 55. [CrossRef]
58. Rašković, B.; Heinke, R.; Rösch, P.; Popp, J. The Potential of Raman Spectroscopy for the Classification of Fish Fillets. *Food Anal. Methods* **2016**, *9*, 1301–1306. [CrossRef]
59. Velioğlu, H.M.; Temiz, H.T.; Boyaci, I.H. Differentiation of fresh and frozen-thawed fish samples using Raman spectroscopy coupled with chemometric analysis. *Food Chem.* **2015**, *173*, 283–290. [CrossRef]
60. Standal, I.B.; Axelson, D.E.; Aursand, M. 13C NMR as a tool for authentication of different gadoid fish species with emphasis on phospholipid profiles. *Food Chem.* **2010**, *121*, 608–615. [CrossRef]
61. Gabr, H.R.; Gab-Alla, A.A.F.A. Comparison of biochemical composition and organoleptic properties between wild and cultured finfish. *J. Fish. Aquat. Sci.* **2007**, *2*, 77–81. [CrossRef]
62. Grigorakis, K.; Taylor, K.D.A.; Alexis, M.N. Organoleptic and volatile aroma compounds comparison of wild and cultured gilthead sea bream (Sparus aurata): Sensory differences and possible chemical basis. *Aquaculture* **2003**, *225*, 109–119. [CrossRef]
63. Grigorakis, K. Compositional and organoleptic quality of farmed and wild gilthead sea bream (Sparus aurata) and sea bass (Dicentrarchus labrax) and factors affecting it: A review. *Aquaculture* **2007**, *272*, 55–75. [CrossRef]
64. Lenas, D.; Chatziantoniou, S.; Nathanailides, C.; Triantafillou, D. Comparison of wild and farmed sea bass (*Dicentrarchus labrax* L) lipid quality. *Procedia Food Sci.* **2011**, *1*, 1139–1145. [CrossRef]

65. Fuentes, A.; Fernández-Segovia, I.; Serra, J.A.; Barat, J.M. Comparison of wild and cultured sea bass (Dicentrarchus labrax) quality. *Food Chem.* **2010**, *119*, 1514–1518. [CrossRef]
66. Ottavian, M.; Facco, P.; Fasolato, L.; Novelli, E.; Mirisola, M.; Perini, M.; Barolo, M. Use of near-infrared spectroscopy for fast fraud detection in seafood: Application to the authentication of wild European sea bass (Dicentrarchus labrax). *J. Agric. Food Chem.* **2012**, *60*, 639–648. [CrossRef]
67. Ghidini, S.; Varrà, M.O.; Dall'Asta, C.; Badiani, A.; Ianieri, A.; Zanardi, E. Rapid authentication of European sea bass (*Dicentrarchus labrax* L.) according to production method, farming system, and geographical origin by near infrared spectroscopy coupled with chemometrics. *Food Chem.* **2019**, *280*, 321–327. [CrossRef]
68. Costa, C.; D'Andrea, S.; Russo, R.; Antonucci, F.; Pallottino, F.; Menesatti, P. Application of non-invasive techniques to differentiate sea bass (Dicentrarchus labrax, L. 1758) quality cultured under different conditions. *Aquac. Int.* **2011**, *19*, 765–778. [CrossRef]
69. Xiccato, G.; Trocino, A.; Tulli, F.; Tibaldi, E. Prediction of chemical composition and origin identification of european sea bass (*Dicentrarchus labrax* L.) by near infrared reflectance spectroscopy (NIRS). *Food Chem.* **2004**, *86*, 275–281. [CrossRef]
70. Trocino, A.; Xiccato, G.; Majolini, D.; Tazzoli, M.; Bertotto, D.; Pascoli, F.; Palazzi, R. Assessing the quality of organic and conventionally-farmed European sea bass (*Dicentrarchus labrax*). *Food Chem.* **2012**, *131*, 427–433. [CrossRef]
71. Dalle Zotte, A.; Ottavian, M.; Concollato, A.; Serva, L.; Martelli, R.; Parisi, G. Authentication of raw and cooked freeze-dried rainbow trout (Oncorhynchus mykiss) by means of near infrared spectroscopy and data fusion. *Food Res. Int.* **2014**, *60*, 180–188. [CrossRef]
72. Masoum, S.; Malabat, C.; Jalali-Heravi, M.; Guillou, C.; Rezzi, S.; Rutledge, D.N. Application of support vector machines to 1H NMR data of fish oils: Methodology for the confirmation of wild and farmed salmon and their origins. *Anal. Bioanal. Chem.* **2007**, *387*, 1499–1510. [CrossRef]
73. Rezzi, S.; Giani, I.; Héberger, K.; Axelson, D.E.; Moretti, V.M.; Reniero, F.; Guillou, C. Classification of gilthead sea bream (Sparus aurata) from1H NMR lipid profiling combined with principal component and linear discriminant analysis. *J. Agric. Food Chem.* **2007**, *55*, 9963–9968. [CrossRef]
74. Capuano, E.; Lommen, A.; Heenan, S.; de la Dura, A.; Rozijn, M.; van Ruth, S. Wild salmon authenticity can be predicted by 1H-NMR spectroscopy. *Lipid Technol.* **2012**, *24*, 251–253. [CrossRef]
75. Melis, R.; Cappuccinelli, R.; Roggio, T.; Anedda, R. Addressing marketplace gilthead sea bream (*Sparus aurata* L.) differentiation by 1H NMR-based lipid fingerprinting. *Food Res. Int.* **2014**, *63*, 258–264. [CrossRef]
76. Picone, G.; Balling, S.; Engelsen, F.; Savorani, S.; Testi, S.; Badiani, A.; Capozzi, F. Metabolomics as a powerful tool for molecular quality assessment of the fish Sparus aurata. *Nutrients* **2011**, *3*, 212–227. [CrossRef] [PubMed]
77. Aursand, M.; Standal, I.B.; Praél, A.; Mcevoy, L.; Irvine, J.; Axelson, D.E. 13C NMR pattern recognition techniques for the classification of atlantic salmon (*salmo salar* L.) according to their wild, farmed, and geographical origin. *J. Agric. Food Chem.* **2009**, *57*, 3444–3451. [CrossRef] [PubMed]
78. Abbas, O.; Zadravec, M.; Baeten, V.; Mikuš, T.; Lešić, T.; Vulić, A.; Prpić, J.; Jemeršić, L.; Pleadin, J. Analytical methods used for the authentication of food of animal origin. *Food Chem.* **2018**, *246*, 6–17. [CrossRef]
79. Liu, Y.; Ma, D.H.; Wang, X.C.; Liu, L.P.; Fan, Y.X.; Cao, J.X. Prediction of chemical composition and geographical origin traceability of Chinese export tilapia fillets products by near infrared reflectance spectroscopy. *LWT–Food Sci. Technol.* **2015**, *60*, 1214–1218. [CrossRef]
80. Guo, X.; Cai, R.; Wang, S.; Tang, B.; Li, Y.; Zhao, W. Non-destructive geographical traceability of sea cucumber (Apostichopus japonicus) using near infrared spectroscopy combined with chemometric methods. *R. Soc. Open Sci.* **2018**, *5*. [CrossRef] [PubMed]
81. Wu, Z.; Tao, L.; Zhang, P.; Li, P.; Zhu, Q.; Tian, Y.; Yang, T. Diffuse reflectance mid-infrared Fourier transform spectroscopy (DRIFTS) for rapid identification of dried sea cucumber products from different geographical areas. *Vib. Spectro.* **2010**, *53*, 222–226. [CrossRef]
82. Locci, E.; Piras, C.; Mereu, S.; Cesare Marincola, F.; Scano, P. 1H NMR metabolite fingerprint and pattern recognition of mullet (Mugil cephalus) bottarga. *J. Agric. Food Chem.* **2011**, *59*, 9497–9505. [CrossRef]
83. Verrez-Bagnis, V.; Sotelo, C.G.; Mendes, R.; Silva, H.; Kappel, K.; Schröder, U. Methods for Seafood Authenticity Testing in Europe. In *Bioactive Molecules in Food*; Springer: Cham, Switzerland; New York, NY, USA, 2018; Volume 1, pp. 1–55.

84. Tokur, B.; Ozkütük, S.; Atici, E.; Ozyurt, G.; Ozyurt, C.E. Chemical and sensory quality changes of fish fingers, made from mirror carp (*Cyprinus carpio* L., 1758), during frozen storage (−18 °C). *Food Chem.* **2006**, *99*, 335–341. [CrossRef]

85. Karoui, R.; Thomas, E.; Dufour, E. Utilisation of a rapid technique based on front-face fluorescence spectroscopy for differentiating between fresh and frozen–thawed fish fillets. *Food Res. Int.* **2006**, *39*, 349–355. [CrossRef]

86. Karoui, R.; Hassoun, A.; Ethuin, P. Front face fluorescence spectroscopy enables rapid differentiation of fresh and frozen-thawed sea bass (Dicentrarchus labrax) fillets. *J. Food Eng.* **2017**, *202*, 89–98. [CrossRef]

87. Gao, Y.; Tang, H.; Ou, C.; Li, Y.; Wu, C.; Cao, J. Differentiation between fresh and frozen-thawed large yellow croaker based on front-face fluorescence spectroscopy technique. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 279–285. [CrossRef]

88. Uddin, M.; Okazaki, E. Classification of fresh and frozen-thawed fish by near-infrared spectroscopy. *J. Food Sci.* **2004**, *69*, C665–C668. [CrossRef]

89. Uddin, M.; Okazaki, E.; Turza, S.; Yumiko, Y.; Tanaka, M.; Fukuda, Y. Non-destructive visible/NIR spectroscopy for differentiation of fresh and frozen-thawed fish. *J. Food Sci.* **2005**, *70*, C506–C510. [CrossRef]

90. Zhang, A.; Cheng, F. Identification of fresh shrimp and frozen-thawed shrimp by Vis/NIR spectroscopy. In Proceedings of the 2nd International Conference on Nutrition and Food Sciences IPCBEE, Singapore, 27–28 July 2013. [CrossRef]

91. Fasolato, L.; Balzan, S.; Riovanto, R.; Berzaghi, P.; Mirisola, M.; Ferlito, J.C.; Serva, L.; Benozzo, F.; Passera, R.; Tepedino, V.; Novelli, E. Comparison of visible and near-infrared reflectance spectroscopy to authenticate fresh and frozen-thawed swordfish (xiphias gladius L). *J. Aquat. Food Prod. Technol.* **2012**, *21*, 493–507. [CrossRef]

92. Fasolato, L.; Manfrin, A.; Corrain, C.; Perezzani, A.; Arcangeli, G.; Rosteghin, M.; Serva, L. Assessment of quality-parameters and authentication in sole (solea vulgaris) by NIRS (Near infrared reflectance spectroscopy). *Ind. Aliment.* **2008**, *47*, 355–361.

93. Reis, M.M.; Martínez, E.; Saitua, E.; Rodríguez, R.; Perez, I.; Olabarrieta, I. Non-invasive differentiation between fresh and frozen/thawed tuna fillets using near infrared spectroscopy (Vis-NIRS). *LWT– Food Sci. Technol. Int.* **2017**, *78*, 129–137. [CrossRef]

94. Ottavian, M.; Fasolato, L.; Facco, P.; Barolo, M. Foodstuff authentication from spectral data: Toward a species-independent discrimination between fresh and frozen-thawed fish samples. *J. Food Eng.* **2013**, *119*, 765–775. [CrossRef]

95. Karoui, R.; Lefur, B.; Grondin, C.; Thomas, E.; Demeulemester, C.; De Baerdemaeker, J.; Guillard, A.S. Mid-infrared spectroscopy as a new tool for the evaluation of fish freshness. *Int. J. Food Sci. Technol.* **2007**, *42*, 57–64. [CrossRef]

96. Sivertsen, A.H.; Kimiya, T.; Heia, K. Automatic freshness assessment of cod (Gadus morhua) fillets by Vis/Nir spectroscopy. *J. Food Eng.* **2011**, *103*, 317–323. [CrossRef]

97. Zhu, F.; Zhang, D.; He, Y.; Liu, F.; Sun, D.W. Application of Visible and Near Infrared Hyperspectral Imaging to Differentiate Between Fresh and Frozen-Thawed Fish Fillets. *Food Bioprocess Technol.* **2013**, *6*, 2931–2937. [CrossRef]

98. Aursand, M.; Veliyulin, E.; Standal, I.B.; Falch, E.; Aursand, I.G.; Erikson, U. Nuclear magnetic resonance. In *Fishery Products, Quality, Safety and Authenticity*; Rehbein, H., Oehlenschläger, J., Eds.; Wiley-Blackwell: Oxford, UK, 2009; Volume 1, pp. 252–266.

99. Nott, K.P.; Evans, S.D.; Hall, L.D. The effect of freeze-thawing on the magnetic resonance imaging parameters of cod and mackerel. *LWT- Food Sci. Technol. Int.* **1999**, *32*, 261–268. [CrossRef]

100. Howell, N.; Shavila, Y.; Grootveld, M.; Williams, S. High-resolution NMR and magnetic resonance imaging (MRI) studies on fresh and frozen cod (Gadus morhua) and haddock (Melanogrammus aeglefinus). *J. Sci. Food Agric.* **1996**, *72*, 49–56. [CrossRef]

101. Foucat, L.; Taylor, R.G.; Labas, R.; Renou, J.P. Characterization of frozen fish by NMR imaging and histology. *Am. Lab.* **2001**, *33*, 38–43.

102. Aursand, I.G.; Veliyulin, E.; Böcker, U.; Ofstad, R.; Rustad, T.; Erikson, U. Water and salt distribution in Atlantic salmon (Salmo salar) studied by low-field 1H NMR, 1H and 23Na MRI and light microscopy: Effects of raw material quality and brine salting. *J. Agric. Food Chem.* **2008**, *57*, 46–54. [CrossRef] [PubMed]

103. Mazzeo, M.F.; Siciliano, R.A. Proteomics for the authentication of fish species. *J. Proteomics* **2016**, *147*, 119–124. [CrossRef] [PubMed]
104. Li, L.; Boyd, C.E.; Sun, Z. Authentication of fishery and aquaculture products by multi-element and stable isotope analysis. *Food Chem.* **2016**, *194*, 1238–1244. [CrossRef] [PubMed]
105. Esteki, M.; Simal-Gandara, J.; Shahsavari, Z.; Zandbaaf, S.; Dashtaki, E.; Vander Heyden, Y. A review on the application of chromatographic methods, coupled to chemometrics, for food authentication (Chromatography-chemometrics in food authentication). *Food Control.* **2018**, *93*, 165–182. [CrossRef]
106. Primrose, S.; Woolfe, M.; Rollinson, S. Food forensics: Methods for determining the authenticity of foodstuffs. *Trends Food Sci. Technol.* **2010**, *21*, 582–590. [CrossRef]
107. Haynes, E.; Jimenez, E.; Pardo, M.A.; Helyar, S.J. The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. *Food Control.* **2019**, *101*, 134–143. [CrossRef]
108. Asensio, L.; González, I.; García, T.; Martín, R. Determination of food authenticity by enzyme-linked immunosorbent assay (ELISA). *Food Control.* **2018**, *19*, 1–8. [CrossRef]
109. Sobolev, A.; Mannina, L.; Proietti, N.; Carradori, S.; Daglia, M.; Giusti, A.M.; Antiochia, R.; Capitani, D. Untargeted NMR-based methodology in the study of fruit metabolites. *Molecules* **2015**, *20*, 4088–4108. [CrossRef] [PubMed]
110. Yang, H.; Irudayaraj, J.; Paradkar, M.M. Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy. *Food Chem.* **2005**, *93*, 25–32. [CrossRef]
111. Dos Santos, C.A.T.; Pascoa, R.N.; Lopes, J.A. A review on the application of vibrational spectroscopy in the wine industry: From soil to bottle. *TrAC Trends Anal. Chem.* **2017**, *88*, 100–118. [CrossRef]
112. Maione, C.; Barbosa, F., Jr.; Barbosa, R.M. Predicting the botanical and geographical origin of honey with multivariate data analysis and machine learning techniques: A review. *Comput. Electron. Agric.* **2019**, *157*, 436–446. [CrossRef]
113. Cozzolino, D. An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. *Food Res. Int.* **2014**, *60*, 262–265. [CrossRef]
114. Kamal, M.; Karoui, R. Analytical methods coupled with chemometric tools for determining the authenticity and detecting the adulteration of dairy products: A review. *Trends Food Sci. Technol.* **2015**, *46*, 27–48. [CrossRef]

molecules

MDPI

Review

Chemometrics Approaches in Forced Degradation Studies of Pharmaceutical Drugs

Benedito Roberto de Alvarenga Junior and Renato Lajarim Carneiro *

Department of Chemistry, Federal University of São Carlos, São Carlos 13565-905, Brazil; benedito.alvarenga@outlook.com

* Correspondence: renato.lajarim@ufscar.br; Tel.: +55-16-3351-9366

Academic Editor: Marcello Locatelli
Received: 11 September 2019; Accepted: 8 October 2019; Published: 22 October 2019

Abstract: Chemometrics is the chemistry field responsible for planning and extracting the maximum of information of experiments from chemical data using mathematical tools (linear algebra, statistics, and so on). Active pharmaceutical ingredients (APIs) can form impurities when exposed to excipients or environmental variables such as light, high temperatures, acidic or basic conditions, humidity, and oxidative environment. By considering that these impurities can affect the safety and efficacy of the drug product, it is necessary to know how these impurities are yielded and to establish the pathway of their formation. In this context, forced degradation studies of pharmaceutical drugs have been used for the characterization of physicochemical stability of APIs. These studies are also essential in the validation of analytical methodologies, in order to prove the selectivity of methods for the API and its impurities and to create strategies to avoid the formation of degradation products. This review aims to demonstrate how forced degradation studies have been actually performed and the applications of chemometric tools in related studies. Some papers are going to be discussed to exemplify the chemometric applications in forced degradation studies.

Keywords: forced degradation; degradation products; stress test; chemometrics

1. Chemometrics

The Swedish word "kemometri" appeared for the first time in 1971 by a combination between the terms chemistry and -metri. In 1972, the English homologous term chemometrics (chemo + metrics) was referred by Prof. Svante Wold that named his group as Forskningsgruppen för Kemometri (Research Group for Chemometrics) or Kemometrigruppen (Chemometrics Group), and in the next year, it was published the first article with the term kemometri [1,2]. The International Chemometrics Society explained the term "chemometrics" for the first time in 1974. International journals, in the 1980s, had special issues on chemometrics. In 1986–1987, the publishers Wiley and Elsevier created the chemometrics journals "The Journal of Chemometrics" and "Chemometrics and Intelligent Laboratory Systems," respectively [3].

The definition of chemometrics is intimately linked to what it is expected to gain from using it. This definition has presented some inconsistencies between authors over the years, once each one belongs to fields with different aims [4].

According to Pure and Applied Chemistry (IUPAC), the full definition of chemometrics, considering no preference of area, is the science of relating measurements performed on a chemical system or process to the state of the system through application of mathematical or statistical methods. IUPAC also highlights that, in chemometrics, the data are treated commonly in a multivariate approach, and although there are cases in theoretical chemistry that use the same mathematical and statistical techniques in some application, it should aim primarily to extract useful chemical information of measured data [5].

This definition evidences clearly the utilization of chemometrics in all stages of the chemical measurement process, from definition of optimal experimental conditions, data collection, and processing of data. Chemometrics has its roots in analytical chemistry [6], but it is totally interdisciplinary and has been applied in many different areas [7], such as food sciences [8–12], assessment of adulteration, geographical origin [13–15], metabolomics [16–18], engineering [19,20], forensics [21–25], pharmaceutical studies [26–30], cultural studies [31–33], environmental chemistry [34], etc. Chemometric tools are fundamental to solve real life problems [35].

In fact, when chemometric is applied appropriately with suitable interpretations, it enables to obtain a better data visualization even from experimental of poor quality (low resolution and high level of noise), making the relations between analytical signals and experimental parameters clearer [36]. The development of methods for analysis of degradation products is a hard work, time consuming, and an expensive task. In this context, chemometric tools are an alternative approach to carry out studies related to impurities in pharmaceutical drugs, contributing for acquiring relevant information from the system or turning the analytical method greener.

2. Degradation Products

The efficacy and safety of drugs are determined by toxicological and pharmacological profiles and adverse side effects due to the dosage and impurities [37–39]. According to the International Council for Harmonization and Technical Requirements for Pharmaceuticals for Humans Use (ICH), a drug impurity is any component that is not a chemical entity defined as an active pharmaceutical ingredient or excipient [40]. The impurities can be classified regarding their origin: inorganic impurities (reagents, ligands and catalysts, heavy metals or other residual metals, and inorganic salts), organic impurities (starting materials, by-products, intermediates, degradation products, reactants, ligands, and catalysts), and solvents (organic and inorganic liquids used in preparation of solutions or in the synthesis of a new drug substance). Therefore, any extra material present in the drug, even if it does not have pharmacological activity, is considered an impurity [39]. Although the term "impurity" is commonly assigned as synonymous of degradation products, it is worth highlighting that these compounds belong to a subgroup inside the impurity definition [41]. The United States Pharmacopoeia adopts the term "Related Compounds" for the main degradation products and impurities from synthesis.

The yielding of degradation products depends of several variables, chemical stability being the most important one. The degradation of APIs involves the formation or breaking of covalent bonds in chemical processes such as oxidation, reduction, thermolysis, and hydrolysis reactions. These processes can usually be accelerated when the drug is exposed to light, high temperatures, acidic or basic conditions, humidity, oxidative environment, incompatible excipients, and even due to its contact with packaging during its shelf-life [41].

2.1. The Generation of Degradation Products

Stability of API is a critical parameter in the development of a drug product, which should be considered in the formulation, analytical methods, package, storage, shelf life determination, safety, and toxicological studies [42,43].

The degradation of an API can result in the loss of effectiveness and can also lead to adverse effects due to degradation products [44]. Therefore, understating the processes that contribute to generation of degradation products is extremely important to create strategies aimed at the prevention and/or minimization of the API's degradation.

The oxidative degradation is one of the leading causes of drugs degradation, once it involves the removal of an electropositive atom, radical, electron, or the addiction of an electronegative atom or radical. The major part of API's oxidation occurs slowly due to the action of molecular oxygen, and some procedures used during manufacturing and storage are employed to stabilize the API in the product. For that, it is necessary to know the variables that increase the extension of oxidation. One form of preventing the oxidation process is to substitute oxygen inside pharmaceutical recipients by nitrogen or dioxide carbon. The contact of drug with metal ions, which can catalyze the oxidation, should be also avoided, as well as high storage temperatures [45].

Temperature is another variable that has significant influence on degradation and is often used in forced degradation studies. The same product can present different shelf lives depending on how and where it is stored. For example, countries in which equatorial climate predominates have higher average temperature than the ones with tropical climate, and this difference promotes different degradation conditions and, consequently, different shelf lives [46].

Several pharmaceutical drugs have low stability in aqueous medium and must be evaluated under hydrolysis conditions. First, to evaluate the hydrolysis of an API, it is necessary to perform tests in a wide range of pH (solution or suspension) once the hydrogen and hydroxyl ions are able to influence the degradation ratio [47–49]. Then, hydrolytic forced degradation studies are performed by submitting the API to acid, basic, and neutral conditions, in a fashion that the experimental variables have to be adapted if it is observed high degradation of API, in order to avoid the formation of secondary degradation products [48].

Photostability studies should also be performed to demonstrate the extension of reactions when the APIs are exposed to light. The photolytic reactions are caused when the drug absorbs the ultraviolet/visible (UV-Vis) light (wavelength 300 to 800 nm), which promote the molecule to an excited state and can increase its reactivity in some sites of the molecule. The UV-Vis radiation also can lead to cleavage of chemical bonds, yielding new molecules. The extension of photodegradation is dependent of the wavelength of the incident radiation and the absorptivity of the molecule. In other words, this process depends of the presence of specific functional groups [50].

Nonetheless, it is worth mentioning that even when an API is shown to be chemically stable in stress tests, the stress conditions can degrade this API when excipients are present.

2.2. Forced Degradation Studies

Since the release of the first guidelines, massive changes to the definition of quality in pharmaceutical drugs have taken place, and several countries are extending the requirements of regulatory agencies to generic drugs and already commercialized products [51]. Forced degradation studies, also called “stress tests,” have been used in the pharmaceutical industry for a long time [50], but the International Conference on Harmonization (ICH) only issued the formal request Q1A with a guideline “Stability Testing of New Drug Substance and Products” in 1993 [52]. In general terms, forced degradation studies are processes that involve the degradation of drugs under extreme conditions to accelerate the yielding of degradation products. The information obtained from these studies are usually used to determine the chemical stability, pathways of degradation, to identify the degradation products, conditions of storage, self-life, excipient compatibility, and also allow the development of selective analytical methods [52–54].

Today, the control of impurities has been established by ICH Q3A and Q3B guidelines, which are addressed for registration applications about the content and qualification of impurities classified as degradation products, which are observed during manufacturing or stability studies of the new drug product. Furthermore, the registration application should present a validated analytical procedure suitable for the detection and quantification of degradation products, which should include or evidence the method’s specificity for specified and unspecified degradation products according to ICH Q2A and Q2B guidelines for analytical validation. When the impurities are available in the validation method phase, the discriminatory capacity of drug and impurities is validated through spiking drug substance with levels of impurities. On the other hand, if impurity or degradation product standards are unavailable, the drug substance should be submitted to stress conditions (light, heat, humidity, acid/base hydrolysis, and oxidation). Therefore, in general, the forced degradation studies are performed in the developing stability-indicating method, and the method validation should take into account the chromatographic separation of the degradation products.

Several works in the literature deal with studies of forced degradation and stability as synonymous, but it is worth highlighting that there are some differences between them. Stability studies consist of submitting the pharmaceutical drug in milder conditions over a long period (months or years) and, besides determining some degradation products, allow the establishment of the product's shelf life. Forced degradation studies are often performed by exposing the API or the product in drastic conditions for some hours or days. These extreme conditions are able to provide, as a general rule, substantial degradation of the API, usually from 10 to 30%. The set of whole degradation products found in every degradation condition composes a "potential" degradation profile. If just few degradation products are found, the degradation profile is then denominated as "real degradation profile." The method to evaluate the degradation products should be selective and developed considering the occurrence of every degradation product [55].

The forced degradation studies are critical in the development of drug products and aims the following points:

- To obtain the potential degradation potential of an API or drug product;
- To discover the degradation mechanism, such as hydrolysis, thermolysis, oxidation, photolysis, etc.;
- To elucidate the molecular structure of degradation product;
- To solve problems regarded to the API stability;
- To identify the conditions where the API or the drug product are more susceptible to degradation in order to ensure the quality of the final product, bringing to pharmaceutical industry enough knowledge for development, packaging, manufacture, manipulation, and storage;
- To obtain more stable formulations;
- To develop analytical methods that can be used to quantify the API without interference of its degradation products and to quantify these degradation products [48,56,57].

The degradation products are commonly analyzed by high-performance liquid chromatography (HPLC) coupled with ultraviolet/visible (UV-Vis) and/or mass spectrometric (MS) detectors. UV-Vis detectors are able to provide only information related to chromophores groups, but they are excellent for quantification. MS detectors are not robust as UV-Vis detectors for quantification, but MS presents high sensitivity (traces level) and gives important data to characterize the degradation products through fragmentation profile, accurate mass (for detectors of High Resolution such as Q-ToF, Orbitrap, and Fourier-transform ion cyclotron resonance (FT-ICR)), as well as information about the origin of fragments using multiple stage (MS^n) and neutral loss scan. When more information is necessary to elucidate a chemical structure, the nuclear magnetic resonance (NMR) technique is required. NMR presents low sensitivity, but it is able to resolve conformational, structural, and optical isomers. All these techniques generate a great amount of data, and the manual data mining is very time and money consuming. In this context, chemometric tools can present a way to organize and pre-process data, optimize parameters of HPLC, MS, and NMR techniques, obtain the maximum knowledge about them, and clarify a lot of useful information [51,58,59].

2.3. Strategies to Select the Degradation Conditions

Forced degradation studies are performed in batches with solutions at different pHs, in the presence of hydrogen peroxide, UV-Vis radiation, metallic cations (Fe^{3+} and Cu^{2+}), and high temperatures [48].

Usually, the influence of pH is evaluated using 0.1 mol L^{-1} of HCl or NaOH [48]. The degradation by radiation is performed under UV-Vis light, which should not be lesser than 1.2 million of lux per hour and a power of 200 Wh m^{-2} [60]. For oxidant condition, the literature recommends using hydrogen peroxide (H_2O_2) in concentration from 0.1% to 3.0% at room temperature (25 °C). The evaluation of temperature is usually performed between 40 to 80 °C, but it could be higher for recalcitrant APIs. Other additional variables can be taken into consideration in the global stability studies of an API or the final product, such as humidity and microbiological stability [22,57,61,62].

According to ICH, in "Expert Committee on Specifications for Pharmaceutical Preparations" document, the recommended degradation should be between 10 to 30% of the API. This degradation range commonly allows for the evaluation of the main degradation products, avoiding the yielding of secondary degradation products [63]. In Brazil, the regulatory agency ANVISA recommends not less than 10% of degradation of API, and a technical justification is needed in the case where such degradation is not obtained [64].

It is worth highlighting that the cited conditions for forced degradation studies are just initial attempts, and the ideal condition could be more extreme or mild, depending of the chemical recalcitrance of the API. Table 1 summarizes degradation conditions of some papers that performed forced degradation studies.

Table 1. Degradation conditions for pharmaceutical drugs in forced degradation studies.

API: Year	Acid	Base	Neutral	Thermolysis	Oxidation	Photolysis
Zidovudine: 2017 [65]	2 M HCl	2 M NaOH	-	Acid/base at 80 °C for 72 h	10% H_2O_2 at room temperature for 10 h	1.2 × 10^6 lx × h of fluorescent light and 200 W h/m^2 UV light
Toloxatone: 2018 [66]	1 M HCl	0.01 M NaOH	H_2O	All hydrolysis at 80 °C for 2 h	0.01% H_2O_2 at room temperature for 2 h	2700 kJ/m^2/h of UV-VIS and UVC 7.5 W/m^2
Amlodipine: 2015 [67]	1 M HCl at 80 °C for 30 min	1 M NaOH at 80 °C for 1 h	H_2O at 80 °C for 2 h	50 °C for 48 h	15% H_2O_2 at room temperature for 48 h	1.2 × 10^6 lx × h of fluorescent light and 200 Wh/m^2 UV-A light for 14 days
Acebutolol: 2018 [68]	1 M HCl	2 M HCl	H_2O	All hydrolysis at 80 °C	3% H_2O_2 at 80 °C	Not less than 1.2 × 10^6 lx × h and ultraviolet energy of not less than 200 W h/m^2
Stevioside: 2018 [69]	0.1 M HCl/0.1 M H_3PO_4	0.1 M NaOH	H_2O	All hydrolysis at 80 °C for 8 h	10% H_2O_2 at 25 °C for 72 h	UV_{254nm} lamp for 48 h
Pentoxifylline: 2013 [70]	2 M HCl at 70 °C for 4 h	2 M NaOH at 70 °C for 4 h	H_2O at 70 °C for 4 h	Dry heat under at 105 °C for 4 h	30% H_2O_2 at 70 °C for 4 h	Sunlight for 8 h
Leflunomide: 2015 [71]	0.1–5 M at 85 °C for 8 h	0.1 M NaOH at 85 °C for 8 h	H_2O at 85 °C for 8 h	50 °C for 30 days	30% H_2O_2 at room temperature for 24 h	UV and white light for 14 days
Actarit: 2014 [72]	0.1 M HCl at 70 °C for 24 h	0.1 M NaOH at 70 °C for 24 h	H_2O at 70 °C for 14 days	Dry heat at 70 °C for 14 days	3% H_2O_2 for 14 days	UV light
Nicardipine: 2014 [73]	1 M HCl at 60 °C for 1 h	0.1–0.5 M NaOH at 50–80 °C for 1 h	-	-	5% H_2O_2 at 30–50 °C for 1 h	$UV_{254–365nm}$ light at room temperature
Clopidogrel bisulfate: 2010 [74]	1 M HCl	1 M NaOH	-	All hydrolysis at 80 °C for 1 h	5% H_2O_2	-
Biapenem: 2009 [75]	pH from 2.5 to 7.5 at 80 °C for 40 min			From room temperature to 100 °C in pH 3.5	-	-
Irbesartan: 2010 [76]	1 M HCl at 80 °C for 24 h	2 M NaOH at 80 °C for 48 h	H_2O at 80 °C for 48 h	50 °C	30% H_2O_2 at room temperature for 2 days	8500 lx fluorescent and 0.05 W/m^2 UV light

2.4. Acceptable Limits of Impurities

After obtaining the degradation profile, a critical analysis should be performed to verify the purity of the chromatographic band of the API and to evaluate the variables that can promote degradation of the API. The degradation products are analyzed according to their amount in relation to the API in the final product, after the regular stability time (without any stress condition). The evaluation considers the maximum amount of API administered per day, and the limit of degradation products are expressed as a percentage (or mass) relative to the API. The amount of degradation products defines if it is necessary to perform notification, identification, or qualification [40,57,77]. Table 2 shows the acceptance criterion used by ICH, FDA, and ANVISA for the amount of impurities found in relation of a daily administrated API. The acceptance criteria have the following meaning:

- Reporting threshold: A limit of impurity that is not necessary to be reported.
- Identification threshold: A limit of impurity does not need to be structurally identified.
- Qualification threshold: The maximum amount of impurity that is not necessary to be qualified. Being "qualified" is the process of acquisition and evaluation of data that establishes biological security of an impurity or a degradation profile at the specified levels [40].

Table 2. Thresholds for degradation products.

	Maximum Daily Dose	Threshold
Reporting Threshold	≤1 g >1 g	0.1% 0.05%
Identification Threshold	<1 mg 1 mg–10 mg >10 mg–2 g >2 g	1.0% or 5 μg TDI, whichever is lower 0.5% or 20 μg TDI, whichever is lower 0.2% or 2 mg TDI, whichever is lower 0.10%
Qualification Threshold	<10 mg 10 mg–100 mg >100 mg–2 g >2 g	1.0% or 50 μg TDI, whichever is lower 0.5% or 200 μg TDI, whichever is lower 0.2% or 3 mg TDI, whichever is lower 0.15%

3. Applications of Chemometric Tools in Forced Degradation Studies

3.1. Design of Experiment (DoE)

In every area is important to know how variables act on the system. In general, processes aim to enhance the quality of the final product, taking into account the minimization of cost and time. To achieve these goals, it is necessary to perform the optimization of variables of the system to gain knowledge about the behavior of variables in order to determine the influence of each variable [78,79]. The optimization of variables in a system is more commonly performed using one-variable-at-a-time approach (OVAT), where one variable, or also called factor, is changed at a time, causing a change in the monitored response. However, this univariate approach does not consider the interactions between variables, and therefore, it does not ensure the discovery of the optimum point in an optimization process [80]. The design of experiments arises as an alternative multivariate approach for studying the behavior of a system [81]. In this approach, the factors are simultaneously evaluated, and the experiments are performed in an organized way in order to acquire information about all the system performing a minimum number of experiments [82,83].

Some terms in DoE must to be clear for better understanding, as variables, levels, and responses. Variables or factors are independent experimental inputs capable of changing the responses of the system. Such factors are temperature, pH, irradiation time, reaction time, concentration of reactants, and so on. It is worth reiterating that variables can be changed independently of each other, but the response is dependent of synergism between them [84].

Levels are different values that a variable can assume within experimental domain. The variable temperature in an optimization process, for example, can be studied at three levels: at 30, 50 and 70 °C.

Responses or independent variables are the monitored parameters. Typical responses are cost, time of analysis, resolution between chromatographic peaks, percentage of API degradation, etc.

The values studied for each variable are coded in levels as high (+1), central (0), low (−1), and other levels, which depend on the design. This codification normalizes the independent variables, avoiding any wrong interpretation of data. The processes involved in DoE allow it to fit the empirical data to a function, creating a linear or quadratic model and considering the interactions between variables of the system [85]. Figure 1 shows the experimental domain of the most common experimental designs for screening and optimization steps.

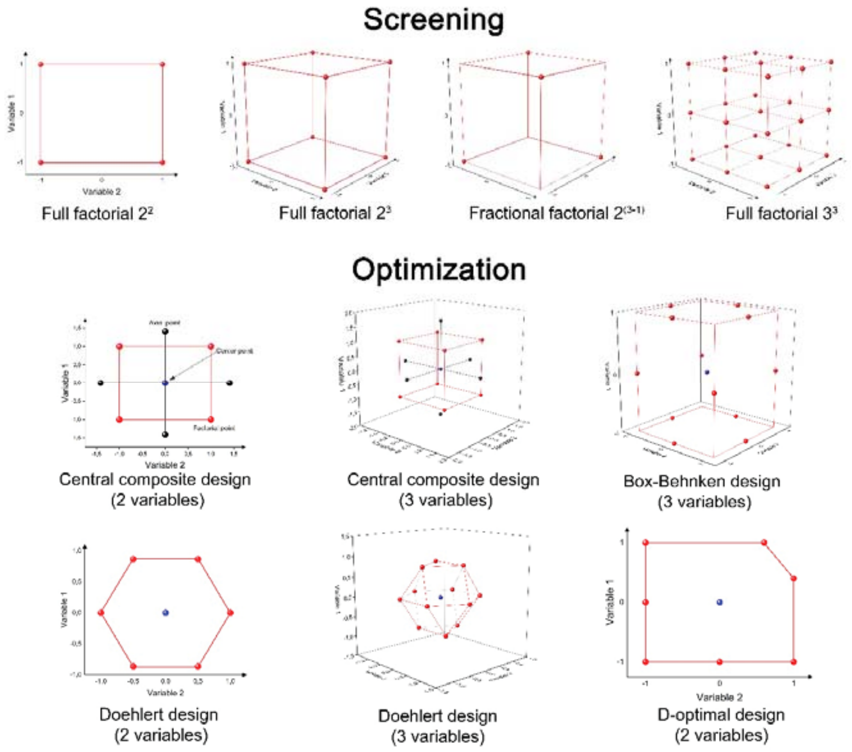


Figure 1. Experimental domain of the most common experimental designs.

In sum, the DoE presents the following advantages:

- Determining how many experiments are necessary to achieve the goal;
- Reducing the number of experiments;
- Observing the synergic and antagonist interactions between variables;
- Allowing for the possibility to create mathematical models and surface response to describe the behavior of the variables and to predict the system's response within an experimental domain;
- Decreasing the time, costs, and generation of lesser amounts of chemical waste, which contributes for the green chemistry principles [79].

In the context of forced degradation studies, the DoE has been mainly used for the development and optimization of chromatographic methods and for multivariate evaluation of stress conditions.

The use of DoE in the development and optimization of chromatographic conditions is not exclusive for forced degradation studies; instead, its application has spread to several fields that use chromatography as a tool [86–88]. Krishna et al. [89] performed forced degradation studies of eberconazole nitrate (EBZ) submitting it to hydrolytic (acid, basic, and neutral), thermal, oxidative, and photolytic degradation. In this work, a full factorial 3^3 design was used to identify the best conditions of the mobile phase for drug analysis. As is already well known in chromatography, the organic modifier in the mobile phase (methanol in this case), pH (10 mM potassium dihydrogen orthophosphate), and ion pair agent (tetra butyl ammonium hydroxide, TBAH) are important variables and alter the capacity factor (k) of the mobile phase. These variables were evaluated in three levels (−1, 0, and +1) following a full factorial design with 27 experiments (3^3 Full Factorial). Table 3 presents the real value of variables, and Table 4 shows the 27 different experiments.

Table 3. Real and coded values of variables considered in design of experiment.

Variable	Level (−1)	Level (0)	Level (+1)
TBHAH (mM)	5	7.5	10
pH	2.6	2.9	3.2
Organic phase (*v*/*v*)	20	25	30

Table 4. Conditions of experiments performed in full factorial 3^3 design.

Experiment	x_1	x_2	x_3	Experiment	x_1	x_2	x_3	Experiment	x_1	x_2	x_3
1	−1	−1	−1	**10**	−1	−1	0	**19**	−1	−1	1
2	0	−1	−1	**11**	0	−1	0	**20**	0	−1	1
3	1	−1	−1	**12**	1	−1	0	**21**	1	−1	1
4	−1	0	−1	**13**	−1	0	0	**22**	−1	0	1
5	0	0	−1	**14**	0	0	0	**23**	0	0	1
6	1	0	−1	**15**	1	0	0	**24**	1	0	1
7	−1	1	−1		−1	1	0	**25**	−1	1	1
8	0	1	−1	**17**	0	1	0	**26**	0	1	1
9	1	1	−1	**18**	1	1	0	**27**	1	1	1

The ranges studied in design were selected according to previous studies and considered the physicochemical properties of EZB. Other chromatographic parameters such as column dimensions, flow rate, injection volume, wavelength for detection, as well as the procedure performed in each degradation condition, can be found in reference [89].

As a result, a Pareto chart of standardized effects showed the quantification of each variable on the capacity factor, where organic phase and TBAH presented the higher influence on the response. Both linear and quadratic regressions showed no significance for pH inside its range of variation. The results of experimental design also allowed the authors to create contour plots, and they emphasized the usefulness of studying the interaction effects of variables on capacity factor. It was observed through contour plots that, by increasing concentration of TBAH, the capacity factor of EBZ was increased, and the same behavior occurred when the organic modifier decreased. Furthermore, pH did not affect the capacity factor in the investigated experimental domain. At the end, the optimum conditions (pH 2.8, 10 mM TBAH, and methanol 25% (*v*/*v*)) made it possible to find a capacity factor equal to 2.06.

Table 5 shows some papers that used the experiment design to optimize the chromatographic conditions to analyze the degradation products yielded in forced degradation studies.

Table 5. Design of experiments used in some papers to optimize chromatographic conditions for analyses of degradation products.

API	Design	Ref
Teriflunomide	Full factorial 3^3	[90]
Simvastatin	Plackett Burman/Box-Behnken	[91]
Linagliptin	Full factorial	[92]
Ticagrelor	Fractional Factorial Resolution V/Central composite	[93]
Imatinib mesylate	Box Behnken	[94]
Fusidic acid	Taguchi/Central Composite	[95]
Cloxacillin	Plackett Burman	[96]
Vilazodone hydrochloride	Central composite experimental	[97]
Darifenacin hydrobromide	Central composite	[98]
Edaravone	Placket Burman/Box Behnken	[99]
Sofosbuvir and Ledipasvir	Box Behnken	[100]

In the papers presented in Table 5, the DoEs were used to evaluate the chromatographic parameters in order to obtain the best chromatographic method. The meaning of the best chromatographic method depends of the intention of the analyst—better resolution for the API, higher number of peaks in order to detect all degradation compounds, cost-and-time saving methods, etc.

Another purpose for forced degradation studies found by Sonawane and Gide [101] was the application of experimental design for the optimization of forced degradation of luliconazole (LCZ), 4-(2,4-dichlorophenyl)-1,3-dithiolan-2-ylidene-1-imidazolylacetonitrile), which is recommended for the treatment of fungal infections. The LCZ was submitted to acidic (HCl), alkaline (NaOH), oxidative (H_2O_2), thermolytic (under reflux), and photolytic (direct sunlight) stress conditions, and a full factorial design was chosen to identify the conditions to obtain a degradation of this API between 10 and 20%. The 2^3 factorial design for acid and alkaline conditions took into account the variables concentration of the degradant agent (x_1), temperature (x_2), and time of exposure (x_3) to achieve the desired degradation. The variable temperature was not included in oxidative degradation, and the design became a 2^2 factorial design. The same design was performed to dry heat and wet heat degradation, but including the variable temperature and discarding the variable concentration. For photolytic degradation, LCZ powder was exposed to direct sunlight for 48 h and compared with control in dark, but DoE was not applied. The level of the variables for each stress condition is presented at Table 6. The 2^3 factorial design was performed in a total of eight experiments, and the 2^2 factorial in a total of four experiments for each degradation (oxidative, dry heat, and wet heat) by design. Table 7 shows the experiments and the obtained results by liquid chromatography.

Table 6. Real values of the variables used in the design of experiments.

Variable	**High Level (+1)**					**Low Level (−1)**				
	Acid	**Basic**	**Oxid.**	**Dry Heat**	**Wet Heat**	**Acid**	**Basic**	**Oxid.**	**Dry Heat**	**Wet Heat**
Conc. (x_1)/mol×L^{-1}	1	0.1	30%	-	-	0.1	0.01	3%	-	-
Time (x_2)/min	75	30	24 h	360	120	15	10	2h	30	30
Temperature (x_3)/°C	100	100	-	200	100	60	60	-	50	60

Table 7. Design of experiments with coded values and % of degradation of active pharmaceutical ingredient (API) for acid, basic, and oxidative conditions.

2^3 Full Factorial Design						2^2 Full Factorial Design			
Exp.	X_1	X_2	X_3	**Acid Condition**	**Basic Condition**	**Exp.**	X_1	X_2	**Oxidative Condition**
1	−1	−1	−1	0%	0%	1	−1	−1	0%
2	+1	−1	−1	4%	3%	2	−1	+1	48%
3	−1	+1	−1	10%	8%	3	+1	−1	51%
4	+1	+1	−1	23%	11%	4	+1	+1	100%
5	−1	−1	+1	8%	19%				
6	+1	−1	+1	32%	26%				
7	−1	+1	+1	21%	38%				
8	+1	+1	+1	41%	43%				

The dry and wet heat degradation did not present any degradation of luliconazole, but photolytic degradation obtained 8%. Concerning acid, alkali and oxidative conditions, the degradation ranges were 0–41%, 0–43%, and 0–100%, respectively. Multivariate regressions were performed on the results for each degradation (acid, alkali, and oxidative) in order to obtain the regression models (equations) for the studied experimental domain. These regression models are used to predict suitable conditions to achieve the desired percentage of degradation. These conditions provided degradation of 11%, therefore, a relative error equal to 9%. More details about the equations in each degradation condition as well as surface response created to better visualization of the results can be found in the reference [101]. The DoE in this work allowed the authors to gain knowledge about stability of LCZ, presenting the degradation condition where LCZ is more susceptible to undergo degradation and indicating the variables that present higher influence on the degradation of LCZ. Finally, the chemometrics tools aid to predict the values of variables to obtain the desired degradation.

Another example was presented by Kurmi et al. [102]. that used DoE to develop the stability-indicating method and also found the stress conditions for forced degradation of furosemide in the range of 20–30%.

Despite the fact that DoE is a very interesting tool to find the most suitable conditions in the degradation studies and avoiding the generation of secondary degradation products, there are few papers presenting such approach.

3.2. *About Fusion QbD®*

As mentioned previously, forced degradation studies are performed in the development stability-indicating method phase. DoE is extremely useful to build a set of screening, optimization and robustness experiments. In this context, some HPLC method development software platforms are commercially available to automatically perform the experimental design. This software, such as Fusion QbD, uses concepts of experimental design and creates a sequence of experiments considering all relevant chromatographic parameters. It is possible to build, for example, a set of screening experiments considering more than one type of chromatography columns, multi-solvents, and other chromatographic variables. After the creation of a set of methods, guided by the DoE principles, and after running the sequence of experiments, the software generates mathematical models and makes predictions to find the better chromatographic method. As Fusion QbD is integrated with the chromatography system, all functions of HPLC are explored, and it allows users to reach maximum efficiency and speed in the method developing process [103]. Others specialized software is also used to create basic designs, such as Origin [104], Matlab [105], Minitab [106], Design-Expert [107], and Statistica [108].

3.3. Principal Component Analysis (PCA)

Principal component analysis (PCA) is one of the most used chemometric tools for data exploration through the reduction of a system's dimensionality [23,109,110]. This technique allows the user to establish the numerical adjustment of a linear model for describing the central relationships among process variables [111]. The PCA aims mainly to extract the most useful information from data. Besides, this chemometric tool helps simplify the description of the data for the analysis of variables [112].

The use of PCA enables the user to represent objects with new variables that are linear combinations of the original variables. These linear combinations, denominated principal components (PCs), are calculated considering directions of maximum variance, in a fashion that they may also be perpendicular to each other [23]. The first PC describes the maximum variance of the sample. The second PC describes the most considerable variability that the first one was not able to describe. The directions of the most dispersed samples are generally described in the first PC, since it corresponds to the vector with more information about the linear combinations of the original variables [113]. Figure 2 presents a graphical representation of PCA, where the axes are changed in order to maximize the explained variance using a smaller number of dimensions.

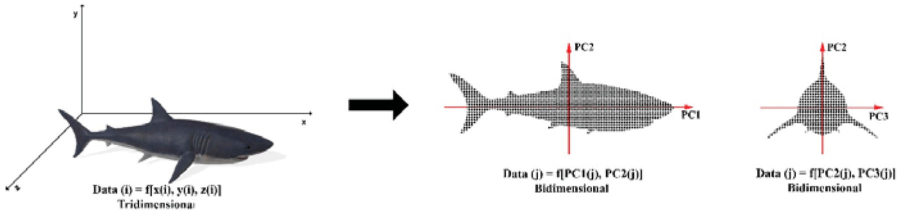

y
x
z
Data (i) = f[x(i), y(i), z(i)]
Tridimensional
PC2
PC1
Data (j) = f[PC1(j), PC2(j)]
Bidimensional
PC2
PC3
Data (j) = f[PC2(j), PC3(j)]
Bidimensional

Figure 2. Representation of principal component analysis (PCA). Original data at left side, PC1 × PC2 in the middle and PC2 × PC3 at right side.

In the literature, three papers were found involving PCA associated with degradation products of pharmaceutical drugs. Two of them will be discussed in the next paragraphs, and the other one will be discussed later, in the MCR-ALS context.

Tôrres et al. [114] performed accelerated degradation studies of captopril and applied Multivariate Statistical Process Control (MSPC) for monitoring and identifying any changes in samples in order to guarantee the product quality. The details of all procedure data treatment can be found in reference [114]. The captopril stability was evaluated leaving 24 blisters of tablets of the same batch in a climatic chamber at 40 ± 2 °C and 75 ± 5% of relative humidity. One blister per week was analyzed by liquid chromatography, for six months, totalizing 24 chromatograms. In order to build the process control chart, a sample set of Captopril was used under normal operation conditions in the calibration (training stage), and in the validation stage, samples were used under normal operation conditions, as were samples presenting expired shelf life. Hotelling's T^2 statistic and Square Prediction Error (SPE) were used for sample monitoring. PCA is a useful tool in the Hotelling's T^2 statistic, since it reduces the number of variables to be monitored, changing the original variables by the scores in the PCA, without significant information loss from dataset. The PCA along with the multivariate control charts contributes to identify possible failures and changes early in the process, making this method useful to ensure the quality control of product [114]. The same authors also performed a similar work using the mid (MIR) and near (NIR) infrared techniques [115].

Skibinski et al. [66] performed forced degradation of toloxatone, which is a pharmaceutical drug used as an antidepressant. These studies were carried out in basic (0.01 M NaOH), acidic (1 M HCl), neutral (water), photo UV-Vis, photo UVC, and oxidative (0.01% H_2O_2) degradation conditions. The samples (including the control solution) were evaluated in a LCMS (ToF) totalizing 21 chromatographic profiles. The stress conditions provided eight unique degradation products of toloxatone [66].

After aligning of chromatographic profiles, PCA analysis showed a visible grouping of the stressed samples. The author noticed that stressed basic samples gave rise to a separated cluster from other stressed samples in the scores analysis obtained from PCA, while neutral and acidic samples were close to the control samples. On the other hand, it was possible to separate in groups the samples carried out under photo UV-VIS, photo UVC, and oxidation conditions. The first three components of PCA model were able to explain almost 71% of the total variance. This work shows that PCA analysis can be used as a tool to characterize the chromatographic profiles.

3.4. Partial Least Squares (PLS)

Partial least squares (PLS) regression is a multivariate regression technique, the most important one in the chemometrics. It is used to stablish quantitative relationships between a vector of information (UV-Vis, Raman, NIR, MID-IR, NMR spectra or chromatogram, diffractogram, etc.) and properties to be quantified (concentration of an analyte, crystalline phase of API, etc.) [116–119].

As example, the concentrations of an analyte in calibration samples are organized in a vector y, and the chemical data (spectra) are organized in a matrix X. In the classic multivariate regression, the regression coefficient **b** is found by $\mathbf{b} = \mathbf{y} \times X^{+}$, where X^{+} is the pseudoinverse of X. The regression equation (model) can be written in the matrix form as $\mathbf{y} = \mathbf{b} \times X$. However, there is some issues related to the use of classical multivariate regression, such as the need of high number of samples and the problem of the correlation among the variables in the matrix X. Then, in a similar way as PCA, PLS calculations simultaneously decomposes X and **y** in order to maximize the correlation among the scores of X and **y**. After defining coefficients **b**, it can be applied to determine the concentration in external samples [120].

Some algorithms have been proposed to perform PLS, and the most common are PLS1 and PLS2, for one response and for multiple responses, respectively. Although PLS2 is used for multiple responses, it is recommended only in the cases where there is high correlation among the responses [121].

Recently, Sayed et al. [122] developed a stability-indicating method using PLS to determine mometasone furoate (MF) pure or in pharmaceutical formulation in the presence of its degradation products. The forced degradation was performed only in basic conditions once other previous works have demonstrated its susceptibility in undergoing alkaline hydrolysis. The multilevel multifactor experimental design was applied to prepare mixtures of calibration set constituted by 14 samples, which were scanned over the range of 220–350 nm. The UV spectra of 11 different mixtures of MF and its degradation products were used to predict the concentration of MF. The PLS model applied in the determination of MF presented good results, obtaining in calibration set mean recovery of 100.2% and RMSEC 0.002% meanwhile validation set presented mean recovery of 97.24% and RMSEP 0.04%. The recoveries in pharmaceutical samples were also satisfactory (98.47–102.66%), demonstrating no interference from excipients or alkaline degradation products in the quantification and the power of PLS method for quantification of MF [122]. Besides, in this same work, a new TLC densitometric method and the chemometric tools CLS and PCR were found, which were applied to develop quantification models for the MF in pharmaceutical samples.

Attia et al. [123] also developed spectrometric methods for determination of cefoxitin-sodium in the presence of its alkaline degradation product using different chemometric tools. PLS was applied to quantify cefoxitin-sodium in pharmaceutical sample. To obtain degradation product, the basic forced degradation was performed using NaOH 0.1 M for 10 min, which was neutralized with HCl 0.1 M. More details about the procedure to prepare the working solution are in reference [123]. The PLS model was built considering 13 mixtures denominated calibration set and 12 mixtures as a validation set obtained through experimental design. The number of factors was optimized through cross-validation method, as performed in reference [122]. The genetic algorithm (GA) was coupled with PLS to improve the prediction capability of models eliminating variables without information. In fact, the efficiency of the calibration of GA-PLS was better than only PLS, given lower RMSEC and RMSEP values for GA-PLS. The analysis of cefoxitin-sodium in presence of degradation products and in the pharmaceutical sample presented mean recovery of 100.54% and 99.86 ± 1.347%, respectively, using GA-PLS. The proposed method presented no significant difference compared to the standard method. Different chemometric tools were proposed and all of them showed a solvent reduction and sample consumption, making the methods greener. Table 8 present papers found in the literature that use in some moment the PLS tool in forced degradation studies of pharmaceutical products.

Table 8. Works involving forced degradation studies and the partial least squares (PLS) tool.

Author	API	Forced Degradation Condition	Chemometric Tool	Year	Ref.
Attia et al.	Cefprozil	Basic hydrolysis	PLS; SRACLS	2016	[124]
Alamein et al.	Pimozide	Acid and basic hydrolysis	CLS; PCR; PLS	2015	[125]
Hegazy et al.	Linezolid	Acid and basic hydrolysis; oxidative	PLS; PCR; Parafac; N-PLS	2014	[126]
Hegazy et al.	Imidapril hydrochloride	Basic hydrolysis; oxidative	PCR; PLS	2014	[127]
Souza et al.	Captopril	Thermolysis	PLS	2012	[128]
Abou Al Alamein	Zafirlukast	Basic hydrolysis	PLS	2012	[129]
Naguib	Bisacodyl	Acid hydrolysis	PLSR; SRACLS	2011	[130]
Abdelwahab	Atenolol; Chlorthalidone	Acid and basic hydrolysis	PCR; PLS	2010	[131]
Wagieh et al.	Oxybutynin hydrochloride	Basic hydrolysis	PCR; PLS	2010	[132]
Moneeb	Rabeprazole sodium	Acid hydrolysis	CLS; PCR; PLS	2008	[133]
S Fayed et al.	Cilostazol	Acid hydrolysis	PLS; CRACLS	2007	[134]
Ragno et al.	Lacidipine	Photodegradation	PLS; PCR; MLRA	2006	[135]
Shehata et al.	Rofecoxib	Basic hydrolysis; photodegradation	PLS; CRACLS	2004	[136]

3.5. Multivariate Curve Resolution (MCR)

Multivariate curve resolution (MCR) has been widely used to analyze several types of data in different application fields [137–139]. MCR constitutes a bilinear model based on the classical least squares (CLS) that decomposes data matrix into two submatrices, which have chemical information of the compounds involved in the system [137,139–141].

This approach is also known to be spectral unmixing tool once it allows mathematically solving analyte signals of a complex mixture where they are overlapped in one or more dimensions of data, as chromatograms and spectra of analyte in the presence of interferents in analysis without resolution. MCR aims to differentiate the individual contributions of components of a mixture providing the pure signals (spectra) and the proportions of analytes through concentration profile [138,139,142]. MCR comes from the Beer's law, where concentration is proportional to the absorbance. In this way, a spectral data set can be deconvoluted in the pure spectra from the analytes and their relative concentration. The general equation for MCR is $X = C \times S^t$, where the spectral matrix X is deconvoluted in the concentration matrix and the pure spectra matrix.

Most papers related to forced degradation studies and MCR-ALS aimed for the evaluation of photodegradation. Except for basic hydrolysis condition, other degradation conditions were not found in the literature.

Marín-García et al. [143] investigated photodegradation of tamoxifen in aqueous medium using Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS). The photodegradation experiments were conducted at 35 °C in a cabinet equipped with light at two different irradiation power conditions (400 and 765 W/m^2) according to ICH requirements. To monitor the photodegradation of tamoxifen, the UV-VIS spectra were collected from 0 to 160 min for irradiation power 400 W/m^2, and from 0 to 120 min for 765 W/m^2. The UV spectra allowed to obtain the evolution of the photodegradation process. MCR-ALS analysis of the UV data allowed to observe the estimation of the kinect profiles for the possible presence of at least four species, three of them being degradation products. Besides, it was possible to obtain the relative concentration of each specie along time.

During photodegradation some molecules cannot be detected by UV-Vis due to the loss of chromophore groups. The authors overcame this situation using a LC-DAD-MS technique to obtain deeper knowledge about species formed in photodegradation. In this case, MCR-ALS analysis provides the C and S matrixes that contain, respectively, the elution profile and pure UV-VIS or MS spectra for each substance. These matrixes showed a new component, which represents a fourth degradation product. This new specie was not observed in the UV-VIS monitoring, it rises during photodegradation but disappears at the end of the process. Furthermore, the authors elucidated the degradation product structures. This work shows MCR-ALS's ability to monitor and solve mixtures of degradation products formed during photodegradation process [143].

Another work reported in the literature was conducted by Feng et. al. [144], which investigated the basic degradation for paracetamol using two-way dimensional UV-Vis associated to MCR-ALS. Forced degradation was performed using a quartz cell where paracetamol and NaOH solutions were added, and the UV-VIS spectra were collected from 1 s to 24 h. Initially, a PCA was applied on UV-VIS data, and it suggested the existence of four components. Later, the concentration profiles were obtained from evolving factor analysis (EFA), and it confirmed the number of chemical components involved in degradation reaction. In the MCR-ALS deconvolution, it was applied to the constraints non-negativity for spectral and concentration profiles and unimodality for the concentration profile. Through the concentration profile and spectra profile plots, it was possible to perform a critical analysis of the formation and consumption of the species during alkaline degradation. It was possible to observe that there were a reactant, a degradation product, and two intermediates. The authors compared the results with HPLC analysis, which proved the existence of two intermediates, and the concentration profile were in agreement with the one recovered by MCR-ALS using UV-Vis. Besides, the authors also proposed a degradation pathway in alkaline media. The use of MCR-ALS in forced degradation studies allowed to verify the drug stability and kinect of degradation of paracetamol [144]. Other papers regarding forced degradation studies and MCR-ALS are presented in Table 9.

Table 9. Works involving forced degradation studies and the Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) tool.

Author	API	Forced Degradation Condition	Chemometric Tool	Year	Ref.
Gómez-Canela	5-Fluorouracil	Photodegradation	MCR-ALS	2017	[145]
Bērziņš et al.	Furazidin	Basic hydrolysis	HS-MCR-ALS	2016	[146]
Luca et al.	Amiloride	Photodegradation	MCR-ALS	2012	[147]
Sílvia Mas et al.	ketoprofen	Photodegradation	MCR-ALS; HSMCR	2011	[148]
Luca et al.	Nitrofurazone	Photodegradation	HS-MCR-ALS	2010	[149]
Javidnia et al.	Nitrendipine and felodipine	Photodegradation	MCR	2008	[150]
Shamsipur et al.	Nifedipine	Photodegradation	MCR	2003	[151]

3.6. Artificial Neural Network (ANN)

Artificial neural networks (ANNs) are powerful chemometric tools based on artificial intelligence. They can model nonlinear data through learning processes in a similar way to the human brain [36,152]. ANN models are able to map the input data in a set of appropriate outputs following a "learning by examples." In other words, the structure of data is learned through training algorithms [153].

To the best of our knowledge, two works regarding to forced degradation studies and artificial neural network are reported in the literature, and only one of them uses ANNs as the main tool [123,154].

Golubović et al. [154] used ANNs to develop quantitative structure-retention relationships (QSRRs) model to optimize isocratic RP-HPLC method of candesartan cilexetil in the presence of seven degradation products obtained from acid, alkaline, neutral hydrolysis, photolysis, and oxidation conditions. QSRRs is able to relate chromatographic retention parameters and molecular structure, and it becomes a valuable tool to the prediction of chromatographic behavior and separation of complex mixtures.

Initially, to investigate the variables that could influence the chromatographic behavior, a 2^{5-1} fractional factorial design was performed. The following variables were included in the design: percentage of acetonitrile in the mobile phase, buffer pH and ionic strength, temperature of the column, and flow rate of the mobile phase. All variables showed to be significant and, therefore, were considered as inputs in the ANN modeling, except flow rate, which was maintained as a constant.

The molecular structure is an essential variable in QSRR model and is encoded by descriptors. Roughly, molecular descriptors are obtained by logic and mathematical procedures that transform chemical information in a useful number of some standardized experiments. The selection of molecular descriptors was based on intermolecular interactions suggested by theory of liquid chromatography. In the ANN modeling it were included the descriptors which present low correlation between them, such as polarizability, H-donor sites, H-acceptor sites, and octanol/water distribution coefficient.

It was used a multi-layer feedforward, the most common ANNs, constituted by one input layer (descriptors and significant chromatographic variables), number of hidden neurons connected to both input and output neurons (retention factor). In the network training stage, the overall agreement between computed and target output for a set training is maximized. In order to avoid overfitting, the predictive power of network was evaluated using a validation set. Both training and validation sets were defined through a Box-Behnken design, varying from −1 to +1 level. A total of 344 cases for ANN optimization were obtained, which were divided into 280 cases for the training set, 32 for external validation, and 32 to validation set. For training, validation, and external validation data sets, coefficients of determination (R^2) were obtained between experimental and predicted retention factor (K_{exp} and K_{ANN} respectively) equal to 0.9993, 0.9969 and 0.9956, respectively. Therefore, high R^2 and low RSME values demonstrate an excellent predictive ability of model and non-occurrence of overfitting during the training process.

This kind of mathematical model is an important tool in forced degradation studies since degradation products derive from the API and, therefore, are chemically similar. The creation of models able to predict the behavior of active substance and all degradation products contribute to defining the optimal chromatographic conditions during the optimization process [154].

4. Conclusions

Chemometric tools can bring considerable gains in forced degradation studies. DoE is the most used chemometric tool in such studies, especially in the development of suitable chromatographic methods to monitor the API. However, the application of DoE directly in stress experiments is also promising, as it is possible to quantify the individual effect of stress variables as well as the synergy between them, simulating what may occur in real life. The other widely used tool is PLS, since its use allows the quantification of the API directly in UV-Vis spectrophotometry analyzes, since it performs multivariate quantification, which makes possible quantification of species without resolution.

The PCA technique is not applied in these studies since it is an exploratory method, and its application is more related to process monitoring and classification methods for raw material identification.

The other tools, despite being very useful in such studies, are more complex, and their application is limited for non-chemometricians.

Author Contributions: Both authors contribute equally to produce this review.

Funding: This research was funded by the São Paulo Research Foundation (FAPESP), grant number 2017/13095-0.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kiralj, R.; Ferreira, M.M.C. The past, present, and future of chemometrics worldwide: Some etymological, linguistic, and bibliometric investigations. *J. Chemom.* **2006**, *20*, 247–272. [CrossRef]
2. Swarbrick, B.; Westad, F. An Overview of Chemometrics for the Engineering and Measurement Sciences. In *Handbook of Measurement in Science and Engineering*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2016; p. 2309.
3. Kumar, R.; Sharma, V. Chemometrics in forensic science. *Trac Trends Anal. Chem.* **2018**, *105*, 191–201. [CrossRef]
4. Brown, S. The chemometrics revolution re-examined. *J. Chemom.* **2017**, *31*, e2864. [CrossRef]
5. Hibbert David, B. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). *Pure Appl. Chem.* **2016**, *88*, 407. [CrossRef]
6. Caballero, B.; Finglas, P.; Toldrá, F. *Encyclopedia of Food and Health*; Academic Press: Cambridge, MA, USA, 2015.
7. Pomerantsev, A.L.; Rodionova, O.Y. Chemometric view on "comprehensive chemometrics". *Chemom. Intell. Lab. Syst.* **2010**, *103*, 19–24. [CrossRef]
8. Ferreira, S.L.C.; Silva Junior, M.M.; Felix, C.S.A.; da Silva, D.L.F.; Santos, A.S.; Santos Neto, J.H.; de Souza, C.T.; Cruz Junior, R.A.; Souza, A.S. Multivariate optimization techniques in food analysis—A review. *Food Chem.* **2019**, *273*, 3–8. [CrossRef]
9. De Luca, S.; De Filippis, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics. *Microchem. J.* **2016**, *129*, 348–361. [CrossRef]
10. Briandet, R.; Kemsley, E.K.; Wilson, R.H. Discrimination of Arabica and Robusta in Instant Coffee by Fourier Transform Infrared Spectroscopy and Chemometrics. *J. Agric. Food Chem.* **1996**, *44*, 170–174. [CrossRef]
11. Santos, P.M.; Pereira-Filho, E.R.; Rodriguez-Saona, L.E. Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis. *Food Chem.* **2013**, *138*, 19–24. [CrossRef]
12. Amorello, D.; Orecchio, S.; Pace, A.; Barreca, S. Discrimination of almonds (Prunus dulcis) geographical origin by minerals and fatty acids profiling. *Nat. Prod. Res.* **2016**, *30*, 2107–2110. [CrossRef]
13. Wu, X.M.; Zuo, Z.T.; Zhang, Q.Z.; Wang, Y.Z. Classification of Paris species according to botanical and geographical origins based on spectroscopic, chromatographic, conventional chemometric analysis and data fusion strategy. *Microchem. J.* **2018**, *143*, 367–378. [CrossRef]
14. Chen, H.; Lin, Z.; Tan, C. Fast discrimination of the geographical origins of notoginseng by near-infrared spectroscopy and chemometrics. *J. Pharm. Biomed. Anal.* **2018**, *161*, 239–245. [CrossRef] [PubMed]
15. Uríčková, V.; Sádecká, J. Determination of geographical origin of alcoholic beverages using ultraviolet, visible and infrared spectroscopy: A review. *Spectrochim. Acta Part A* **2015**, *148*, 131–137. [CrossRef] [PubMed]
16. Kanginejad, A.; Mani-Varnosfaderani, A. Chemometrics advances on the challenges of the gas chromatography– mass spectrometry metabolomics data: A review. *J. Iran. Chem. Soc.* **2018**, *15*, 2733–2745. [CrossRef]
17. Liu, S.; Liang, Y.Z.; Liu, H.T. Chemometrics applied to quality control and metabolomics for traditional Chinese medicines. *J. Chromatogr. B* **2016**, *1015–1016*, 82–91. [CrossRef]
18. Savorani, F.; Rasmussen, M.A.; Mikkelsen, M.S.; Engelsen, S.B. A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Res. Int.* **2013**, *54*, 1131–1145. [CrossRef]

19. Bhushan, N.; Hadpe, S.; Rathore, A.S. Chemometrics applications in biotech processes: Assessing process comparability. *Biotechnol. Prog.* **2012**, *28*, 121–128. [CrossRef] [PubMed]
20. Xu, Q.S.; Xu, Y.D.; Li, L.; Fang, K.T. Uniform experimental design in chemometrics. *J. Chemom.* **2018**, *32*, e3020. [CrossRef]
21. Gadžurić, S.B.; Podunavac Kuzmanović, S.O.; Vraneš, M.B.; Petrin, M.; Bugarski, T.; Kovačević, S.Z. Multivariate Chemometrics with Regression and Classification Analyses in Heroin Profiling Based on the Chromatographic Data. *Iran. J. Pharm. Res. IJPR* **2016**, *15*, 725–734.
22. Materazzi, S.; Gregori, A.; Ripani, L.; Apriceno, A.; Risoluti, R. Cocaine profiling: Implementation of a predictive model by ATR-FTIR coupled with chemometrics in forensic chemistry. *Talanta* **2017**, *166*, 328–335. [CrossRef]
23. Muehlethaler, C.; Massonnet, G.; Esseiva, P. The application of chemometrics on Infrared and Raman spectra as a tool for the forensic analysis of paints. *Forensic Sci. Int.* **2011**, *209*, 173–182. [CrossRef] [PubMed]
24. Thanasoulias, N.C.; Parisis, N.A.; Evmiridis, N.P. Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra. *Forensic Sci. Int.* **2003**, *138*, 75–84. [CrossRef] [PubMed]
25. Brereton, R.G. Pattern recognition in chemometrics. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 90–96. [CrossRef]
26. Roggo, Y.; Degardin, K.; Margot, P. Identification of pharmaceutical tablets by Raman spectroscopy and chemometrics. *Talanta* **2010**, *81*, 988–995. [CrossRef]
27. da Silva, V.H.; Soares-Sobrinho, J.L.; Pereira, C.F.; Rinnan, Å. Evaluation of chemometric approaches for polymorphs quantification in tablets using near-infrared hyperspectral images. *Eur. J. Pharm. Biopharm.* **2019**, *134*, 20–28. [CrossRef]
28. Dinç, E.; Büker, E. Spectrochromatographic determination of dorzolamide hydrochloride and timolol maleate in an ophthalmic solution using three-way analysis methods. *Talanta* **2019**, *191*, 248–256. [CrossRef]
29. Sakr, M.; Hanafi, R.; Fouad, M.; Al-Easa, H.; El-Moghazy, S. Design and optimization of a luminescent Samarium complex of isoprenaline: A chemometric approach based on Factorial design and Box-Behnken response surface methodology. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *208*, 114–123. [CrossRef]
30. Rodionova, O.Y.; Titova, A.V.; Demkin, N.A.; Balyklova, K.S.; Pomerantsev, A.L. Qualitative and quantitative analysis of counterfeit fluconazole capsules: A non-invasive approach using NIR spectroscopy and chemometrics. *Talanta* **2019**, *195*, 662–667. [CrossRef]
31. Visco, G.; Avino, P. Employ of multivariate analysis and chemometrics in cultural heritage and environment fields. *Environ. Sci. Pollut. Res.* **2017**, *24*, 13863–13865. [CrossRef]
32. Musumarra, G.; Fichera, M. Chemometrics and cultural heritage. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 363–372. [CrossRef]
33. Madariaga, J.M. Analytical chemistry in the field of cultural heritage. *Anal. Methods* **2015**, *7*, 4848–4876. [CrossRef]
34. Barreca, S.; Mazzola, A.; Orecchio, S.; Tuzzolino, N. Polychlorinated Biphenyls in Sediments from Sicilian Coastal Area (Scoglitti) using Automated Soxhlet, GC-MS, and Principal Component Analysis. *Polycycl. Aromat. Compd.* **2014**, *34*, 237–262. [CrossRef]
35. Granato, D.; Santos, J.S.; Escher, G.B.; Ferreira, B.L.; Maggio, R.M. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends Food Sci. Technol.* **2018**, *72*, 83–90. [CrossRef]
36. Panchuk, V.; Yaroshenko, I.; Legin, A.; Semenov, V.; Kirsanov, D. Application of chemometric methods to XRF-data—A tutorial review. *Anal. Chim. Acta* **2018**, *1040*, 19–32. [CrossRef] [PubMed]
37. Jain, D.; Basniwal, P.K. Forced degradation and impurity profiling: Recent trends in analytical perspectives. *J. Pharm. Biomed. Anal.* **2013**, *86*, 11–35. [CrossRef]
38. Rao, R.N.; Nagaraju, V. An overview of the recent trends in development of HPLC methods for determination of impurities in drugs. *J. Pharm. Biomed. Anal.* **2003**, *33*, 335–377.
39. Holm, R.; Elder, D.P. Analytical advances in pharmaceutical impurity profiling. *Eur. J. Pharm. Sci.* **2016**, *87*, 118–135. [CrossRef]
40. ICH. *Impurities in New Drug Substances Q3a(R2)*; Published by food and Drug Administration: Silver Spring, MD, USA, 2008. Available online: https://www.fda.gov/media/71272/download (accessed on 10 August 2019).
41. Melo, S.R.d.O.; Mello, M.H.d.; Silveira, D.; Simeoni, L.A. Advice on Degradation Products in Pharmaceuticals: A. *PDA J. Pharm. Sci. Technol.* **2014**, *68*, 221–238. [CrossRef]

42. Pan, C.; Liu, F.; Motto, M. Identification of pharmaceutical impurities in formulated dosage forms. *J. Pharm. Sci.* **2011**, *100*, 1228–1259. [CrossRef]
43. Sastry, R.P.; Venkatesan, C.; Sastry, B.; Mahesh, K. Identification and characterization of forced degradation products of pralatrexate injection by LC-PDA and LC–MS. *J. Pharm. Biomed. Anal.* **2016**, *131*, 400–409. [CrossRef]
44. Skibiński, R.; Trawiński, J.; Komsta, Ł.; Bajda, K. Characterization of forced degradation products of clozapine by LC-DAD/ESI-Q-TOF. *J. Pharm. Biomed. Anal.* **2016**, *131*, 272–280. [CrossRef] [PubMed]
45. Attwood, D.; Florence, A.T.; Rothschild, Z. *Princípios Físico-Químicos em Farmácia Volume 4*; Edusp: São Paulo, Brazil, 2003.
46. Gallardo, C.; Rojas, J.J.; Flórez, O.A. La temperatura cinética media en los estudios de estabilidad a largo plazo y almacenamiento de los medicamentos. *Vitae* **2004**, *11*, 67–72.
47. Allen Jr, L.V.; Popovich, N.G.; Ansel, H.C. *Formas Farmacêuticas e Sistemas de Liberação de Fármacos-9*; Artmed Editora: Porto Alegre, Brazil, 2013.
48. Blessy, M.; Patel, R.D.; Prajapati, P.N.; Agrawal, Y.K. Development of forced degradation and stability indicating studies of drugs—A review. *J. Pharm. Anal.* **2014**, *4*, 159–165. [CrossRef]
49. Qiu, F.; Norwood, D.L. Identification of pharmaceutical impurities. *J. Liq. Chromatogr. Relat. Technol.* **2007**, *30*, 877–935. [CrossRef]
50. Ahmad, I.; Ahmed, S.; Anwar, Z.; Sheraz, M.A.; Sikorski, M. Photostability and photostabilization of drugs and drug products. *Int. J. Photoenergy* **2016**, *2016*. [CrossRef]
51. Singh, S.; Handa, T.; Narayanam, M.; Sahu, A.; Junwal, M.; Shah, R.P. A critical review on the use of modern sophisticated hyphenated tools in the characterization of impurities and degradation products. *J. Pharm. Biomed. Anal.* **2012**, *69*, 148–173. [CrossRef]
52. ICH. *Stability Testing of New Drug Substances and Products Q1A (R2)*; Published by Food and Drug Administration: Silver Spring, MD, USA, 2003. Available online: https://www.fda.gov/media/71707/download (accessed on 11 August 2019).
53. Singh, S.; Junwal, M.; Modhe, G.; Tiwari, H.; Kurmi, M.; Parashar, N.; Sidduri, P. Forced degradation studies to assess the stability of drugs and products. *Trac Trends Anal. Chem.* **2013**, *49*, 71–88. [CrossRef]
54. Chen, W.-H.; Lin, Y.-Y.; Chang, Y.; Chang, K.-W.; Hsia, Y.-C. Forced degradation behavior of epidepride and development of a stability-indicating method based on liquid chromatography–mass spectrometry. *J. Food Drug Anal.* **2014**, *22*, 248–256. [CrossRef]
55. ANVISA Perguntas & Respostas. Assunto: RDC 53/2015 e Guia 4/2015. Available online: http://portal.anvisa.gov.br/documents/33836/418522/Perguntas+e+Respostas+-+RDC+53+2015+e+Guia+04+2015/6b3dec42-546c-4953-943f-4047b8b50f87 (accessed on 10 August 2019).
56. Canavesi, R.; Aprile, S.; Varese, E.; Grosa, G. Development and validation of a stability-indicating LC-UV method for the determination of pantethine and its degradation product based on a forced degradation study. *J. Pharm. Biomed. Anal.* **2014**, *97*, 141–150. [CrossRef]
57. Bhardwaj, S.P.; Singh, S. Study of forced degradation behavior of enalapril maleate by LC and LC-MS and development of a validated stability-indicating assay method. *J. Pharm. Biomed. Anal.* **2008**, *46*, 113–120. [CrossRef]
58. Palaric, C.; Molinié, R.; Cailleu, D.; Fontaine, J.-X.; Mathiron, D.; Mesnard, F.; Gut, Y.; Renaud, T.; Petit, A.; Pilard, S. A Deeper Investigation of Drug Degradation Mixtures Using a Combination of MS and NMR Data: Application to Indapamide. *Molecules* **2019**, *24*, 1764. [CrossRef] [PubMed]
59. Fatima, S.; Beg, S.; Samim, M.; Ahmad, F.J. *Application of Chemometric Approach for Development and Validation of High Performance Liquid Chromatography Method for Estimation of Ropinirole Hydrochloride*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2019.
60. ICH. *Photostability Testing of New Drug Substances and Products Q1B*; Published by Food and Drug Administration: Silver Spring, MD, USA, 1996. Available online: https://www.fda.gov/media/71713/download (accessed on 15 August 2019).
61. Bakshi, M.; Singh, S. Development of validated stability-indicating assay methods—Critical review. *J. Pharm. Biomed. Anal.* **2002**, *28*, 1011–1040. [CrossRef]
62. Bansal, G.; Singh, M.; Jindal, K.C.; Singh, S. Ultraviolet-photodiode array and high-performance liquid chromatographic/mass spectrometric studies on forced degradation behavior of glibenclamide and development of a validated stability-indicating method. *J. Aoac Int.* **2008**, *91*, 709–719. [PubMed]

63. World Health Organization. *WHO Expert Committee on Specifications for Pharmaceutical Preparations: Thirty-Ninth Report*; World Health Organization: Geneva, Switzerland, 2005; Volume 39.
64. Sanitária, A.N.d.V. *Resolução De Diretoria Colegiada—RDC Nº 53*; Diário Oficial da União: Brasília, Brazil, 2015.
65. Devrukhakar, P.S.; Shankar, M.S.; Shankar, G.; Srinivas, R. A stability-indicating LC–MS/MS method for zidovudine: Identification, characterization and toxicity prediction of two major acid degradation products. *J. Pharm. Anal.* **2017**, *7*, 231–236. [CrossRef] [PubMed]
66. Skibiński, R.; Trawiński, J.; Komsta, Ł.; Murzec, D. Characterization of forced degradation products of toloxatone by LC-ESI-MS/MS. *Saudi Pharm. J.* **2018**, *26*, 467–480. [CrossRef] [PubMed]
67. Tiwari, R.N.; Shah, N.; Bhalani, V.; Mahajan, A. LC, MSn and LC–MS/MS studies for the characterization of degradation products of amlodipine. *J. Pharm. Anal.* **2015**, *5*, 33–42. [CrossRef]
68. Rakibe, U.; Tiwari, R.; Mahajan, A.; Rane, V.; Wakte, P. LC and LC–MS/MS studies for the identification and characterization of degradation products of acebutolol. *J. Pharm. Anal.* **2018**, *8*, 357–365. [CrossRef]
69. Martono, Y.; Rohman, A.; Martono, S.; Riyanto, S. Degradation study of stevioside using RP-HPLC and ESI-MS/MS. *Malays. J. Fundam. Appl. Sci.* **2018**, *14*, 138–141. [CrossRef]
70. Korany, M.A.; Haggag, R.S.; Ragab, M.A.A.; Elmallah, O.A. A validated stability indicating DAD–HPLC method for determination of pentoxifylline in presence of its pharmacopeial related substances. *Bull. Fac. Pharm.* **2013**, *51*, 211–219. [CrossRef]
71. Saini, B.; Bansal, G. Isolation and characterization of a degradation product in leflunomide and a validated selective stability-indicating HPLC–UV method for their quantification. *J. Pharm. Anal.* **2015**, *5*, 207–212. [CrossRef]
72. Abiramasundari, A.; Joshi, R.P.; Jalani, H.B.; Sharma, J.A.; Pandya, D.H.; Pandya, A.N.; Sudarsanam, V.; Vasu, K.K. Stability-indicating assay method for determination of actarit, its process related impurities and degradation products: Insight into stability profile and degradation pathways. *J. Pharm. Anal.* **2014**, *4*, 374–383. [CrossRef] [PubMed]
73. Al-Ghannam, S.M.; Al-Olayan, A.M. Stability-indicating HPLC method for the determination of nicardipine in capsules and spiked human plasma. Identification of degradation products using HPLC/MS. *Arab. J. Chem.* **2014**, in press. [CrossRef]
74. Alarfaj, N.A. Stability-indicating liquid chromatography for determination of clopidogrel bisulfate in tablets: Application to content uniformity testing. *J. Saudi Chem. Soc.* **2012**, *16*, 23–30. [CrossRef]
75. Xia, M.; Hang, T.J.; Zhang, F.; Li, X.M.; Xu, X.Y. The stability of biapenem and structural identification of impurities in aqueous solution. *J. Pharm. Biomed. Anal.* **2009**, *49*, 937–944. [CrossRef] [PubMed]
76. Shah, R.P.; Sahu, A.; Singh, S. Identification and characterization of degradation products of irbesartan using LC–MS/TOF, MSn, on-line H/D exchange and LC–NMR. *J. Pharm. Biomed. Anal.* **2010**, *51*, 1037–1046. [CrossRef]
77. US Department of Health and Human Services. *Guidance for Industry ANDAs: Impurities in Drug Substances*; US Department of Health and Human Services, Food and Drug Administration: Washington, DC, USA, 1999.
78. Robinson, T.J.; Borror, C.M.; Myers, R.H. Robust parameter design: A review. *Qual. Reliab. Eng. Int.* **2004**, *20*, 81–101. [CrossRef]
79. Breitkreitz, M.C.; Souza, A.M.d.; Poppi, R.J. Experimento didático de quimiometria para planejamento de experimentos: Avaliação das condições experimentais na determinação espectrofotométrica de ferro II com o-fenantrolina. *Química Nova* **2014**, *37*, 564–573.
80. Dejaegher, B.; Vander Heyden, Y. Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *J. Pharm. Biomed. Anal.* **2011**, *56*, 141–158. [CrossRef]
81. Mäkelä, M. Experimental design and response surface methodology in energy applications: A tutorial review. *Energy Convers. Manag.* **2017**, *151*, 630–640. [CrossRef]
82. Tye, H. Application of statistical 'design of experiments' methods in drug discovery. *Drug Discov. Today* **2004**, *9*, 485–491. [CrossRef]
83. Altekar, M.; Homon, C.A.; Kashem, M.A.; Mason, S.W.; Nelson, R.M.; Patnaude, L.A.; Yingling, J.; Taylor, P.B. Assay Optimization: A Statistical Design of Experiments Approach. *Clin. Lab. Med.* **2007**, *27*, 139–154. [CrossRef] [PubMed]

84. Bezerra, M.A.; Santelli, R.E.; Oliveira, E.P.; Villar, L.S.; Escaleira, L.A. Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta* **2008**, *76*, 965–977. [CrossRef] [PubMed]
85. Sahu, P.K.; Ramisetti, N.R.; Cecchi, T.; Swain, S.; Patro, C.S.; Panda, J. An overview of experimental designs in HPLC method development and validation. *J. Pharm. Biomed. Anal.* **2018**, *147*, 590–611. [CrossRef] [PubMed]
86. Lafossas, C.; Benoit-Marquié, F.; Garrigues, J.C. Analysis of the retention of tetracyclines on reversed-phase columns: Chemometrics, design of experiments and quantitative structure-property relationship (QSPR) study for interpretation and optimization. *Talanta* **2019**, *198*, 550–559. [CrossRef] [PubMed]
87. Valente, J.F.A.; Sousa, A.; Queiroz, J.A.; Sousa, F. DoE to improve supercoiled p53-pDNA purification by O-phospho-l-tyrosine chromatography. *J. Chromatogr. B* **2019**, *1105*, 184–192. [CrossRef]
88. Mahrouse, M.A.; Lamie, N.T. Experimental design methodology for optimization and robustness determination in ion pair RP-HPLC method development: Application for the simultaneous determination of metformin hydrochloride, alogliptin benzoate and repaglinide in tablets. *Microchem. J.* **2019**, *147*, 691–706. [CrossRef]
89. Krishna, M.V.; Dash, R.N.; Jalachandra Reddy, B.; Venugopal, P.; Sandeep, P.; Madhavi, G. Quality by Design (QbD) approach to develop HPLC method for eberconazole nitrate: Application oxidative and photolytic degradation kinetics. *J. Saudi Chem. Soc.* **2016**, *20*, S313–S322. [CrossRef]
90. Nadella, N.P.; Ratnakaram, V.N.; Srinivasu, N.; Technologies, R. Quality-by-design-based development and validation of a stability-indicating UPLC method for quantification of teriflunomide in the presence of degradation products and its application to in-vitro dissolution. *J. Liquid Chromatogr. Relat. Technol.* **2017**, *40*, 517–527. [CrossRef]
91. Hadzieva Gigovska, M.; Petkovska, A.; Acevska, J.; Nakov, N.; Antovska, P.; Ugarkovic, S.; Dimitrovska, A. Comprehensive Assessment of Degradation Behavior of Simvastatin by UHPLC/MS Method, Employing Experimental Design Methodology. *J. Int. J. Anal. Chem.* **2018**, *2018*, 17. [CrossRef]
92. Jadhav, S.B.; Reddy, P.S.; Narayanan, K.L.; Bhosale, P.N. Development of RP-HPLC, Stability Indicating Method for Degradation Products of Linagliptin in Presence of Metformin HCl by Applying 2 Level Factorial Design; and Identification of Impurity-VII, VIII and IX and Synthesis of Impurity-VII. *Sci. Pharm.* **2017**, *85*, 25. [CrossRef]
93. Wingert, N.R.; Ellwanger, J.B.; Bueno, L.M.; Gobetti, C.; Garcia, C.V.; Steppe, M.; Schapoval, E.E.S. Application of Quality by Design to optimize a stability-indicating LC method for the determination of ticagrelor and its impurities. *Eur. J. Pharm. Sci.* **2018**, *118*, 208–215. [CrossRef] [PubMed]
94. Ren, Z.; Zhang, X.; Wang, H.; Jin, X. Using an innovative quality-by-design approach for the development of a stability-indicating UPLC/Q-TOF-ESI-MS/MS method for stressed degradation products of imatinib mesylate. *RSC Adv.* **2016**, *6*, 13050–13062. [CrossRef]
95. Sharma, G.; Thakur, K.; Raza, K.; Katare, O.P. Stability kinetics of fusidic acid: Development and validation of stability indicating analytical method by employing Analytical Quality by Design approach in medicinal product(s). *J. Chromatogr. B* **2019**, *1120*, 113–124. [CrossRef] [PubMed]
96. Zhang, X.; Hu, C. Application of quality by design concept to develop a dual gradient elution stability-indicating method for cloxacillin forced degradation studies using combined mixture-process variable models. *J. Chromatogr. A* **2017**, *1514*, 44–53. [CrossRef]
97. Kalariya, P.D. Experimental Design Approach for Selective Separation of Vilazodone HCl and Its Degradants by LC-PDA and Characterization of Major Degradants by LC/QTOF–MS/MS. *Chromatographia* **2014**, *77*, 1299–1313. [CrossRef]
98. Murthy, M.V.; Krishnaiah, C.; Srinivas, K.; Rao, K.S.; Kumar, N.R.; Mukkanti, K. Development and validation of RP-UPLC method for the determination of darifenacin hydrobromide, its related compounds and its degradation products using design of experiments. *J. Pharm. Biomed. Anal.* **2013**, *72*, 40–50. [CrossRef]
99. Baghel, M.; Rajput, S.J. Stress degradation of edaravone: Separation, isolation and characterization of major degradation products. *Biomed. Chromatogr.* **2018**, *32*, e4146. [CrossRef]
100. Yeram, P.; Hamrapurkar, P.; Mukhedkar, P. Implementation of Quality by Design approach to develop and validate stability indicating assay method for simultaneous estimation of sofosbuvir and ledipasvir in bulk drugs and tablet formulation. *Int. J. Pharm. Sci.* **2019**, *10*, 180–188.
101. Sonawane, S.; Gide, P. Application of experimental design for the optimization of forced degradation and development of a validated stability-indicating LC method for luliconazole in bulk and cream formulation. *Arab. J. Chem.* **2016**, *9*, S1428–S1434. [CrossRef]

102. Kurmi, M.; Kumar, S.; Singh, B.; Singh, S. Implementation of design of experiments for optimization of forced degradation conditions and development of a stability-indicating method for furosemide. *J. Pharm. Biomed. Anal.* **2014**, *96*, 135–143. [CrossRef]
103. Fusion QbD Quality by Design Software Solutions. Available online: http://www.smatrix.com/ (accessed on 10 March 2019).
104. Originlab. Available online: https://www.originlab.com/ (accessed on 10 March 2019).
105. MATLAB for Artificial Intelligence. Available online: https://au.mathworks.com/ (accessed on 10 March 2019).
106. Minitab 19. Available online: https://www.minitab.com/ (accessed on 10 March 2019).
107. StatEase Statistics Made Easy. Available online: https://www.statease.com (accessed on 10 March 2019).
108. Accelerate Innovation with Data Science. Available online: https://www.tibco.com/products/data-science (accessed on 10 March 2019).
109. Dégardin, K.; Guillemain, A.; Guerreiro, N.V.; Roggo, Y. Near infrared spectroscopy for counterfeit detection using a large database of pharmaceutical tablets. *J. Pharm. Biomed. Anal.* **2016**, *128*, 89–97. [CrossRef] [PubMed]
110. The basic building block of chemometrics. *Analytical Chemistry*. Available online: https://www.intechopen.com/books/analytical-chemistry/pca-the-basic-building-block-of-chemometrics (accessed on 15 August 2019).
111. Godoy, J.L.; Vega, J.R.; Marchetti, J.L. Relationships between PCA and PLS-regression. *Chemom. Intell. Lab. Syst.* **2014**, *130*, 182–191. [CrossRef]
112. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
113. Rutledge, D.N.; Jouan-Rimbaud Bouveresse, D. Independent Components Analysis with the JADE algorithm. *Trac Trends Anal. Chem.* **2013**, *50*, 22–32. [CrossRef]
114. Tôrres, A.R.; Grangeiro, S.; Fragoso, W.D. Multivariate control charts for monitoring captopril stability. *Microchem. J.* **2015**, *118*, 259–265. [CrossRef]
115. Tôrres, A.R.; Grangeiro, S.; Fragoso, W.D. Vibrational spectroscopy and multivariate control charts: A new strategy for monitoring the stability of captopril in the pharmaceutical industry. *Microchem. J.* **2017**, *133*, 279–285. [CrossRef]
116. Bro, R. Multiway calibration. *Multilinear PLS J. Chemom.* **1996**, *10*, 47–61.
117. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom. A J. Chemom. Soc.* **2002**, *16*, 119–128. [CrossRef]
118. Krishnan, A.; Williams, L.J.; McIntosh, A.R.; Abdi, H. Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage* **2011**, *56*, 455–475. [CrossRef]
119. Medina, S.; Perestrelo, R.; Silva, P.; Pereira, J.A.M.; Câmara, J.S. Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends Food Sci. Technol.* **2019**, *85*, 163–176. [CrossRef]
120. Rosipal, R.; Krämer, N. *In Overview and Recent Advances in Partial Least Squares, International Statistical and Optimization Perspectives Workshop Subspace, Latent Structure and Feature Selection*; Springer: New York, NY, USA, 2005; pp. 34–51.
121. Biancolillo, A.; Marini, F. Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. *Front. Chem.* **2018**, *6*, 576. [CrossRef] [PubMed]
122. Sayed, R.A.; El-Masri, M.M.; Hassan, W.S.; El-Mammli, M.Y.; Shalaby, A. Validated Stability-Indicating Methods for Determination of Mometasone Furoate in Presence of its Alkaline Degradation Product. *J. Chromatogr. Sci.* **2017**, *56*, 254–261. [CrossRef] [PubMed]
123. Attia, K.A.-S.M.; Abdel-Aziz, O.; Magdy, N.; Mohamed, G.F. Development and validation of different chemometric-assisted spectrophotometric methods for determination of cefoxitin-sodium in presence of its alkali-induced degradation product. *Future J. Pharm. Sci.* **2018**, *4*, 241–247. [CrossRef]
124. Attia, K.A.M.; Nassar, M.W.I.; El-Zeiny, M.B.; Serag, A. Stability indicating methods for the analysis of cefprozil in the presence of its alkaline induced degradation product. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2016**, *159*, 1–6. [CrossRef] [PubMed]

125. Alamein, A.M.A.A.; Hussien, L.A.E.A.; Mohamed, E.H. Univariate spectrophotometry and multivariate calibration: Stability-indicating analytical tools for the quantification of pimozide in bulk and pharmaceutical dosage form. *Bull. Fac. Pharm.* **2015**, *53*, 173–183. [CrossRef]
126. Hegazy, M.A.E.-M.; Eissa, M.S.; Abd El-Sattar, O.I.; Abd El-Kawy, M. Two and three way spectrophotometric-assisted multivariate determination of linezolid in the presence of its alkaline and oxidative degradation products and application to pharmaceutical formulation. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2014**, *128*, 231–242. [CrossRef] [PubMed]
127. Hegazy, M.A.-M.; Eissa, M.S.; Abd El-Sattar, O.I.; Abd El-Kawy, M.M. Determination of a novel ACE inhibitor in the presence of alkaline and oxidative degradation products using smart spectrophotometric and chemometric methods. *J. Pharm. Anal.* **2014**, *4*, 132–143. [CrossRef]
128. Souza, J.A.L.; Albuquerque, M.M.; Grangeiro, S.; Pimentel, M.F.; de Santana, D.P.; Simões, S.S. Quantification of captopril disulphide as a degradation product in captopril tablets using near infrared spectroscopy and chemometrics. *Vib. Spectrosc.* **2012**, *62*, 35–41. [CrossRef]
129. Abou Al Alamein, A.M. Validated stability-indicating methods for the determination of zafirlukast in the presence of its alkaline hydrolysis degradation product. *Bull. Fac. Pharm.* **2012**, *50*, 111–119. [CrossRef]
130. Naguib, I.A. Stability indicating analysis of bisacodyl by partial least squares regression, spectral residual augmented classical least squares and support vector regression chemometric models: A comparative study. *Bull. Fac. Pharm.* **2011**, *49*, 91–100. [CrossRef]
131. Abdelwahab, N.S.J.A.M. Determination of atenolol, chlorthalidone and their degradation products by TLC-densitometric and chemometric methods with application of model updating. *Anal. Methods* **2010**, *2*, 1994–2001. [CrossRef]
132. Wagieh, N.E.; Hegazy, M.A.; Abdelkawy, M.; Abdelaleem, E.A. Quantitative determination of oxybutynin hydrochloride by spectrophotometry, chemometry and HPTLC in presence of its degradation product and additives in different pharmaceutical dosage forms. *Talanta* **2010**, *80*, 2007–2015. [CrossRef] [PubMed]
133. Moneeb, M.S. Chemometric determination of rabeprazole sodium in presence of its acid induced degradation products using spectrophotometry, polarography and anodic voltammetry at a glassy carbon electrode. *Pak. J. Pharm. Sci.* **2008**, *21*, 214–224. [PubMed]
134. Fayed, A.S.; Shehata, M.; Ibrahim, A.; Hassan, N.; Weshahy, S.A. Validated stability-indicating methods for determination of cilostazol in the presence of its degradation products according to the ICH guidelines. *J. Pharm. Biomed. Anal.* **2007**, *45*, 407–416. [CrossRef] [PubMed]
135. Ragno, G.; Ioele, G.; De Luca, M.; Garofalo, A.; Grande, F.; Risoli, A. A critical study on the application of the zero-crossing derivative spectrophotometry to the photodegradation monitoring of lacidipine. *J. Pharm. Biomed. Anal.* **2006**, *42*, 39–45. [CrossRef] [PubMed]
136. Shehata, M.A.; Ashour, A.; Hassan, N.Y.; Fayed, A.S.; El-Zeany, B.A. Liquid chromatography and chemometric methods for determination of rofecoxib in presence of its photodegradate and alkaline degradation products. *Anal. Chim. Acta* **2004**, *519*, 23–30. [CrossRef]
137. Jaumot, J.; de Juan, A.; Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1–12. [CrossRef]
138. Ruckebusch, C.; Blanchet, L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Anal. Chim. Acta* **2013**, *765*, 28–36. [CrossRef]
139. Firmani, P.; Hugelier, S.; Marini, F.; Ruckebusch, C. MCR-ALS of hyperspectral images with spatio-spectral fuzzy clustering constraint. *Chemom. Intell. Lab. Syst.* **2018**, *179*, 85–91. [CrossRef]
140. Devos, O.; Schröder, H.; Sliwa, M.; Placial, J.P.; Neymeyr, K.; Métivier, R.; Ruckebusch, C. Photochemical multivariate curve resolution models for the investigation of photochromic systems under continuous irradiation. *Anal. Chim. Acta* **2019**, *1053*, 32–42. [CrossRef]
141. Alcaraz, M.R.; Aguirre, A.; Goicoechea, H.C.; Culzoni, M.J.; Collins, S.E. Resolution of intermediate surface species by combining modulated infrared spectroscopy and chemometrics. *Anal. Chim. Acta* **2019**, *1049*, 38–46. [CrossRef] [PubMed]
142. Cook, D.W.; Oram, K.G.; Rutan, S.C.; Stoll, D.R. Rational Design of Mixtures for Chromatographic Peak Tracking Applications via Multivariate Selectivity. *Anal. Chim. Acta* **2019**, *2*, 100010. [CrossRef]
143. Marín-García, M.; Ioele, G.; Franquet-Griell, H.; Lacorte, S.; Ragno, G.; Tauler, R. Investigation of the photodegradation profile of tamoxifen using spectroscopic and chromatographic analysis and multivariate curve resolution. *Chemom. Intell. Lab. Syst.* **2018**, *174*, 128–141. [CrossRef]

144. Feng, X.; Zhang, Q.; Cong, P.; Zhu, Z. Determination of the paracetamol degradation process with online UV spectroscopic and multivariate curve resolution-alternating least squares methods: Comparative validation by HPLC. *Anal. Methods* **2013**, *5*, 5286–5293. [CrossRef]
145. Gómez-Canela, C.; Bolivar-Subirats, G.; Tauler, R.; Lacorte, S. Powerful combination of analytical and chemometric methods for the photodegradation of 5-Fluorouracil. *J. Pharm. Biomed. Anal.* **2017**, *137*, 33–41. [CrossRef]
146. Bērziņš, K.; Kons, A.; Grante, I.; Dzabijeva, D.; Nakurte, I.; Actiņš, A. Multi-technique approach for qualitative and quantitative characterization of furazidin degradation kinetics under alkaline conditions. *J. Pharm. Biomed. Anal.* **2016**, *129*, 433–440. [CrossRef]
147. De Luca, M.; Ioele, G.; Mas, S.; Tauler, R.; Ragno, G. A study of pH-dependent photodegradation of amiloride by a multivariate curve resolution approach to combined kinetic and acid–base titration UV data. *Analyst* **2012**, *137*, 5428–5435. [CrossRef]
148. Mas, S.; Tauler, R.; de Juan, A. Chromatographic and spectroscopic data fusion analysis for interpretation of photodegradation processes. *J. Chromatogr. A* **2011**, *1218*, 9260–9268. [CrossRef]
149. De Luca, M.; Mas, S.; Ioele, G.; Oliverio, F.; Ragno, G.; Tauler, R. Kinetic studies of nitrofurazone photodegradation by multivariate curve resolution applied to UV-spectral data. *Int. J. Pharm.* **2010**, *386*, 99–107. [CrossRef]
150. Javidnia, K.; Hemmateenejad, B.; Miri, R.; Saeidi-Boroujeni, M. Application of a self-modeling curve resolution method for studying the photodegradation kinetics of nitrendipine and felodipine. *J. Pharm. Biomed. Anal.* **2008**, *46*, 597–602. [CrossRef]
151. Shamsipur, M.; Hemmateenejad, B.; Akhond, M.; Javidnia, K.; Miri, R. A study of the photo-degradation kinetics of nifedipine by multivariate curve resolution analysis. *J. Pharm. Biomed. Anal.* **2003**, *31*, 1013–1019. [CrossRef]
152. Arabzadeh, V.; Sohrabi, M.R.; Goudarzi, N.; Davallo, M. Using artificial neural network and multivariate calibration methods for simultaneous spectrophotometric analysis of Emtricitabine and Tenofovir alafenamide fumarate in pharmaceutical formulation of HIV drug. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *215*, 266–275. [CrossRef] [PubMed]
153. Marini, F.; Bucci, R.; Magrì, A.L.; Magrì, A.D. Artificial neural networks in chemometrics: History, examples and perspectives. *Microchem. J.* **2008**, *88*, 178–185. [CrossRef]
154. Golubović, J.B.; Protić, A.D.; Zečević, M.L.; Otašević, B.M. Quantitative structure retention relationship modeling in liquid chromatography method for separation of candesartan cilexetil and its degradation products. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 92–101. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Molecules Editorial Office
E-mail: molecules@mdpi.com
www.mdpi.com/journal/molecules

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03936-786-3