Klassifizierung und Charakterisierung des Metaboloms von Lebensmitteln mit Random Forest Methoden

Dissertation

Zur Erlangung des akademischen Grades

Doctor rerum naturalium

Dr. rer. nat

aus dem Fachbereich Chemie der Universität Hamburg

vorgelegt von

Soeren Hendrik Wenck

Hamburg, den 30.03.2025

1. Gutachter der Dissertation: Prof. Dr. Stephan Seifert

2. Gutachter der Dissertation: Prof. Dr. Maria Buchweitz

1. Mitglied der Prüfungskommission der Disputation: Prof. Dr. Stephan Seifert

2. Mitglied der Prüfungskommission der Disputation: Prof. Dr. José A.C. Broekaert

3. Mitglied der Prüfungskommission der Disputation: Prof. Dr. Volkmar Vill

Datum der Disputation: 18.07.2025

Die vorliegende Arbeit wurde im Zeitraum von November 2020 bis April 2024 am Institut für Lebensmittelchemie und der *Hamburg School of Food Science* des Fachbereichs Chemie der Fakultät für Mathematik, Informatik und Naturwissenschaften der Universität Hamburg unter der Anleitung von Herrn Prof. Dr. Stephan Seifert angefertigt.

Inhalt

I. Publikationsliste

Veröffentlichungen als Erstautor

1. Opening the Random Forest Black Box of the Metabolome by the Application of Surrogate Minimal Depth – **Soeren Wenck**, Marina Creydt, Jule Hansen, Florian Gärber, Markus Fischer, Stephan Seifert; *Metabolites*, 2022, 12, 5 (13 pp)
2. Opening the Random Forest Black Box of ^{1}H NMR Metabolomics Data by the Exploitation of Surrogate Variables – **Soeren Wenck**, Thorsten Mix, Markus Fischer, Thomas Hackl, Stephan Seifert; *Metabolites*, 2023, 13, 1075 (16 pp)
3. Authentication of apples (*Malus domestica* BORKH.) according to geographical origin, variety and production method using ^{1}H NMR spectroscopy and random forests – **Soeren Wenck**, Rene Bachmann, Sarah-Marie Barmbold, Anna Lena Horns, Nele Paasch, Stephan Seifert; *Food Control*, 2024

Veröffentlichungen als Koautor

1. Impact of Freeze-Drying on the Determination of the Geographical Origin of Almonds (*Prunus dulcis* MILL.) by Near-Infrared (NIR) Spectroscopy – Henri Lösel, Navid Shakiba, **Soeren Wenck**, Phat Le Tan, Maike Arndt, Stephan Seifert, Thomas Hackl, Markus Fischer; *Food Analytical Methods*, 2022, 15, 2847-2857
2. Determination of the geographical origin of hazelnuts (*Corylus avellana* L.) by Near-Infrared spectroscopy (NIR) and a Low-Level Fusion with nuclear magnetic resonance (NMR) – Navid Shakiba, Annika Gerdes, Nathalie Holz, **Soeren Wenck**, René Bachmann, Tobias Schneider, Stephan Seifert, Markus Fischer, Thomas Hackl; *Microchemical Journal*, 2022, 174, 107066 (8 pp)
3. Food Monitoring: Limitations of Accelerated Storage to Predict Molecular Changes in Hazelnuts (*Corylus avella*na L.) under Realistic Conditions Using UPLC-ESI-IM-QTOF-MS – Henri Lösel, Navid Shakiba, **Soeren Wenck**, Phat Le Tan, Tim-Oliver Karstens, Marina Creydt, Stephan Seifert, Thomas Hackl, Markus Fischer; *Metabolites*, 2023, 13, 1031 (17 pp)
4. Exploring the potential of high-resolution LC-MS in combination with ion mobility separation and surrogate minimal depth for en- hanced almond origin authentication – Henri Lösel, Maike Arndt, **Soeren Wenck**, Lasse Hansen, Marie Oberpottkamp, Stephan Seifert, Markus Fischer; *Talanta*, 2023, 271, 125598 (9 pp)
5. Analysis of hazelnuts (Corylus avellana L.) stored for extended pe- riods by ^{1}H NMR spectroscopy monitoring storage-induced changes in the polar and nonpolar metabolome – Navid Shakiba, Henri Lösel, **Soeren Wenck**, Leif Kumpmann, René Bachmann, Marina Creydt, Stephan Seifert, Markus Fischer, Thomas Hackl; *Journal of Agricultural and Food Chemistry*, 2023, 71, 3093-3101
6. Rapid testing in the food industry: the potential of Fourier trans- form near-infrared (FTIR) spectroscopy

and spatially offset Raman spectroscopy (SORS) to detect raw material defects in hazelnuts (Corylus avellana L.) – Henri Lösel, Navid Shakiba, René Bachmann, **Soeren Wenck**, Phat Le Tan, Marina Creydt, Stephan Seifert, Thomas Hackl, Markus Fischer; *Food Analytical Methods*, 2024

7. Detection of almonds (Prunus dulcis) adulteration by genotyping of sweet and bitter almonds with double-mismatch allele-specific qPCR (DMAS-qPCR) – Nils Wax, Lucas F. Voges, **Soeren Wenck**, Jana L. Herold, Stephan Seifert, Markus Fischer; *Food Control*, 2023, 152, 109866

Vorträge

1. Klassifizierung und Charakterisierung des Metaboloms von weißem Spargel mit Random Forest Methoden – Regionalverbandstagung Nord, 2022, Bremerhaven
2. Charakterisierung des Metaboloms von weißem Spargel mit Random Forest Methoden – KI Tag des Arbeitskreises Chemometrik und Qualitätssicherung der GDCh – Chemometrics meets AI, 2022, Berlin
3. Untersuchung von Äpfeln – Regionalverbandstagung Nord, 2024, Bremerhaven

Poster

1. Charakterisierung des Metaboloms von Lebensmitteln mit Surrogate Minimal Depth – **Soeren Wenck**, Stephan Seifert; 49. Lebensmittelchemikertag, 2021, digital
2. Characterization of the LC-MS Asparagus Metabolome with Surrogate Minimal Depth – **Soeren Wenck**, Stephan Seifert; Summer School of the School of Analytical Sciences Adlershof (SALSA) - Make and Measure ... and Machines, 2021, Berlin
3. Opening the Random Forest Black Box of the Asparagus Metabolome by Random Forest Approaches – CAC2022 - Chemometrics in Analytical Chemistry, 2022, Rom
4. Klassifizierung und Charakterisierung von weißem Spargel mit Random Forest Methoden – **Soeren Wenck**, Stephan Seifert; 50. Lebensmittelchemikertag, 2022, Hamburg
5. Classification and characterization of truffles with ^{1}H NMR and random forest methods – **Soeren Wenck**, Stephan Seifert; sensorFINT conference, 2023, Berlin

II. Abkürzungsverzeichnis

Abkürzung	Bedeutung	Englische Bedeutung
EG	Europäische Gemeinschaft	European Union
BfR	Bundesinstitut für Risikobewertung	(German) federal institution for risk assessment
LC	Flüssigchromatographie	liquid chromatography
MS	Massenspektrometrie	mass spectrometry
LC-MS	Flüssigchromatographie gekoppelt mit Massenspektrometrie	liquid chromatography coupled with mass spectrometry
m/z	Masse zu Ladung	mass to charge
RP-HPLC	Umkehrphasen-Hochleistungsflüssig-Chromatographie	reverse phase high performance liquid chromatography
MS	Massenspektrometer	mass spectrometer
COSY	Korrelationsspektroskopie	correlation spectroscopy
TOCSY	Vollständige Korrelationsspektroskopie	total correlation spectroscopy
STOCSY	Statistische vollständige Korrelationsspektroskopie	statistical total correlation spectroscopy
ML	Maschinelles Lernen	machine learning

PCA	Hauptkomponentenanalyse	principal component analysis
RF	Random Forest	random forest
OOB	Out-of-bag Fehler	out of bag error
SV	Schattenvariablen	shadow variables
SMD	Surrogate Minimal Depth	surrogate minimal depth
MAA	*mean adjusted agreement*	mean adjusted agreement
SVM	*support vector machine*	support vector machine
ANN	Künstliche neuronale Netzwerke	artificial neural networks
IR	Infrarotspektroskopie	infrared spectroscopy
FTIR	Fourier-Transform-Infrarotspektroskopie	Fourier transform infrared spectroscopy
PLS-DA	*partial least square discriminant analysis*	partial least square discriminant analysis
MFI	*mutual forest impact*	mutual forest impact

III. Abbildungsverzeichnis

IV. Tabellenverzeichnis

1. Zusammenfassung

Im Rahmen dieser Arbeit wurden große analytische Datensätze des Metaboloms mit *random forest* (RF) Verfahren untersucht. Dabei wurden ausgewählte Lebensmittel hinsichtlich verschiedener Eigenschaften klassifiziert, relevante Variablen mit Variablenselektionsmethoden ausgewählt und deren gemeinsamer Einfluss auf das Klassifikationsmodell analysiert. Die auf diese Weise gefundenen Zusammenhänge wurden bezüglich des analytischen und biologischen Hintergrunds interpretiert und damit gezeigt, dass anhand der hier angewendeten Methoden eine detaillierte Analyse der untersuchten Proben, die weit über die bei *machine learning*-Verfahren häufig angewendete „*black box*" Untersuchung hinaus geht, ermöglicht wird.

Die Untersuchungen erfolgten an Metabolom-Daten aus ^{1}H Kernspinresonanz-spektroskopie (engl.: *nuclear magnetic resonance*, NMR-Spektroskopie) und gekoppelter Flüssigchromatographie mit Massenspektrometrie (engl.: *liquid chromatography coupled with mass spectrometry*, LC-MS) von Apfel-, Spargel- und Trüffelproben. Die Daten wurden dabei zuerst mit der oft eingesetzten Hauptkomponentenanalyse (*principal component analysis*, PCA) untersucht, um die Hauptunterschiede in den Datensätzen zu analysieren. Dabei zeigte sich, dass diese meistens keine klare Unterscheidung der analysierten Klassen ermöglichte und somit überwachte Verfahren angewendet werden sollten. RF zeigte sich als sehr gut geeignet, um die Datensätze mit teilweise recht geringen Stichprobengrößen einzelner Klassen zu klassifizieren, da durch die interne Validierung in Kombination mit dem Verzicht auf eine Optimierung der Modellparameter ein unabhängiger Validierungsfehler erhalten werden konnte, ohne zusätzliche Daten zu benötigen. Dabei konnten Klassifizierungsgenauigkeiten über 70 %, meist zwischen 80-100 %, erreicht werden.

Die Anwendung von *surrogate minimal depth* (SMD) zur Selektion relevanter Variablen und deren Beziehungsanalyse, zusammen mit der anschließenden Identifizierung mit Datenbankabgleich und LC-MS-MS Analyse, bzw. der zusätzlichen Analyse mit weiteren

Methoden der NMR-Spektroskopie und *spike-in*-Experimenten erwies sich als ein leistungsfähiger Ansatz zur Untersuchung der Wirkung von Variablen in den RF Modellen und damit deren Beitrag zur erfolgreichen Klassifizierung von Lebensmitteln. Dabei konnten sowohl Signale der gleichen Metabolite als auch biologisch sinnvolle Beziehungen zwischen einzelnen Metaboliten aufgedeckt werden.

2. Abstract

In the presented work large data sets of metabolomics data were analyzed with random forest (RF) methods. Data of chosen foods were classified concerning different characteristics. Relevant variables were selected with different approaches and their shared impact on the research outcome was investigated. The resulting connections were interpreted in light of their biological and analytical background. It was shown, that the applied methods provide a more detailed analysis of the examined samples, which went beyond the commonly applied black box approach that is associated with machine learning methods.

The examinations were conducted on different metabolomics data sets of ^{1}H nuclear magnetic resonance spectroscopy (NMR-spectroscopy) or liquid chromatography coupled with mass spectrometry (LC-MS) of apples, asparagus and truffles. In the first step all data sets were analyzed with the commonly applied principal component analysis (PCA) to show the main differences in the datasets. It was shown, that no clear indication of the analyzed classes was represented in the main variance of the data sets. Based on these findings, supervised machine learning approaches needed to be applied. RF presented itself as a powerful classification algorithm to analyze data sets especially with smaller sample sizes, since the combination of internal validation with the omission of parameter optimization resulted in independent validation errors without the need for further data. With this approach classification accuracies of above 70 %, but mostly 80-100 % were reached.

The application of surrogate minimal depth (SMD) for the selection of relevant variables and their relation analysis, combined with a subsequent comparison to data banks and LC-MS-MS analyses or additional NMR-spectroscopy methods and spike-in experiments resulted in a powerful approach for the analysis of the effect of variables to the classification model, and therefore for the classification of foods. With this combination signals from the same metabolite as well as logical biological relations between metabolites could be identified.

3. Einleitung

Die Informatik, allen voran Methoden im Bereich des *machine learnings*, sowie die Lebenswissenschaften gelten aktuell als die Wissenschaftsdisziplinen mit der größten Dynamik.[1,2] Sprunghafter Fortschritt entsteht oft an der Schnittstelle von Wissenschaftsdisziplinen.[3] Die vorliegende Arbeit entstand an der Schnittstelle modernster *machine learning* (ML)-Methoden, die auf komplexe Datensätze von Metabolom-Analysen ausgesuchter Lebensmittel angewendet wurden.

Die Lebensmittelanalytik dient der Qualitätskontrolle und verhindert überdies falsche Kennzeichnungen bzw. deckt diese auf. Dazu erhebt man vermehrt detaillierte analytische Fingerabdrücke der natürlichen Lebensmittelbestandteile[4], um bspw. deren Herkunft, Anbaumethode und Sorte eindeutig zu klassifizieren.[5–7] Hier kommt der Metabolom-Analyse ein steigender Stellenwert zu; jedoch sind hier die resultierenden Datensätze hochgradig komplex und ziehen langwierige, aufwändige Datenanalytik nach sich.[8] Dabei können neueste ML-Verfahren wertvoll unterstützen, indem sie Gemeinsamkeiten, Unterschiede und damit ungewünschte oder unerlaubte Abweichungen der üblichen Zusammensetzung von Lebensmitteln herausarbeiten. Ein oft genanntes Problem dieser Methoden besteht in der mangelnden Nachvollziehbarkeit selbstlernender Systeme, deren intransparente Ergebnisfindung deshalb oft als „*black box*“ bezeichnet wird.[9] Diesem Defizit soll in der vorliegenden Arbeit nachgegangen werden, indem Random Forest (RF) Datenanalyseverfahren auf unterschiedliche, große Datensätze ausgewählter Analysen des Metaboloms von Lebensmitteln angewandt werden. Dabei sollen insbesondere Variablenselektionsverfahren verwendet werden, um relevante Variablen zu identifizieren und deren Wirken im RF-Modell aufzuklären. Zu Beginn dieser Arbeit sollen zunächst die relevanten Grundlagen bezüglich der untersuchten Lebensmittel (3.1), aus dem Bereich Metabolomics (3.2), sowie der Methode des maschinellen Lernens (3.3) und der verwendeten Software (3.4) ausgeführt werden.

3.1. Lebensmittel

Da diese Arbeit im Zusammenhang mit der Entwicklung von Verfahren zur Authentifizierung von Lebensmitteln steht, wird im Folgenden zunächst allgemein auf Lebensmittel (3.1.1), Lebensmittelbetrug (3.1.2) und Lebensmittelauthentifizierung (3.1.3) sowie auf die Wahl der untersuchten Lebensmittel (3.1.4), Apfel (3.1.5), Trüffel (3.1.6) und Spargel (3.1.7) eingegangen werden.

3.1.1. Definition Lebensmittel

Lebensmittel sind gemäß Artikel 2 Basisverordnung (EG) Nr. 178/2002 alle Stoffe und Erzeugnisse, von denen nach vernünftigem Ermessen erwartet wird, dass sie im verarbeiteten, teilweise verarbeiteten oder unverarbeiteten Zustand vom Menschen aufgenommen werden. Hierzu zählen auch alle Stoffe, die für die Zubereitung eines Lebensmittels verwendet werden, um bestimmte sensorische Faktoren wie Aroma oder Geschmack zu bewirken. Allgemein gehören hierzu alle pflanzlichen und tierischen Lebensmittel sowie die Fruchtkörper von Speisepilzen, welche epi- oder hypogäisch wachsen können. Gemäß Artikel 14 Abs. (1) der Basisverordnung (EG) Nr. 178/2002 ist es verboten, Lebensmittel in den Verkehr zu bringen, die nicht sicher sind, also für den Endverbraucher gesundheitliche Risiken bergen. Die Richtlinien für das Inverkehrbringen ist die Lebensmittelinformationsverordnung (EG) Nr. 1169/2011.

Lebensmittel sind im Allgemeinen komplexe Matrices, die aus diversen Stoffen zusammengesetzt sind. Aus diesen Matrices werden während des Abbaus durch chemische, enzymatische und physikalische Verdauungsprozesse kleinere Moleküle extrahiert, die der menschliche Organismus für verschiedenste Bioprozesse verwendet.

So entstehen bspw. aus Polysacchariden über mehrere Schritte zunächst Monosaccharide, aus denen dann durch Spaltungs- und Umlagerungsreaktionen Pyruvat gebildet wird.[10] Pyruvat wird unter anderem in den Mitochondrien direkt in den

Citratzyklus zur Energiegewinnung eingespeist oder dient in anderen Stoffwechselprozessen zur Biosynthese von Grundbausteinen für Lipid- oder Proteinbiosynthese.[11,12]

Auch Proteine aus Lebensmittelmatrices werden während des Verdauungsprozesses in Aminosäuren oder kleinere Bausteine abgebaut. Einige Aminosäuren können vom menschlichen Körper nicht synthetisiert werden, weswegen ihre Aufnahme über die Ernährung überlebenswichtig ist.[12]

Lipide werden in der β_1-Oxidation in Pyruvat-Bausteine umgesetzt. Wie bei Aminosäuren gibt es ebenfalls essentielle Fettsäuren, die vom Körper nicht synthetisiert werden.[12]

Überdies gewinnt der Organismus aus Lebensmitteln wichtige Vitamine, sekundäre Pflanzenstoffe sowie Spurenelemente, u.a. als Zentralatome in Proteinen.[12]

Da Lebensmittel überlebenswichtig sind, stellt die Rechtsprechung einen hohen Anspruch an die Echt- und die Sicherheit von Lebensmitteln.

3.1.2. Lebensmittelbetrug

Wenn ein Lebensmittel nicht die Ansprüche von Echt- und Sicherheit erfüllt, kann es sich um Lebensmittelbetrug handeln. Gemäß Artikel 8 der Basisverordnung (EG) Nr. 178/2002 zum Schutz der Verbraucherinteressen sowie der Kontrollverordnung (EG) Nr. 2017/625 wird unter dem Begriff “Lebensmittelbetrug” das vorsätzliche Inverkehrbringen von Lebensmitteln verstanden, um durch gezielte Täuschung einen finanziellen oder wirtschaftlichen Vorteil zu erreichen.

Betrugsansätze gemäß Bundesinstitut für Risikobewertung (BfR) umfassen den Zusatz eines Fremdstoffs zur Vortäuschung einer besseren Qualität oder zur Streckung, Falschdeklarationen von Art, Sorte oder Herkunft, illegitime Güte- und/oder Qualitätssiegel sowie unerlaubte Herstellungsprozesse.[13] Insbesondere hochpreisige Produkte oder Lebensmittel beliebter Marken sind von Lebensmittelbetrug betroffen. Für viele

Lebensmittel gibt es spezifische Vorschriften und Richtlinien, um die Echtheit zu überprüfen, wie z.B. für Fruchtsäfte[14], Honig[15,16], Olivenöl[17] und Wein[18].

3.1.3. Lebensmittelauthentifizierung

Die Lebensmittelauthentifizierung erfüllt den Zweck, Lebensmittel auf ihre Echtheit und die Erfüllung der rechtlichen Grundlagen zu untersuchen. Dies geschieht klassischerweise mit etablierten *targeted* Methoden. Bei *targeted* Methoden handelt es sich häufig um nasschemische Verfahren, bei denen einzelne Bestandteile eines untersuchten Lebensmittels bestimmt werden und deren Echtheit anhand von Erfahrungswerten oder Datenbanken überprüft wird.[19] Eine große Schwachstelle ist, dass auch Betrüger diese Methoden nutzen und ein Lebensmittel gezielt modifizieren können, damit dieses in *targeted* Untersuchungen unverdächtig wirkt. Es können also bewusst verfälschte Produkte in Verkehr gebracht werden, die eine Prüfsubstanz in der für das Lebensmittel bekannten Konzentration enthalten und deshalb in einer *targeted*-Analyse nicht als solche identifiziert werden. Ein Beispiel hierfür ist der populäre Betrugsfall von chinesischer Babynahrung im Jahr 2008. Hierbei wurde die stickstoffreiche Verbindung Melamin zugesetzt, um bei der Analyse einen höheren Proteingehalt vorzutäuschen.[20]

Für die Authentifizierung von Lebensmitteln kann es daher sinnvoll sein, einen möglichst großen Anteil der enthaltenen Inhaltstoffe gleichzeitig mit *non-targeted* Methoden zu erfassen. Hierbei wird ein Datensatz aus authentischen Proben erhoben und ein Modell trainiert, welches auf den Unterschieden der untersuchten Klassen basiert. Dieser Datensatz muss auf einer ausreichenden Anzahl an Stichproben für jede der Klassen basieren, um die Varianz dieser Klassen im Modell abzubilden. Zur Analyse einer potenziell gefälschten Probe wird deren Klassenzugehörigkeit mit dem entwickelten Modell vorhergesagt und, falls die Vorhersage auf eine Fälschung hindeutet, das Ergebnis mit weiteren Verfahren validiert.[19]

3.1.4. Wahl der analysierten Lebensmittel

Im Rahmen dieser Arbeit wurden Datensätze von Lebensmitteln untersucht, die sich biologisch möglichst stark unterscheiden sollten: Trüffel gehören zum taxonomischen Reich der *Fungi*, Äpfel und Spargel zum Reich der *Plantae*. Bei Äpfeln handelt es sich um eine ausgewachsene Frucht, bei Spargel um eine noch im Wachstum befindliche Sprossachse mit starker Zellproliferation. Im Folgenden wird jedes der drei untersuchten Lebensmittel im Zusammenhang mit der Relevanz der jeweiligen Authentifizierung dargestellt.

3.1.5. Trüffel

Der Genus *Tuber* fasst die "echten Trüffel" zusammen, welche eine Gattung innerhalb der Familie der Trüffelähnlichen (*Tuberaceae*) definieren. Diese Gattung besteht aus weltweit mehr als 200 Arten von Schlauchpilzen[21], von welchen 30 auch in Europa vorkommen.[22,23]

Im Allgemeinen wird zwischen weißen und schwarzen Speisetrüffeln unterschieden. Die wirtschaftlich wichtigsten Vertreter der Speisetrüffel sind die drei schwarzen Arten *Tuber melanosporum* VITTADINI, "Périgord-Trüffel", *T. aestivum* VITTADINI, "Sommertrüffel" und *T. indicum*, der "Himalaya-" oder "China-Trüffel", sowie die zwei weißen Arten *T. magnatum pico*, "Alba-Trüffel" und *T. borchii* VITTADINI, "Bianchetti/Bianchetto-Trüffel".

Innerhalb der weißen und schwarzen Trüffel gibt es aufgrund des verschieden stark ausgeprägten Aromaprofils preislich große Unterschiede.[24,25] *T. magnatum pico* erreicht Marktpreise zwischen 1000-5000 €/kg[26,27], während *T. borchii* zwischen 105 – 305 €/kg gehandelt wird.[26] Innerhalb der schwarzen Trüffel gilt der *T. melanosporum* mit Preisen zwischen 950 – 4000 €/kg als der teuerste Vertreter.[25,28] Der schwarze Sommertrüffel *T. aestivum* hingegen wird mit Preisen von etwa 200 €/kg gehandelt[29], während *T. indicum* Preise zwischen 70 – 200 €/kg erzielt[30]. Auf verschiedenen Onlineportalen wird der Marktpreis von Speisetrüffeln verzeichnet und kann täglich eingesehen werden. Auch

hier werden die Arten *T. melanosporum* und *T. magnatum pico* als die teuersten Vertreter aufgeführt mit einem Durchschnittlichen Preis von 984 € für *T. melanosporum* und 1960 € für *T. magnatum pico* für je 20-50 g.[31]

Aufgrund der großen Preisunterschiede besteht daher ein finanzieller Anreiz, Lebensmittelbetrug mit Trüffeln zu betreiben.

3.1.6. Apfel

Der Genus Apfel, *Malus* sp., beschreibt mehr als 38 Arten von laubbildenden Kernobstgewächsen (*Pyrinae*) innerhalb der Familie der Rosengewächse (*Rosaceae*) mit einer Verbreitung in überwiegend gemäßigten Klimazonen in Asien, Europa, Nordamerika und Südafrika.[32]

Die Authentifizierung von Äpfeln ist auf mehreren Ebenen von Relevanz, da Verbraucher an regionalen und ökologischen Lebensmitteln interessiert sind[33] und es Hinweise darauf gibt, dass sich die verschiedenen Sorten hinsichtlich ihres allergischen Potenzials unterscheiden.[34] Bisher konnte der Zusammenhang zwischen Sorte und Allergenität jedoch nicht eindeutig bestätigt werden, sondern vielmehr gezeigt werden, dass die Allergenität von Äpfeln durch die Lagerungsbedingungen und -dauer beeinflusst werden.[35–37]

Der größte Preisunterschied bei Äpfeln und Apfelprodukten besteht zwischen herkömmlichen Mischapfelsäften und sortenreinen Bio-Apfelsäften. So sind Mischapfelsäfte ohne Bio Siegel bereits für 1 – 2 € je Liter erhältlich, während sortenreine Direktsäfte mit Biosiegel Preise von 2,50 – 10 €[38] je Liter erreichen können. Aufgrund dieser Unterschiede und der potenziellen Relevanz von Lebensmittelallergien ist für Verbraucher eine ausführliche Untersuchung auf Lebensmittelbetrug zu gewährleisten.

3.1.7. Spargel

Die Gattung Spargel (*Asparagus*) innerhalb der Familie der Spargelgewächse (*Asparagaceae*) enthält etwa 150 Arten von krautig wachsenden Pflanzen oder Halbsträuchern.[39]

Es gibt diverse Spargelsorten, die verschiedene Produkteigenschaften, und Erntezeitpunkte erzielen sowie unterschiedliche Ansprüche an ihre Anbaubedingungen haben. Die in der Arbeit untersuchten Sorten Backlim, Cumulus, Gijnlim und Grolim sind weit verbreitet.

Backlim ist eine Standardsorte, welche überwiegend in Gewächshäusern mit beheizten Böden angebaut wird und mit einer Pflanzdichte von 4-5 Pflanzen pro m in Reihe gute Erträge in der zweiten Hälfte der Erntesaison liefert.[40] Cumulus ist als der zarteste Spargel bekannt, der überwiegend in der ersten Hälfte der Saison einen sehr hohen Ertrag bringt.[41] Gijnlim ist eine Sorte, die sehr früh bereits hohe Erträge erzielt.[42] Grolim ist eine Hochertragssorte, welche insbesondere in gemäßigterem Klima in humösen entwässerten Böden wächst.[43]

Spargel erzielt je nach Jahr und Zeitpunkt innerhalb der Erntezeit Marktpreise zwischen 15 und 30 €/kg. Wegen der hohen Nachfrage in Deutschland wird Spargel neben dem Anbau vor Ort auch importiert. Im Vergleich zu regionalem Spargel ist dieser i.d.R. günstiger, bspw. war peruanischer und griechischer Importspargel zu Beginn der Saison 2024 für ca. 6 €/kg erhältlich, als regionale deutsche Ware noch ca. 20 €/kg kostete.[44] 2023 wurde Spargel aus Griechenland, Spanien, Italien, Mexiko, Peru und den Niederlanden importiert; dabei wird Spargel auch über die Deutsch-Polnische Grenze z.T. illegal eingeführt.[45] Das offenkundige Potential von Lebensmittelbetrug besteht darin, günstig im Ausland produzierten Spargel als deutsche Ware auszugeben.

3.2. Metabolomics

In der vorliegenden Arbeit wurden *Metabolomics*-Daten untersucht, und in diesem Kapitel soll der biologische Hintergrund (3.2.1) sowie die analytischen Verfahren, welche zur Generierung der entsprechenden Daten (3.2.2) angewendet wurden, erläutert werden.

3.2.1. Biologischer Hintergrund

Das Metabolom ist der letzte Schritt in der in Abbildung 1 dargestellten *omics*-Kaskade, und *metabolomics* beschreibt die Methoden die zu seiner Untersuchung eingesetzt werden. Da alle vorhergehenden Schritte das Metabolom somit beeinflussen, soll hier kurz auf diese eingegangen werden.

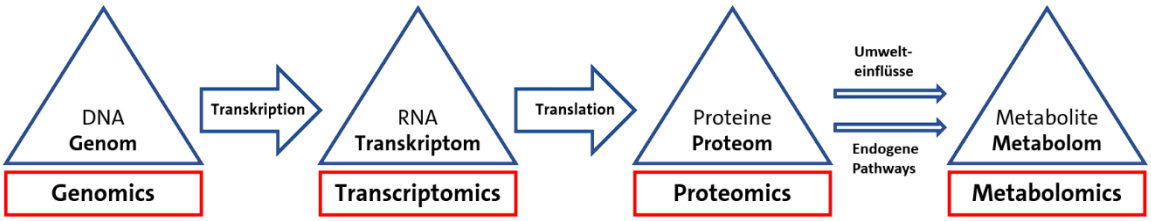


Abbildung 1: Darstellung der *omics*-Kaskade.

Genomics beschreibt die Methoden, mit denen das Genom, also die Gesamtheit der Gene mit ihren Informationen zum Bauplan eines Organismus, untersucht werden. In diesem Forschungsfeld finden verschiedenste Verfahren Anwendung wie bspw. DNA-Extraktion, diverse Sequenzierungsmethoden, Schmelzpunktuntersuchungen, Restriktionsverdau oder die Polymerase-Kettenreaktion. Diese Methoden sind oftmals Zeit- und Kosten-intensiv, weshalb *genomics*-Datensätze, besonders im Bereich der Lebensmittelanalytik, häufig aus vergleichsweise wenigen Proben bestehen.[46]

Das Proteom beschreibt die Gesamtheit der Proteine, welche in der Zelle durch die Transkription exprimierter Gene (Gegenstand von *transcriptomics*) synthetisiert werden. Im Forschungsfeld *proteomics* werden alle Proteinanalytikmethoden zusammengefasst,

die die Sequenz, Struktur und quantitativen Expression untersuchen. Da Proteine aus Aminosäureketten aufgebaut sind, kann ihre Sequenz bestimmt werden, darüber hinaus aber auch ihre inter- und intramolekularen Wechselwirkungen. Klassischerweise fand die Strukturaufklärung insbesondere durch Kristallisation mit anschließender Röntgendiffraktion statt.[47] Proteine können auch massenspektrometrisch untersucht werden, indem diese auf flüchtigen Matrizes aufgebracht unter Zufuhr von Energie mittels Laserlicht fragmentiert und vaporisiert werden.[48] Die *Deep-Learning*-Technik *AlphaFold*[49] entwickelte ein Modell, welches zur Strukturaufklärung von mittlerweile 200 Millionen Proteinstrukturen beitrug.[50]
Proteine katalysieren biochemische Reaktionen, die Umwandlung intra- und extrazellulärer Substanzen in Kaskaden von Folgeprodukten und fungieren als Signalmoleküle.[51] Diese zelleigenen Stoffwechselwege bilden den Metabolismus, die beteiligten und resultierenden Stoffe werden als Metabolite bezeichnet und ihre Gesamtheit als Metabolom.

Das Metabolom besteht aus hunderten bis mehreren tausend Stoffwechselprodukten mit einer kleineren molekularen Masse als 1500 Da.[52] Die strukturell stark unterschiedlichen Metabolite decken ein Spektrum verschiedenster Eigenschaften ab und erfüllen vielfältige biologische Funktionen. Die Konzentration der Metabolite in einem Organismus resultiert aus dem Zusammenspiel von Genom, Transkriptom und Proteom. Diese Kaskade ist jedoch nicht nur endogen gesteuert, sondern auch dadurch, dass sich Organismen an ihre Umwelt anpassen.[53] Deswegen können auch exogene Faktoren einen starken Einfluss auf das Metabolom haben. Die externen Faktoren können klimatischer, geographischer oder ökologischer Natur sein. Zu den klimatischen Effektoren zählen die Menge an zugeführtem Licht, die Außen- und Bodentemperatur, sowie die Menge an Niederschlag. Geographische Einflüsse sind u.a. die Nähe zu großen Gewässern und die Lage in Höhenmetern. Ökologische Einflüsse stellen Faktoren wie Bodenbeschaffenheit, die Anwesenheit anderer Organismen und die Menge an verfügbaren Nährstoffen dar.[53]

Das Metabolom repräsentiert somit umfassend den Phänotyp eines Organismus und *metabolomics*-Untersuchungen bieten den Vorteil gegenüber bspw. *genomics*-Untersuchungen, dass sie Umwelteinflüsse oder andere externe Stressfaktoren umfassend berücksichtigen.[54] Die chemische Analyse der vielfältigen Metabolite ist offenkundig äußerst komplex, nicht nur auf Seiten der eingesetzten Analyseverfahren, sondern auch in Bezug auf die Interpretation der großen Datensätze.[8]

3.2.2. Analytische Verfahren zur Metabolomanalyse

Folgerichtig werden für die Metabolom-Untersuchung Hochdurchsatz-Analytik Methoden mit statistischer bioinformatischer Auswertung kombiniert, um qualitative und quantitative Aussagen über Zwischen- und Endprodukte treffen zu können.[53,55] In der Humanmedizin identifiziert man auf diese Weise krankheitsrelevante Markermoleküle[56] oder verfolgt die Wirkung therapeutischer Maßnahmen[57].

In der Lebensmittelanalytik wird *metabolomics* dafür eingesetzt, Lebensmittel basierend auf einem metabolischen Fingerabdruck bezüglich Eigenschaften, wie Art, Sorte, Herkunft oder Anbaubedingungen zu authentifizieren. Da Metabolite sehr unterschiedliche Eigenschaften aufweisen, werden dazu verschiedene Verfahren angewendet, am häufigsten LC-MS Methoden[58–60], NMR-[61–63] sowie die Fourier-Transform-Infrarotspektroskopie (FTIR)[4,62,64]. Im Rahmen dieser Arbeit wurde die ^{1}H NMR-Spektroskopie und LC-MS verwendet, weshalb diese Methoden hier kurz erläutert werden sollen.

3.2.2.1. ^{1}H Kernresonanz-Spektroskopie

Die NMR-Spektroskopie, basiert auf der Umkehr des Eigendrehimpulses (engl.: *spin*) von erregbaren Atomkernen unter Einfluss von Radiowellen in einem stationären Magnetfeld.[65] Die geläufigsten Atomkerne mit einem Eigendrehimpuls, welche mit NMR-Spektroskopie untersucht werden, sind ^{1}H, ^{13}C, ^{15}N und ^{35}S.[65] Es können aber noch weitere Kerne mit NMR-Spektroskopie untersucht werden, jedoch fokussierte die

vorliegende Arbeit auf ^{1}H NMR-Spektren, weshalb sich dieses Kapitel auf diese Untersuchungen konzentriert.

Der Aufbau eines NMR-Spektroskops ist schematisch in Abbildung 2 dargestellt.

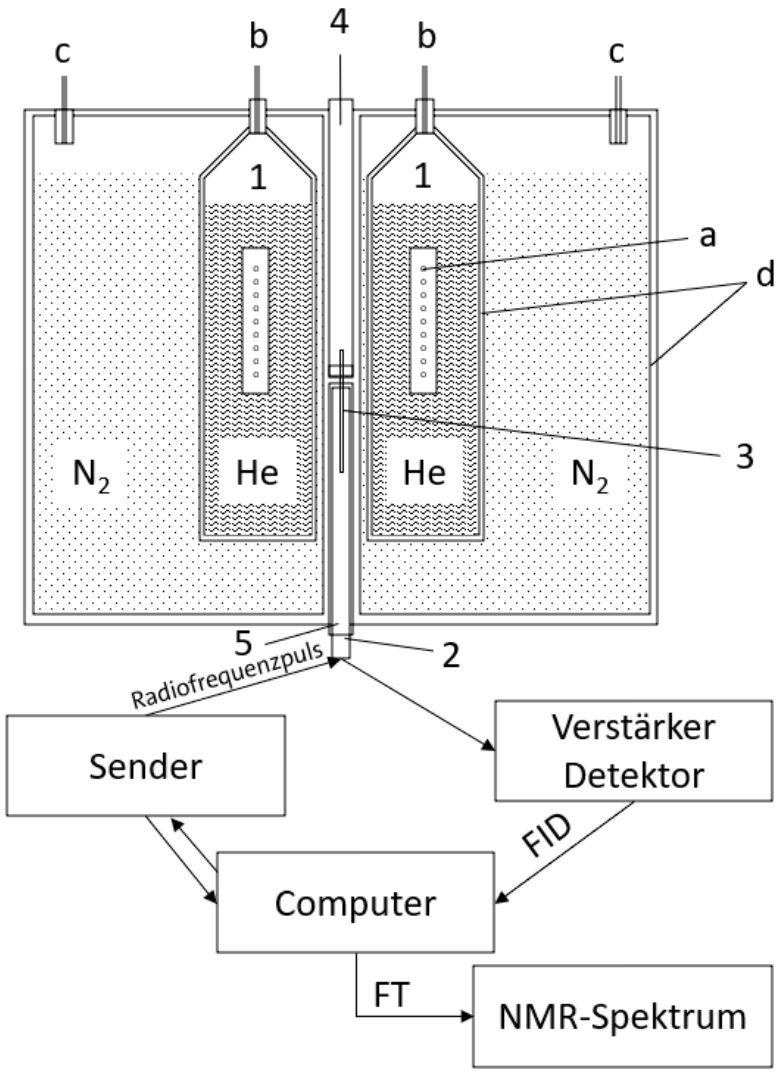


Abbildung 2: Schematischer Querschnitt eines NMR-Spektroskops mit Kryomagnet mit 1: Magnet mit Magnetspulen (a) mit Einfüllstutzen für flüssiges Helium (He) (b) und flüssigen Stickstoff (N_2) (c) und innerer und äußerer Vakuumkammer (d); 2: rotierender Probenkopf; 3: Probenröhrchen; 4: Probenwechsler; 5: *Shim*-Einheit.[66]

In einem mit flüssigem Helium gefüllten Gefäß befindet sich eine Spule, welche unter Spannung gesetzt wird. Diese ist auf etwa 4 K gekühlt und wird zusätzlich von außen mit flüssigem Stickstoff gekühlt. Die Spule umgibt einen sich rotierenden Probenkopf, in

welchen ein Glasröhrchen gestellt wird.[66] Dieses enthält die zu untersuchende Probenlösung. In einem Supramagnetfeld richten sich die Kerne entsprechend ihres *spins* entlang des Magnetfeldes aus. Die *Shim*-Einheit gleicht Magnetfeldschwankungen aus. Über einen Radiofrequenzemitter, die Sendereinheit, wird nun die Probe mit einem Radiofrequenzpuls angeregt. Der Detektor misst anschließend das Abklingverhalten der angeregten Kerne als *free ion decay*-Spektrum, kurz FID-Spektrum. Dieses FID-Spektrum wird von einem Computer einer Fouriertransformation (FT) unterzogen und es resultiert das NMR-Spektrum.[66] In einem NMR-Spektrum wird die Intensität gegen die chemische Verschiebung aufgetragen. Einer Probe wird vor einer Messung i.d.R. ein interner Standard hinzugeben, dessen resultierendes NMR-Signal für verschiedenste Normalisierungsschritte verwendet wird.[66]

Die NMR-Spektroskopie wird in Metabolom-Analysen meist als ^{1}H NMR-Spektroskopie eingesetzt. Sie dient zur Strukturaufklärung organischer Substanzen, denn aus der chemischen Verschiebung der Protonenbanden und aus der Multiplizität ihrer Signale kann man auf deren räumliche Nähe schließen; überdies reflektiert das Integral die Zahl der Protonen und die Konzentration des Moleküls.[67] Bei größeren Molekülen kombiniert man die Methode mit MS, welche die Molekülmasse und Fragment-Muster ergibt. In Kombination von MS mit den Protonenresonanzdaten gelingt in der Praxis oftmals die angestrebte Substanz-Identifikation und im Falle nicht zu komplexer Mischungen auch die Quantifizierung der Metabolite.[67]

In *metabolomics*-Untersuchungen werden eine Vielzahl von Substanzen gleichzeitig gemessen werden, weswegen NMR-Spektren in diesem Forschungsfeld hoch komplex ausfallen.[68–72] Ebenfalls werden insbesondere für Klassifizierungsfragestellungen in der Lebensmittelüberwachung eine große Anzahl an Proben gemessen, die alle jeweils ein individuelles NMR-Spektrum hervorbringen. Für eine reproduzierbare Untersuchung müssen daher Spektren in mehreren, in hier nicht näher aufgeführten Schritten normalisiert werden.[73]

Anschließend findet das sog. *bucketing* statt, im Zuge dessen man NMR-Spektren in einzelne Spektralabschnitte unterteilt und das Integral jedes einzelnen Abschnittes berechnet. So werden die Spektren zum einen vergleichbarer, zum anderen dient *bucketing* der Datenreduktion, da NMR-Spektren aus mehreren tausend Datenpunkten zusammengesetzt sind.[74] Diese Reduktion ist für die Auswertung mit zeit- und rechenintensiven statistischen Methoden erstrebenswert. Da allerdings *bucketing* lediglich ein einzelnes Integral für mehrere Datenpunkte ergibt, somit die Informationen mehrerer Signale zusammenfasst, geht das *bucketing* immer auch mit einem Verlust an Informationen einher. Es ist daher bei der Untersuchung von in *buckets* unterteilten NMR-Daten immer empfehlenswert, zusätzlich die korrespondierenden NMR-Spektren zu berücksichtigen, damit Signale relevanten Spektralbereichen zugeordnet werden können.[75,76]

^{1}H NMR-Spektren können bei *metabolomics*-Untersuchungen verschieden ausgewertet werden. Für eine Klassifizierung von Proben sieht man ein Spektrum oft als Fingerabdruck an, ohne dass Informationen über die vorhandenen Metabolite existieren. Anhand von Spektralbereichen, die für Klassenunterschiede relevant sind, kann dann eine gezielte Metabolit-Identifikation erfolgen.[77]

Da die Komplexität von *metabolomics*-Datensätzen wegen der großen Zahl der vorhandenen Metabolite sehr stark erhöht ist, gelingt eine Metabolit-Identifikation allein anhand von Kopplungskonstanten und Datenbankabgleich nur selten, zumal sich die Signale mehrerer Metabolite meist überlagern. Für die Zuordnung der Signale zu Metaboliten sind zweidimensionale NMR-Experimente hilfreich, welche die Korrelationsmuster verschiedener Signale darstellen. Dabei werden die Korrelationsspektroskopie (co*rrelation spectroscopy*, COSY) und die vollständige Korrelationsspektroskopie (*total correlation spectroscopy*, TOCSY) angewendet.[78]

Das zweidimensionale COSY-Spektrum bildet auf der X- und Y-Achse jeweils das ^{1}H NMR-Spektrum der Probe ab. Auf der Diagonalen wird die sogenannte Selbstkorrelation jedes

einzeln angeregten Protons dargestellt. Zusätzlich erhält man die sogenannten *off-diagonal peaks*, aus deren Position man korrelierende Protonensignale ableitet. Jeder dieser *off-diagonal peaks* stellt eine direkte Kopplung zwischen zwei Protonen dar, was auf ihre unmittelbare chemische Nähe innerhalb eines Moleküls hinweist.[79]

Auch beim TOCSY Spektrum handelt es sich um ein zweidimensionales Spektrum, welches entlang der Diagonale alle angeregten Protonen enthält. Zusätzlich zu den *off-diagonal peaks* der benachbarten Protonen sind hier ebenfalls Signale der nicht-benachbarten Protonen sichtbar, sofern sich diese im selben *spin*-System befinden.[80]

Für Strukturaufklärungen insbesondere komplexerer Metabolite in *metabolomics*-Datensätzen bietet TOCSY den Vorteil gegenüber COSY, dass bedeutend mehr Korrelationsmuster dargestellt und entsprechend leichter eine Zuordnung zu einer Einzelsubstanz stattfinden kann. Zwar bildet COSY übersichtlichere Spektren ab, die Auswertungsmöglichkeiten sind jedoch begrenzt.[81] Bei der Untersuchung von *metabolomics*-Daten ist es daher erstrebenswert, TOCSY-Experimente durchzuführen. Beispielhaft sind in Abbildung 3 jeweils ein COSY- und TOCSY-Spektrum einer Mischung aus 2-Pentanal und 5-Epoxyhexen nach Paul et al. dargestellt.[82]

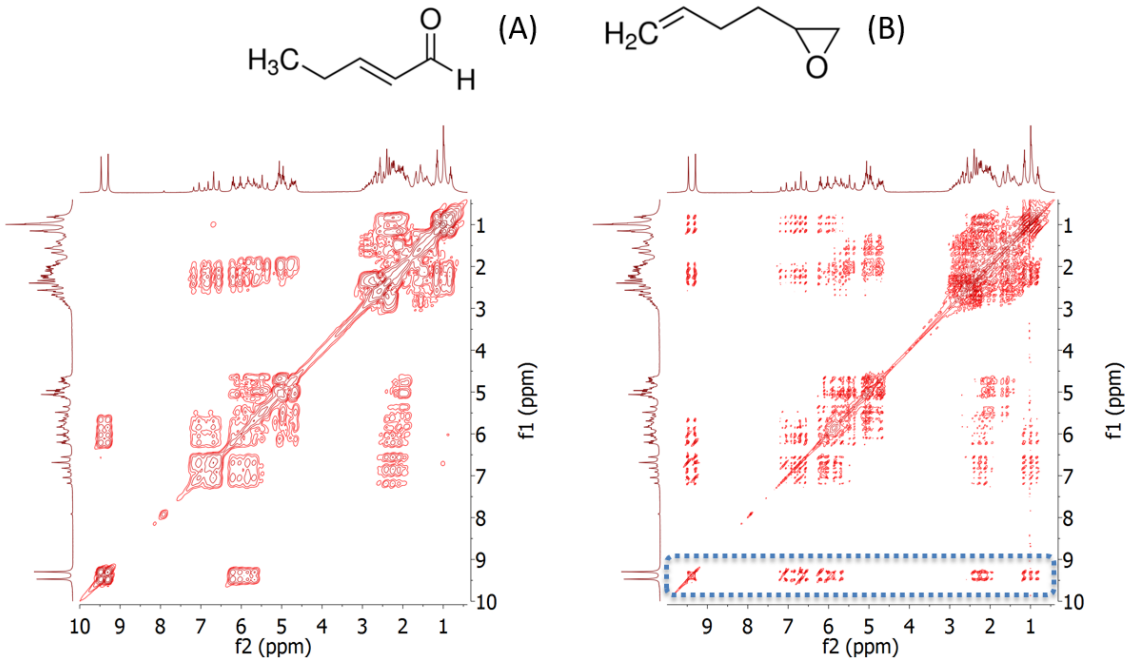


Abbildung 3: COSY(links) und TOCSY- Spektrums (rechts) einer Mischung aus 2-Pentanal und 5-Epoxyhexen nach Magritek und Paul et al.[82]

Für die Untersuchung von *metabolomics*-Daten kann zusätzlich das Verfahren der statistischen vollständigen Korrelationsspektroskopie, engl. *statistical total correlation spectroscopy, STOCSY,* hilfreich sein, da der Informationsgewinn verglichen zu COSY und TOCSY nicht auf intramolekulare Wechselwirkungen begrenzt ist. Vielmehr kann eine STOCSY auch die intermolekularen Verknüpfungen von Metaboliten darstellen, die bspw. an demselben Stoffwechselweg beteiligt sind.[83] Hierfür wird die Korrelation aller Variablen aller Spektren berechnet und eine einzige Korrelationsmatrix erhalten. Diese wird anschließend grafisch dargestellt, weswegen aus allen vorhandenen Spektren eine einzige STOCSY-Grafik erhalten wird. Somit sorgt die STOCSY im Vergleich zu COSY und TOCSY zeitgleich auch für eine Reduktion des analytischen Aufwands. In NMR-Spektren sind jedoch viele Signale mit niedrigen Intensitäten vorhanden, welche durch vorhandenes Rauschen auch zufällig korrelieren können. Deswegen ist es zielführend bei Metabolomics Analysen, eine multivariate Datenanalysemethode vorzuschalten, um klassifikationsrelevante Signale zu selektieren.[83] Beispielhaft sind in Abbildung 4 die Ergebnisse einer STOCSY-Analyse nach Cloarec et al. dargestellt.[83]

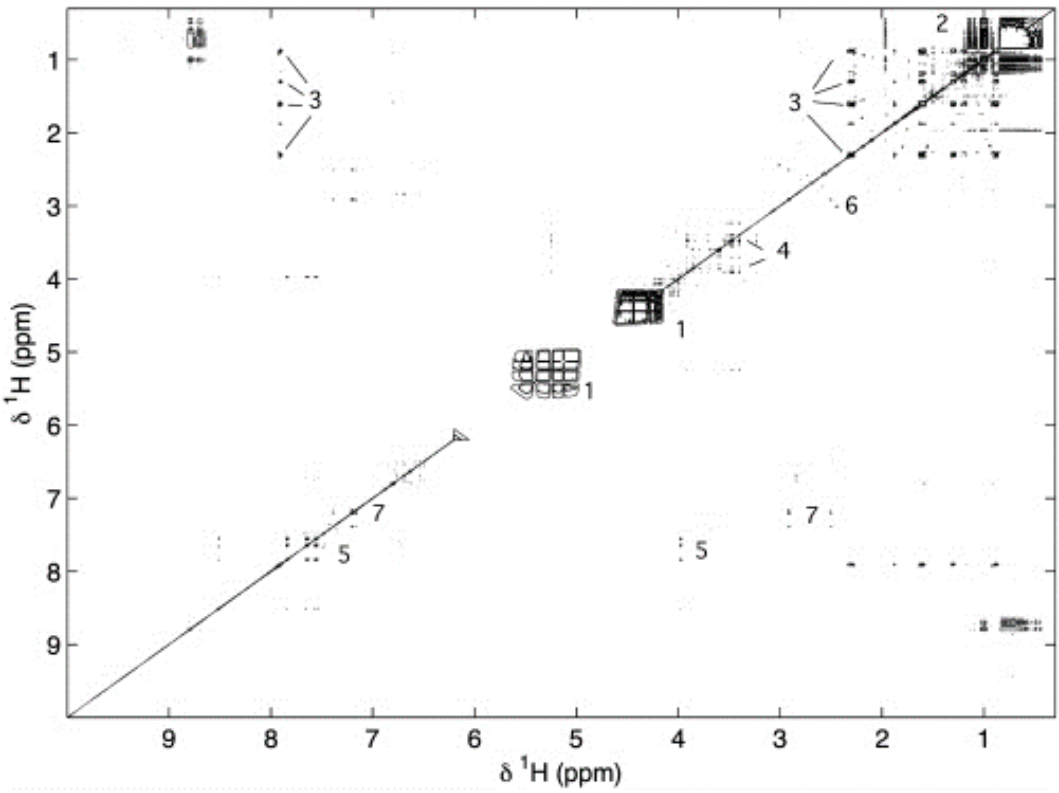


Abbildung 4: Beispielhafte Darstellung einer zweidimensionalen STOCSY Auswertung von ^{1}H NMR Spektren von drei verschiedenen Mausarten nach Cloarec et al. mit (1) Verunreinigung durch Wassersignalunterdrückung, (2) Proteinen, (3) Valeramid, (4) Glucose; (5) Hippurat, (6) 2-Oxoglutarat und (7) 3-Hydroxyphenylpropionat.[83]

Für eine sichere Identifizierung von Metaboliten in einem NMR Spektrum einer komplexen Probe ist der Abgleich der Signale mit denen in Datenbanken oder selbst durchgeführten Referenzmessungen und/oder die Anwendung von zweidimensionalen NMR Experimenten nicht ausreichend. Dafür kann die potentiell identifizierte Substanz gezielt zur untersuchten Probe in einem sogenannten *spike-in*-Experiment zugegeben werden. Wenn die entsprechenden Signale im NMR-Spektrum ansteigen, gilt der entsprechende Metabolit als nachgewiesen.[83]

3.2.2.2. Flüssigchromatographie gekoppelt mit Massenspektrometrie

Die Flüssigchromatographie gekoppelt mit Massenspektrometrie, engl.: *liquid chromatography coupled with mass spectrometry*, LC-MS, basiert auf der

flüssigchromatographischen Auftrennung einer Probe in ihre Einzelkomponenten mit einer anschließenden Detektion an einem Massenspektrometer.[84]

Der schematische Aufbau eines LC-MS ist in Abbildung 5 dargestellt.

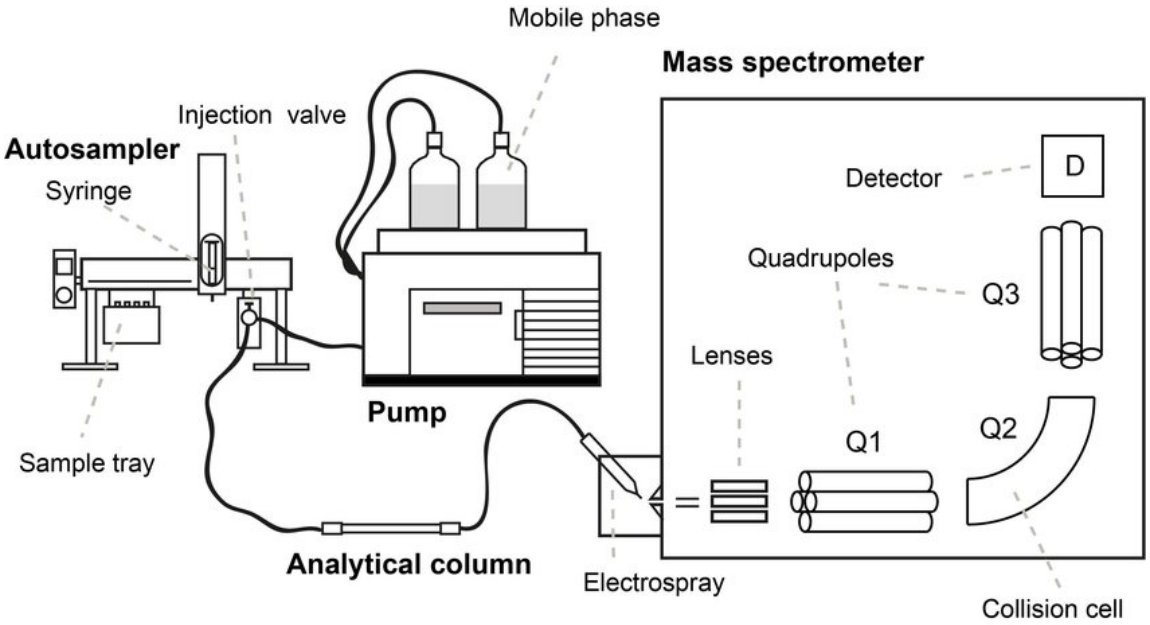


Abbildung 5: Schematischer Aufbau eines LC-MS mit einem dreifach gekoppelten Quadrupol nach Östman et al.[85]

Die LC ist eine Verteilungschromatographie, welche Substanzen anhand ihrer chemischen und physikalischen Eigenschaften verschieden stark an einer stationären Phase retardiert, während bei konstantem Fluss ein flüssiges Lösungsmittelgemisch als mobile Phase agiert. Durch Änderung der Polaritätsverhältnisse im Lösungsmittel werden während einer Messung an die stationäre Phase gebundene Analyten abgelöst und erreichen zeitlich verzögert einen Detektor. In der LC werden verschiedene Detektoren verwendet, wie bspw. die Absorption ultravioletter Strahlung eines Analyten oder ein entstehendes Signal an einem MS.[84,86]

Ein MS erlaubt, die Masse von Ionen zu bestimmen und ist i.A. aus vier Untereinheiten aufgebaut; der Eingabe, der Ionisationsquelle, dem Massenanalysator und einem Massendetektor.[87]

Als Eingabe wird in der LC-MS die LC verstanden, da sie eine Probe in ihre Einzelkomponenten aufgetrennt an das MS weitergibt. Die Ionisationsquelle ionisiert ein Molekül, damit dieses vom MS erfassbar wird. Hierbei können verschiedene Ansätze verwendet werden, als Schonendste wird allerdings die Elektronensprayionisation, ESI verstanden. Im Vergleich zu anderen Ionisationsmethoden sorgt sie für eine nur geringe Fragmentierung der Analyten, wodurch Analyten kaum zersetzt den Detektor erreichen. Insbesondere in *metabolomics*-Untersuchungen wird die ESI-Methode am häufigsten angewendet, da die erhaltene Molekülmasse und die Massen der größten Fragmente wertvolle Informationen über den Analyten gibt.[88] Als Massenanalysator können ebenfalls verschiedene Methoden verwendet werden. Am häufigsten werden hierbei Quadrupole verwendet. Diese sind aus vier Metallstangen aufgebaut, welche verschieden unter Spannung gesetzt sind. Durch das entstehende elektromagnetische Feld werden Ionen in Spiralbahnen beschleunigt und erreichen dann den Detektor, welcher letztendlich die Masse eines Ions als Masse zu Ladungsverhältnis (m/z) aufnimmt. Auch wird die Flugzeit eines Ions gemessen, engl.: *time off light*, ToF. Es basiert auf der Eigenschaft, dass schwerere Ionen eine längere Zeit für eine feste Distanz brauchen als Leichtere und dass mehrfach geladene Ionen stärker in einem elektromagnetischen Feld beschleunigt werden als einfach Geladene. Die Detektion findet in der MS anhand der Messung von Strömen statt, die ein Ionenpaket bei Erreichen erzeugt.[87]

LC-MS ist ein oft angewendetes Verfahren für Untersuchungen des Metaboloms von Lebensmitteln.[89] Die Kombination des ermittelten m/z Verhältnisses mit der Retentionszeit und der Signalintensität der Detektion charakterisiert die einzelnen Metabolite. Die Kombination der Informationen aller analysierten Metabolite ermöglicht eine detaillierte Analyse der Zusammensetzung der biologischen Proben, bspw. von Lebensmittelextrakten.[89] Für die Metabolit-Identifikation wird in *metabolomics*-Untersuchungen ein LC-MS mit einem weiteren MS kombiniert (LC-MS/MS), um eine stärkere Fragmentierung eines Metabolit-Ions zu verursachen. Hierfür

werden bspw. drei in Reihe geschaltete Quadrupole derart konfiguriert werden, dass der erste als Filter für das Metabolit-Ion agiert, der Zweite der weiteren Fragmentierung des Ions dient und der Dritte für eine m/z-spezifische Auftrennung der entstandenen Fragment-Ionen sorgt. Aus dem entstehenden MS-Spektrum lassen sich dann einzelne Gruppen und Seitenketten eines Metaboliten identifizieren und auf die Struktur des zu untersuchenden Ziel-Metaboliten schließen.[90]

3.3. Maschinelles Lernen

Bei multidimensionalen Datensätzen, wie sie typischerweise in *metabolomics*-Untersuchungen vorkommen, stoßen univariate statistische Ansätze an ihre Grenzen. Dies ist insbesondere der Fall, wenn es darum geht, hochkomplexe Zusammenhänge darzustellen.[91] Im Allgemeinen werden für diese Fragestellungen multivariate Verfahren angewendet, welche auch als maschinelle Lernmethoden bezeichnet werden. Dieses Kapitel soll daher einen Einblick in die Grundlagen des maschinellen Lernens geben (3.3.1) und auf die multivariaten Verfahren der PCA (Abschnitt 3.3.2) und RF (3.3.3), sowie allgemein auf die RF basierte Analyse der *variable importance* (3.3.4), der Variablenselektion (3.3.5) und die Verfahren Boruta (3.3.6) und SMD (3.3.7) eingehen.

3.3.1. Grundlagen

Es wird zwischen überwachten (*supervised*), unüberwachten (*unsupervised*) und halbüberwachten (*semi-supervised*) maschinellen Lernalgorithmen sowie dem *reinforcement learning* unterschieden.[92]

Bei überwachten Methoden wird ein Modell basierend auf den *input* bzw. unabhängigen Variablen trainiert, welche die *output* oder abhängigen Variablen vorhersagen sollen. Hierfür werden für den Trainingsprozess Daten bereitgestellt, bei denen die entsprechende Zuordnung bekannt ist, diese also mit einem sog. *label* versehen sind. Innerhalb der unabhängigen Variablen werden Muster oder einzelne Marker erkannt, die spezifisch für Vorhersage der abhängigen Variablen sind. Im Allgemeinen spricht man dann von einem überwachten Ansatz, wenn der Hauptfokus auf dem Beantworten

einer konkreten Fragestellung liegt.[93] Dabei kann zwischen Klassifikations- oder Regressionsanalysen unterschieden werden. Bei Klassifikationsfragestellung wird eine kategorielle *output* Variable basierend auf den unabhängigen Variablen vorhergesagt und ein konkretes Beispiel dafür ist die Vorhersagte des Herkunftslandes oder der Sorte von Lebensmitteln .[59,94] Bei Regressionsfragestellungen wiederum werden quantitative *outcome* Variablen vorhergesagt. Ein Beispiel im Zusammenhang der Lebensmittel-analyse ist die Vorhersage der Frische von Fisch.[95]

Bei allen diesen Ansätzen ist die Validierung des entstandenen Modells notwendig, welche mit einem unabhängigen Datensatz durchgeführt wird. Hierfür wird vor dem Training eines Modells ein repräsentativer Teil des Datensatzes nicht in den Trainingsprozess einbezogen, um mit diesem eine Validierung durchzuführen.[96] Auch können mehrere Modelle mit unterschiedlichen Anteilen der Datensätze trainiert werden und die nicht verwendeten Anteile anschließend für die Validierung.[97]
Bei unüberwachten maschinellen Lernmethoden wird, im Gegensatz zu überwachten Verfahren, ein Modell ohne Einbeziehung der abhängigen Variablen und damit ohne externe Anreize trainiert.[98] Es können daher aussagekräftige Trends oder Strukturen und Gruppierungen für explorative Zwecke in den Daten identifiziert werden. Unüberwachte Verfahren werden für Clusteranalysen, Dichteabschätzungen oder Dimensions-reduktion, sowie Assoziationsregeln und Ausreißer-Detektion verwendet. Cluster-analysen sind hilfreich, um ähnliche oder abweichende Proben zu identifizieren oder auch Eigenschaften von unabhängigen Variablen miteinander zu vergleichen.[93] Halbüberwachte ML Methoden können als eine Kombination von überwachten und unüberwachten Methoden verstanden werden. Sie sind sowohl für den Umgang mit gelabelten als auch nicht-gelabelten Daten geeignet. Gelabelte Daten kommen in Echtweltszenarien seltener vor als nicht-gelabelte Daten. Das Ziel halbüberwachter Verfahren ist ein besseres Vorhersagemodell im Vergleich zu überwachten Ansätzen. Anwendungsbereiche sind u.a. bei maschineller Übersetzung von Texten, Betrugs-erkennung und Kennzeichnung unbekannter Daten in einem beschrifteten Datensatz,

sowie Textklassifizierungen.[93,99]

Das *reinforcement learning* ist ein umgebungsgesteuerter Ansatz. Der Algorithmus passt sich an bestimmte Fragestellungen an und identifiziert jene Sequenz, die das beste Ergebnis erzielt. Es basiert auf einem Belohnungssystem, was nach dem „*trial and error*"-Prinzip den analytischen Fortschritt einer Fragestellung überwacht. Hierbei wird ein Fortschritt für die Beantwortung der Fragestellung als Belohnung angesehen (bspw. eine höhere Vorhersagegenauigkeit), während kein Fortschritt als Strafe angesehen wird. Der Algorithmus ist darauf ausgelegt, die Belohnungspunkte zu maximieren und die Strafpunkte zu minimieren. Die Implementierung solcher Ansätze ist jedoch herausfordernder als die Anwendung von überwachten oder unüberwachten Algorithmen.[100,101] Es wird überwiegend bei der Automatisierung von Robotiksystemen, Kraftfahrzeugen, Herstellungsprozessen und in der Logistik von Versorgungsketten angewendet.[102]

Eine häufige Vorgehensweise bei der Analyse von *metabolomics* Daten ist die Anwendung sowohl von unüberwachten als auch überwachten ML Ansätzen. Ein unüberwachtes Verfahren wird typischerweise dafür eingesetzt, zu analysieren ob die gefundene Klassenzugehörigkeit bereits aus der Untersuchung der Hauptunterschiede des Datensatzes ersichtlich wird. Falls dies der Fall ist, entfällt oftmals die weitere Untersuchung mit überwachten Verfahren, da unüberwachte Verfahren i.d.R. weniger rechenaufwändig und somit zeitsparender sind. Sollte die Hauptvarianz der Proben nicht den abhängigen Variablen entsprechen, wird auf überwachte Verfahren zurückgegriffen, um eine genaue Vorhersage zu erreichen. Halbüberwachte Verfahren werden angewendet, wenn entsprechende Metainformationen nicht für alle Proben vorliegen. Erste Ansätze existieren bereits, bei denen *reinforcement learning* für *metabolomics*-Fragestellungen angewendet wird.[103,104]

3.3.2. Hauptkomponentenanalyse

Die PCA ist ein unüberwachtes ML Verfahren und ein Standardwerkzeug in der modernen Datenanalyse. Die PCA ist in der Lage, versteckte Strukturen und Zusammenhänge innerhalb der Daten hervorzuheben, ohne dabei durch Rauschen beeinflusst zu werden[105], indem die größten Unterschiede durch eine Dimensionsreduktion dargestellt werden. Dies geschieht durch eine Transformation der unabhängigen Variablen in sogenannte Hauptkomponenten, die voneinander unabhängig sind und jeweils die größtmögliche Varianz darstellen. Diese Hauptkomponenten können als Richtungsvektoren gesehen werden, die orthogonal zueinander sind und an denen sich die Daten der Proben wie in einem Koordinatensystem anordnen.[105] Die Anwendung der PCA lässt sich in mehrere Schritte unterteilen: die Vorverarbeitung, die Bildung der Kovarianzmatrix und schließlich die Bestimmung der Hauptkomponenten. Im ersten Schritt wird eine Zentrierung der Daten durchgeführt, bei der von allen *input* Variablen der jeweilige Mittelwert subtrahiert wird. Dieser Schritt ist wichtig, damit die erste Hauptkomponente nicht primär durch die Mittelwerte der Variablen beeinflusst wird, sondern die gewünschten relevanten Unterschiede innerhalb der Daten widerspiegelt.[105] Zusätzlich dazu kann eine Skalierung der Daten durchgeführt werden, bei der die zentrierten Variablen durch die Standardabweichung dividiert werden. Dies führt dazu, dass jede Variable gleichermaßen die Ergebnisse der PCA beeinflusst.

Nach der Datenvorbereitung wird im zweiten Schritt die Kovarianzmatrix berechnet, welche den Zusammenhang der Variablen untereinander repräsentiert: Hierbei bedeutet ein positiver Wert der Kovarianz zweier Variablen, dass sie in die gleiche Richtung variieren.[105] Die Kovarianzmatrix C mit n Proben wird berechnet nach:

$$C = \frac{1}{n\text{-}1} X^T \cdot X$$

mit:

X – standardisierte Datenmatrix X

X^T – transponierte Datenmatrix X

Bei der PCA werden nun im dritten Schritt die Eigenvektoren und Eigenwerte der Kovarianzmatrix berechnet, um die Hauptkomponenten zu bestimmen. Die Eigenvektoren geben dabei die Richtung der jeweils maximalen Varianz der Daten und damit der neu gebildeten Hauptkomponenten an, während die Eigenwerte die Größe der Varianz in dieser Richtung darstellen. Dabei werden mehrere Hauptkomponenten berechnet, deren Anzahl oft so gewählt wird, dass sie einen großen Anteil der Varianz repräsentieren (üblicherweise entsprechend 90-95% der Gesamtvarianz). Anschließend werden die Daten auf die neu gebildeten Hauptkomponenten projiziert, indem die zentrierte Datenmatrix mit den ausgewählten Eigenvektoren multipliziert wird. Die neu gebildeten Hauptkomponenten sind aufgrund ihrer Orthogonalität voneinander unabhängig und repräsentieren zeitgleich die größte Varianz der ursprünglichen Daten.[105]

Zwei zentrale Begriffe sind für die Interpretation von Ergebnissen der PCA relevant: die *scores* und die *loadings*. Die *scores* repräsentieren die transformierten Daten und entsprechen den Koordinaten einer Probe bezüglich der Hauptkomponenten. Die *loadings* entsprechen den jeweiligen Eigenvektoren der Kovarianzmatrix und geben somit an, wie stark jede der ursprünglichen Variablen zur Bildung der jeweiligen Hauptkomponente beiträgt. Sowohl *scores* als auch *loadings* werden als Ergebnis einer PCA graphisch dargestellt und interpretiert. Die *scores* werden üblicherweise in eine Abbildung dargestellt, welcher die Werte zweier oder dreier Hauptkomponenten enthält, um Gruppierungen innerhalb der Daten aufzuzeigen. Proben, die sich hier sehr weit von allen anderen Proben befinden, können dabei als Ausreißer identifiziert werden. Im Bereich der *Metabolomics* Analysen von Lebensmitteln könnten diese bspw. auf eine verdorbene Probe hinweisen. Die Auftragung der *loadings* geschieht in

Zusammenhang mit den unabhängigen Variablen oder den *scores*, um den Einfluss der einzelnen Variablen auf die PCA zu analysieren.[105]

3.3.3. Random Forest

In der vorliegenden Arbeit wurde der multivariate, überwachte maschinelle Lernalgorithmus RF zur Klassifizierung von Lebensmitteln angewendet. Dieser basiert auf einer Vielzahl binärer Entscheidungsbäume, die unabhängig voneinander trainiert werden und deren Funktionsweise zunächst erläutert werden soll. Ein Entscheidungsbaum ist eine Anordnung hierarchischer Entscheidungen, bei denen die Proben durch die Festlegung eines spezifischen Splits, d.h. einer Splitvariable und eines Grenzwertes, jeweils in zwei Gruppen eingeteilt werden. Diese Gruppen sollten möglichst rein sein, d.h. möglichst ausschließlich Proben aus einer der vordefinierten Klassen enthalten.[106] Es werden im Zusammenhang mit RFs die Begriffe *root node*, *parent node*, *child node* und *leaf node* verwendet. Die *root node* ist jener Knotenpunkt, an dem der erste Split stattfindet. Sie ist zeitgleich die erste *parent node* für die darunterliegenden *child nodes*. Die *child nodes* sind immer einer einzelnen *parent node* untergeordnet. Am Ende eines Entscheidungsbaum befinden sich die *leaf nodes*, an welchen keine weiteren Knotenpunkte gebildet werden und eine Zuordnung der Proben zu den Klassen erfolgt.

Der schematische Aufbau eines Entscheidungsbaumes eines RF ist in der Abbildung 6 dargestellt.

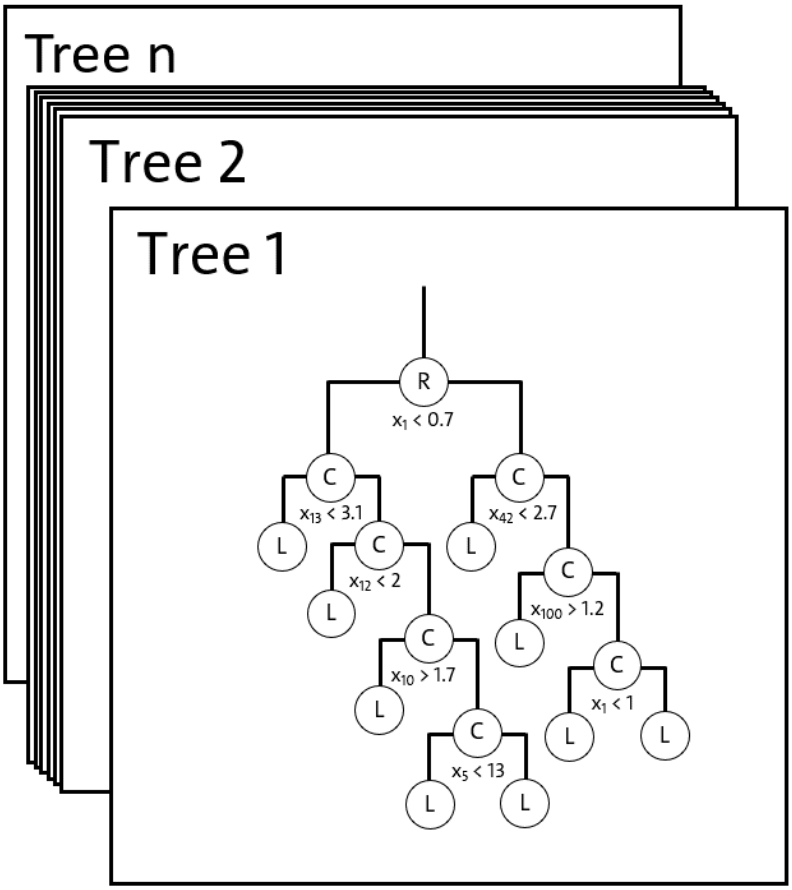


Abbildung 6: Schematische Darstellung eines RF mit R: *root node*, C: *child node* und L: *leaf node*.

Um jeweils den bestmöglichen Split zu erreichen, wird die *Gini impurity* G berechnet, welche ein Maß für die Reinheit eines Knotenpunktes darstellt.[107,108] Diese summiert die Reinheit der einzelnen Klassen p_i auf, welche durch den Quotienten der Anzahl an Proben dieser Klasse n_i und der Gesamtprobenzahl n im Knotenpunkt t berechnet wird. Die *Gini impurity* kann Werte zwischen 0 und 0.5 annehmen, wobei ein Wert von 0 bedeutet, dass alle Proben an einem Knotenpunkt zu einer Klasse gehören. Sie wird berechnet nach[107]:

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2$$

mit:

t – *node* im Entscheidungsbaum

C – Anzahl der Klassen

p_i – Reinheit der Klasse *i*

Um den optimalen Split für einen Knotenpunkt zu bestimmen, wird der *Gini gain* ΔG berechnet, welcher die *Gini impurity* vor (G(t)) und nach dem entsprechenden Split, d.h. für die *parent* und die *child node*, verwendet:[106,107]

$$\Delta G = G(t) - \frac{n_L}{n} G(t_L) + \frac{n_R}{n} G(t_R)$$

mit:

$t_{L,R}$ – linke bzw. rechte *child node*

n_L, n_R – Anzahl der Proben der linken bzw. rechten *child node*

Der *Gini gain* wird nun für jeden möglichen Split, d.h. alle Variablen und Grenzwerte, berechnet und der ausgewählt, welcher den größten Wert aufweist. Dieser Prozess wird dann jeweils mit jeder der *child nodes* wiederholt, bis ein spezifisches Kriterium erfüllt ist. Meistens ist dieses durch den Parameter der *minimal node size* definiert, was bedeutet das nur noch eine festgelegte Anzahl an Proben im Knotenpunkt vorhanden ist.[106]

Bei einem RF wird nun eine Vielzahl von Entscheidungsbäumen trainiert, wobei möglichst unterschiedliche Bäume erhalten werden sollen. Dies wird einerseits dadurch erreicht, dass nur eine zufällig ausgewählte Teilmenge der abhängigen Variablen als potentielle Split-Variablen ausgewählt wird. Die Größe dieser Teilmenge wird über den Parameter *mtry* festgelegt.[106] Andererseits wird für jeden Baum auch eine zufällig selektierten Teilmenge der Trainingsdaten verwendet. Diese Proben werden durch

bootstrapping, d.h. zufällig mit Zurücklegen gezogen, sodass eine Probe auch mehrmals im Trainingsdatensatz eines Baumes vorkommen kann. Da bei einem RF eine Vielzahl von Entscheidungsbäumen, deren Anzahl mit dem Parameter *ntree* festgelegt werden, mit verschiedenen *bootstrap*-Proben aggregiert werden, spricht man in diesem Zusammenhang auch von *bootstrap aggegation* oder *bagging*.[109]

Bei der Anwendung eines RF zur Klassifikation wird für jeden Baum eine Klassifizierungsentscheidung getroffen und diese gemittelt, um die Entscheidung des RF zu bestimmen. Ein wesentlicher Vorteil von RF ist, dass für jeden Baum ca. 37% der Proben im Trainingsdatensatz vorhanden sind, welche beim *bootstrapping* nicht gezogen wurden und somit für eine unabhängige, interne Validierung verwendet werden können.[106] Der Vorhersagefehler wird dabei als *out-of-bag*-Fehler (OOB) bezeichnet und für jeden Baum und den gesamten RF bestimmt. Dieser Fehler ist besonders dann von Vorteil, wenn auf eine Optimierung der Parameter *ntree*, *mtry* und *minimal node size* verzichtet wird. Dieser ist somit vergleichbar mit der Vorhersage eines Testdatensatzes. Bezüglich der Parameter hat sich gezeigt, dass eine Optimierung einen vergleichsweise kleinen Einfluss auf die Leistungsfähigkeit eines RF hat und diese für *ntree* ab einem bestimmten Wert konvergiert.[106,110,111]

Um den generellen Nachteil von Klassifizierungsalgorithmen anzugehen, dass keine Information darüber vorliegt, wie klar eine Klassifizierungsentscheidung getroffen wird, wurden für RF sogenannte *probability machines* entwickelt. Hierfür wird der Algorithmus im Regressionsmodus angewendet, um für die vorhergesagten Proben eine Wahrscheinlichkeit der Klassenzugehörigkeit zu den einzelnen Klassen zu generien.[112] Diese kann sowohl bezüglich der Klarheit der Entscheidung, aber auch zur Untersuchung der generellen Ähnlichkeit der untersuchten Klassen interpretiert werden.

3.3.4. Variable Importance

RF bietet verschiedene Möglichkeiten zur Analyse der Wichtigkeit einzelner Variablen. Die *Gini importance*, welche auch als *impurity importance* bezeichnet wird, summiert

den *Gini gain* aller Splits, die auf der entsprechenden Variablen basieren auf. Da die *Gini importance* allerdings eine Verzerrung aufweist und z.B. bezüglich des Vergleichs von kategoriellen Variablen fehlerhaft sein kann[113], wurde zusätzlich die sog. *permutation importance* entwickelt. Diese basiert auf der Permutation, also zufälligen Durchmischung der Werte einer Variablen über die Proben und wird aus der Differenz des OOB-Fehlers des RF Modells, welches aus den Originaldaten erhalten wurde, und den Daten, bei denen die entsprechende Variable permutiert wurde, erhalten.[113]

Zur Veranschaulichung ist eine Permutation beispielhaft anhand von drei Proben mit drei gemessenen Variablen in Tabelle 1 dargestellt.

Tabelle 1: Permutation von drei Proben mit den Originalvariablen x_{1-3} und den permutierten Variablen perm. x_{1-3}

	x_1	x_2	x_3	perm. x_1	perm. x_2	perm. x_3
Probe 1	1.34	5.43	3.42	2.09	3.87	2.61
Probe 2	2.14	4.99	2.61	1.34	4.99	0.17
Probe 3	2.09	3.87	0.17	2.14	5.43	3.42

Eine Weiterentwicklung der *Gini impurity* stellt die *actual impurity reduction* (AIR) dar, bei der die *Gini impurity* einer Variablen mit der *Gini impurity* einer permutierten Version dieser Variablen korrigiert wird:[113]

$$AIR_{X_i} = VIM_{original} - VIM_{reordered}$$

mit:

AIR_{Xi} – Neuberechnete Wichtigkeit einer Variable Xi

$VIM_{original}$ – Wichtigkeit einer betrachteten Originalvariable

$VIM_{reordered}$ – Wichtigkeit einer permutierten Originalvariable

3.3.5. Variablenselektion

Die Variablenselektion beschreibt den Prozess, die im vorigen Abschnitt ausgeführte *variable importance* dafür zu verwenden, wichtige von unwichtigen Variablen zu separieren. Sie kann zwei Ziele haben: Die Verbesserung, bzw. Vereinfachung des Modells und das Verständnis der Funktionsweise des Modells. Für ersteres Ziel wird die sog. *minimal-optimal* Selektion angewendet. Das Ziel hierbei ist es eine möglichst geringe Anzahl an Variablen zu finden, um die Rechenzeit zu reduzieren und die Performance eines ML-Algorithmus zu verbessern, welcher durch eine große Zahl irrelevanter Variablen beeinflusst wird.[114]

Im Rahmen dieser Arbeit haben wir uns auf das zweite Ziel fokussiert, welches als *all relevant selection* bezeichnet wird. Hier sollen alle relevanten Variablen selektiert werden, z.B. um die Mechanismen des Modells zu verstehen. Eine Studie von Degenhardt et al. hat in diesem Zusammenhang RF basierte Variablenselektionsverfahren unter Nutzung der *permutation importance* verglichen und die Methoden Vita[115] und Boruta[116] als die leistungsfähigsten Selektionsverfahren identifiziert.[117] Die Vita-Methode berechnet P-Werte basierend auf einer empirischen Null-Verteilung, welche nur nicht-positive *importance*-Werte verwendet. Daher wird für eine sinnvolle Anwendung dieser Methode ein Datensatz benötigt, der eine sehr große Menge an unwichtigen Variablen enthält. Da dies bei unseren Analysen nicht sicher der Fall ist, haben wir im Rahmen dieser Arbeit nur Boruta angewendet und mit dem neuartigen Verfahren *Surrogate Minimal Depth* (SMD)[118] verglichen. Die Funktionsweise dieser beiden Methoden soll im Folgenden erläutert werden.

3.3.6. Boruta

Boruta ist ein RF basierter Variablenselektionsalgorithmus, welcher auf der Permutation aller unabhängigen Variablen zur Generierung sogenannter *shadow variables* (SVs) basiert. Dabei wird sowohl für die Originalvariablen als auch die SVs eine *variable importance* berechnet und nur die Originalvariablen selektiert, die einen signifikant höheren Wert aufweisen. Um dies zu bestimmen, wird die Generierung der SVs und die Berechnung der *importance* Werte iterativ, in diversen sogenannten *runs*, durchgeführt und ein *Z-score* aus der durchschnittlichen *importance* geteilt durch dessen Standard-abweichungen für alle Variablen berechnet.[119] Zur Selektion relevanter Variablen wird dann ein statistischer Test angewendet, um die Variablen zu bestimmen, die signifikant höhere *importance* Werte haben als der jeweils höchste Werte der SVs eines Runs. In Abbildung 7 ist das Ergebnis der Anwendung von Boruta auf einen Beispieldatensatz dargestellt[116]. In dieser Abbildung sind der *Z-score* von den Variablen V_1-V_{13} sowie der des jeweils höchsten, mittlere und niedrigsten SVs als *Box-Whisker-Plot* dargestellt.

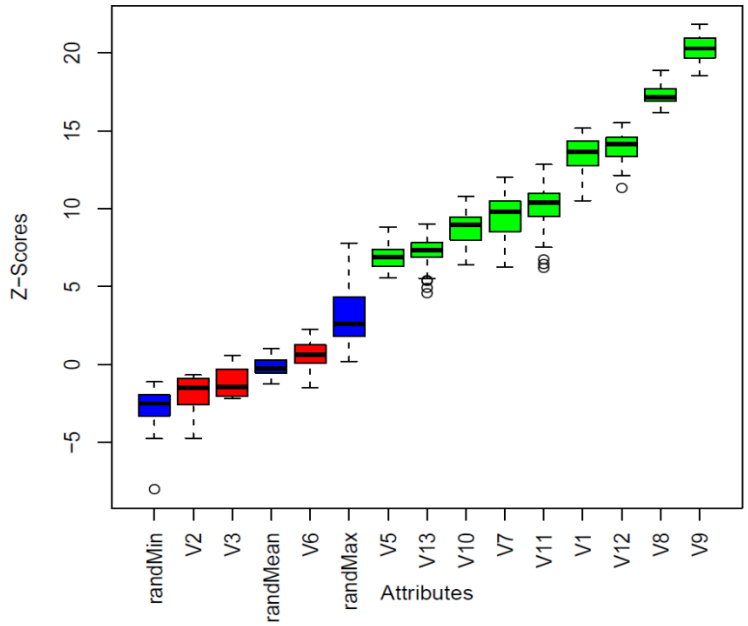


Abbildung 7: Ergebnis der Anwendung von Boruta auf einen Beispieldatensatz aus Kursa et al.[116] Es sind die Z-scores der Variablen und SVs (blau), sowie in Grün die selektierten und in Rot die nicht-selektierten Variablen gezeigt.

3.3.7. Surrogate Minimal Depth

Surrogate Minimal Depth (SMD) ist ein RF basiertes Maß für die Wichtigkeit von Variablen, welches auch zur Variablenselektion angewendet werden kann.[118] Der Unterschied zu anderen Verfahren besteht darin, dass hier die Wichtigkeit der Variablen nicht einzeln sondern in Zusammenhang mit anderen Variablen bewertet wird. Des Weiteren kann SMD genutzt werden, um die Beziehung zwischen Variablen bezüglich deren Einfluss auf das RF Modell zu untersuchen.

Grundlage für SMD ist das *variable importance* Maß *Minimal Depth (MD)*[120], welches die Wichtigkeit von Variablen nach deren jeweils erstem Auftreten in den Entscheidungsbäumen evaluiert und damit auf der Struktur der Entscheidungsbäume basiert. MD ist definiert als der Quotient aus der Summe der Ebene des ersten Auftretens über alle Entscheidungsbäume des RF und der Gesamtzahl an Entscheidungsbäumen[120]:

$$MD = \frac{\sum_{i=1}^{n} D}{n}$$

mit:

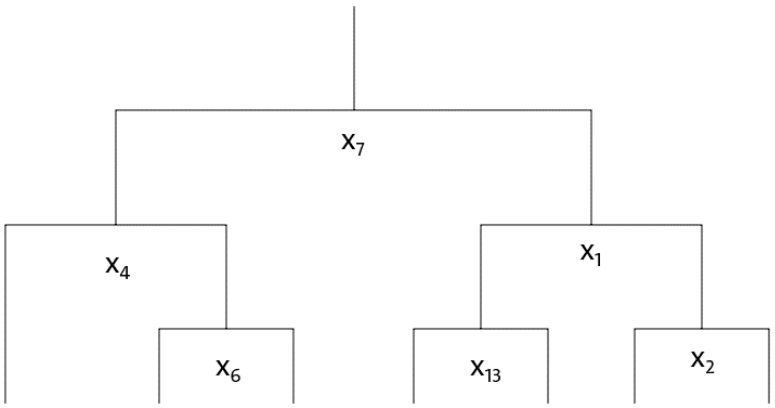
MD – *minimal depth* einer Variablen

i – Zahl des aktuellen Entscheidungsbaums

n – Anzahl aller Entscheidungsbäume

D – Tiefe des ersten Erscheinens einer Variablen in dem Entscheidungsbaum i

Das Konzept der *minimal depth* ist in der Abbildung 8 dargestellt.



Variable	*minimal depth*
x_1	1
x_2	2
x_4	1
x_6	2
x_7	0
x_{13}	2

Abbildung 8: Schematische Darstellung der *minimal depth* von Variablen anhand eines beispielhaften Entscheidungsbaums mit drei Ebenen.

Für die Variablenselektion wird die MD dann mit einem Grenzwert T_{MD} verglichen, welcher über die folgende Gleichung berechnet wird[118]:

$$T_{MD} = \sum j \cdot \pi_j \cdot n_j$$

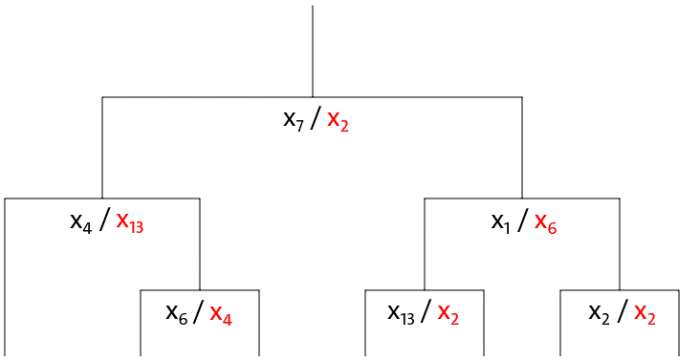
mit:

j – betrachtete Ebene

π_j –Wahrscheinlichkeit für die zufällige Wahl einer Variablen in Ebene j

n_j – Anzahl der *nodes* in j

Bei SMD ist nun nicht nur das erste Auftreten einer Variablen p für den Split in der *node*, sondern auch als Surrogatvariable q relevant. Diese wurden ursprünglich dafür etabliert, einzelne fehlende Variablen bei der Anwendung von RF Modellen durch alternative Splits mit einer ähnlichen Auftrennung der Proben auszugleichen.[107] Das Konzept von Surrogatsplits und SMD ist in Abbildung 9 dargestellt.



Variable	*minimal depth*	*surrogate minimal depth*
x_1	1	1
x_2	2	0
x_4	1	1
x_6	2	1
x_7	0	0
x_{13}	2	1

Abbildung 9: Schematische Darstellung der *minimal depth* und der *surrogate minimal depth, sowie* der Originalsplitvariablen (schwarz) und Surrogatvariablen (rot) für einen beispielhaften Entscheidungsbaum mit drei Ebenen.

Als Alternative zu Surrogatvariablen kann auch die sogenannte *majority rule* angewendet werden, welche bei fehlenden Variablen die Proben der *child node* zuordnen, die am meisten Proben enthält. Da Surrogatvariablen eine bessere Leistung als die *majority rule* aufweisen sollten, werden diese über den Parameter *adjusted agreement (*agree(p,q)) bestimmt, welcher für alle möglichen alternativen Variablen und Splits berechnet wird[118]:

$$\text{agree}(p,q) = \frac{n_{surr}\text{-}n_{maj}}{n_{total}\text{-}n_{maj}}$$

mit:

n_{total} – Anzahl an Proben an einer *node*

n_{maj} – Anzahl an Proben, die durch die *majority rule* der gleichen *child node* zugeordnet werden wie durch den Originalsplit

n_{Surr} – Anzahl an Proben, die durch den Surrogatsplit der gleichen *child node* zugeordnet werden wie durch den Originalsplit

Das *adjusted agreement* kann Werte zwischen 0 und 1 einnehmen. Ein Wert von 1 bedeutet eine exakte Übereinstimmung der Klassifizierungsleistung von Surrogat- und Originalsplit, während ein Wert von 0 eine entsprechende Übereinstimmung von Surrogatsplit und *majority rule* bedeutet. Die Surrogatsplits werden nun durch die

größten Werte der *adjusted agreement* für jeden nicht-terminale Knotenpunkt des RF bestimmt.[121] Dies ist in Abbildung 10 beispielhaft dargestellt.

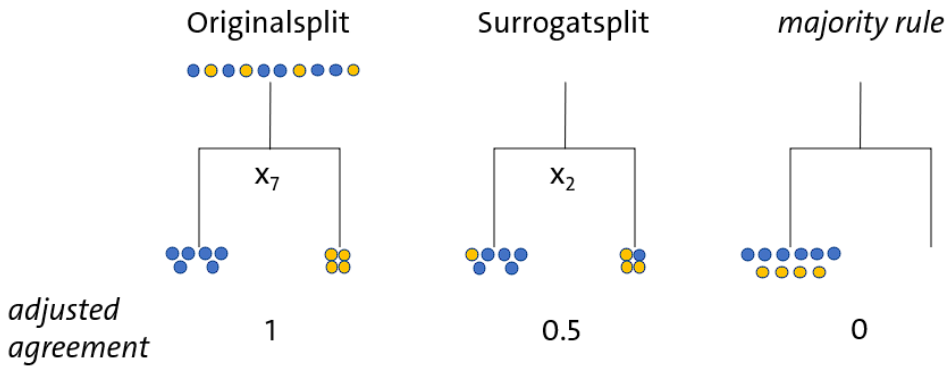


Abbildung 10: Visualisierung des *agreements* und des *adjusted agreements* innerhalb eines Splits anhand einer Original- und Surrogatvariable sowie der Vergleich mit einem Split entsprechend der der *majority rule*.

Da SMD nun nicht nur das erste Auftreten der Variable als Originalsplit, sondern auch als Surrogatsplit betrachtet, muss der Schwellenwert T_{SMD} für die Variablenselektion mit SMD nach Seifert et al. angepasst werden:[118]

$$T_{SMD} = \sum_{j} j \cdot \pi_j \cdot (n_j + \bar{s}_j)$$

mit:

$\bar{s}_j$ – Durchschnittliche Anzahl an Surrogatvariablen auf der Ebene j

Um die Beziehung zweier Variablen A und B bezüglich deren Einfluss auf das RF Modell zu analysieren, kann nun jeder Split betrachtet werden, der Variable A in seinem Originalsplit und Variable B in einem Surrogatsplit verwendet. Das *mean adjusted agreement* MAA ist nach Seifert et al. wie folgt definiert:[118]

$$MAA = \frac{\sum_{i=1}^{|nodes(A,B)|} agree(p_i^A, q_i^B)}{|nodes(A)|}$$

mit:

$agree(p_i^A, q_i^B)$ – Split, in welchem A die Originalsplitvariable darstellt und B die Surrogatvariable

nodes(A) – Anzahl aller *nodes*, in welchen A als Primärsplitvariable agiert

Der Beziehungsparameter MAA kann Werte zwischen 0 und 1 einnehmen. Ein Wert von 0 ist gleichbedeutend mit der Aussage, dass zwischen den Variablen kein Zusammenhang vorliegt und ein Wert von 1 bedeutet, dass die Variablen bezüglich deren Wirkung im RF exakt gleich sind. Im Gegensatz zu einem Korrelationskoeffizienten ist MAA nicht symmetrisch, was bedeutet, dass unterschiedliche Werte erhalten werden können, abhängig davon ob eine Variable für die Berechnung als Originalsplit- oder Surrogatvariable betrachtet wird.

3.4. Software

Für diese Arbeit wurde die *open-source*-Programmiersprache und -umgebung R für statistische Berechnungen, Datenauswertung und -visualisierung genutzt. In R sind diverse Pakete vorhanden, welche Funktionen durch Einbindung in die Programmierumgebung integrieren.[122] Details zu den verwendeten Paketen sind im kumulativen Teil der Dissertation (Abschnitt 5) aufgeführt.

4. Zielsetzung der Arbeit

Ziel dieser Arbeit ist die Untersuchung verschiedener *Metabolomics*-Daten von Lebensmitteln mit RF Methoden, welche dadurch bezüglich verschiedener Authentifizierungsfragen klassifiziert werden sollen. Des Weiteren werden relevante Variablen selektiert und identifiziert, um deren Einfluss auf die entwickelten Modelle detailliert zu analysieren. Dabei liegt ein Hauptschwerpunkt auf der Anwendung des neuartigen Variablenselektionsverfahrens SMD, welches die Einbeziehung und Analyse von Variablenbeziehungen ermöglicht.[118] Durch die erstmalige Anwendung dieses Verfahrens auf verschiedene *metabolomics* Datensätze, soll insbesondere analysiert werden, wie die gefundenen Variablenbeziehungen bezüglich des analytischen und biologischen Hintergrunds interpretiert werden können.

5. Kumulativer Teil der Dissertation

5.1. Authentifizierung von Äpfeln bezüglich verschiedener Fragestellungen durch ^{1}H Kernresonanz-Spektroskopie

In der ersten Publikation wurden die ^{1}H NMR-Daten von 217 authentischen Apfelproben aus den Erntejahren 2020 – 2022 untersucht und hinsichtlich deren geographischer Herkunft, Sorte und Anbaubedingung mit RF klassifiziert. Die Daten entstammen einer Erhebung des Landeslabors Schleswig-Holstein.

Für die Klassifikation nach Herkunftsländern wurde eine Vorhersagegenauigkeit für Äpfel aus Deutschland, Frankreich, Italien, Neuseeland, Chile und Südafrika von insgesamt 75.8 % erreicht, wobei Klassen mit einer großen Anzahl an Proben eine höhere Genauigkeit als Klassen mit wenigen Proben aufwiesen. In einem weiteren Ansatz wurde für die Unterscheidung von deutschen und nicht-deutschen Äpfeln eine Vorhersagegesamtgenauigkeit von 88.5 % erreicht.

Des Weiteren wurden die deutschen Apfelproben hinsichtlich ihrer regionalen Herkunft untersucht. Hierbei wurden die zwei Klassen “Nord” und “Süd” aus Proben aus Nord- (Schleswig-Holstein, Niedersachsen, Hamburg), und Süddeutschland (Baden-Württemberg, Bayern) gebildet und jeweils nur Daten von Apfelsorten verwendet, die in beiden Klassen vorlagen. Es wurde eine Vorhersagegenauigkeit von 80.7 % erzielt.

Für die Unterscheidung von biologisch und konventionell angebauten Äpfeln konnte eine Vorhersagegenauigkeit von 79.5 % erreicht werden. Allerdings wurde dabei die größere Klasse der konventionell angebauten Äpfel mit einer Sensitivität von 87.5 % mit einer viel höheren Wahrscheinlichkeit richtig klassifiziert als die kleinere Klasse der biologisch angebauten Äpfel, welche nur eine Sensitivität von 52.9 % aufwies. Ein ähnlicher Effekt zeigte sich auch bei der Klassifizierung der Apfelsorte, bei der eine Vorhersagegenauigkeit von 73.2 % erreicht wurde. Auch hier wurde bei größeren Klassen

höhere Genauigkeiten und häufig eine Sensitivität über 90 % erreicht, während kleinere Klassen teilweise eine Sensitivität unter 50 % zeigten.

Durch die Anwendung der Variablenselektionsverfahren SMD und Boruta konnte gezeigt werden, dass die ^{1}H NMR Spektroskopie ein vielversprechender Ansatz zur simultanen Untersuchung verschiedener Klassifizierungsfragen von Äpfeln darstellt, da verschiedene Teile des Spektrums für einzelne Fragestellungen relevant sind. Die vergleichsweise niedrige Klassifizierungsgenauigkeit kleinerer Klassen wurde darauf zurückgeführt, dass in diesen Fällen die Varianz der untersuchten Klasse nicht ausreichend im Modell repräsentiert werden konnte.

Food Control 167 (2025) 110817

Contents lists available at ScienceDirect

Food Control

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

Authentication of apples (*Malus* × *domestica* BORKH.) according to geographical origin, variety and production method using ^{1}H NMR spectroscopy and random forest

Soeren Wenck [a], René Bachmann [b,**], Sarah-Marie Barmbold [b], Anna Lena Horns [b], Nele Paasch [b], Stephan Seifert [a,*]

[a] Institute of Food Chemistry, Hamburg School of Food Science, University of Hamburg, Grindelallee 117, 20146, Hamburg, Germany
[b] Landelabor Schleswig-Holstein, Max-Eyth-Str. 5, 24537, Neumünster, Germany

ARTICLE INFO

Keywords:
Random forest
^{1}H NMR spectroscopy
Apples
Food authentication
Classification
Variable selection
Machine learning
Metabolomics

ABSTRACT

When buying apples, there are several characteristics that are important to the consumer. These include taxonomic variety, which influences the desired flavor and the allergenicity, and, increasingly, whether the cultivation is regional and organic. In this study, we analyzed 217 apple samples with ^{1}H NMR spectroscopy and verified the authentication of the apples in terms of country of origin, local origin within Germany, taxonomic variety and whether they were organically produced. The obtained random forest models showed classification accuracies ranging from about 73 % for the discrimination of seven taxonomic varieties to 88.5 % for the separation of German and non-German samples. The results demonstrate that ^{1}H NMR spectroscopy is promising for comprehensive authentication of apples. This is achieved by using different parts of the NMR fingerprint depending on the desired outcome, which was confirmed by the application of variable selection methods and the comparison of the selected variables.

1. Introduction

Apples, *Malus* × *domestica* BORKH., are the third most produced fruit in the world after bananas and watermelons (FAO Statista, 2023) and the most consumed fruit in Germany (Statista, 2023) with a per capita consumption of 21.9 kg in 2019/20 (BLE, 2020). They are associated with a healthy lifestyle, due to their richness in vitamins and antioxidants like phenolic acid, which have been shown to reduce the risk of cancer and DNA damage (Kaur & Kapoor, 2001). In addition to price, other criteria are also playing an increasingly important role for consumers, such as whether the apples are organically grown or the country of origin. Especially regarding European countries, regionally produced food is becoming increasingly popular with consumers (Denver & Jensen, 2014). Furthermore, the taxonomic variety is relevant, as this is decisive for the flavor of the apple and also influences allergenicity (Chitarrini et al., 2020; Siekierzynska et al., 2021), which is relevant to 7.5 million people in Germany alone (Becker et al., 2021). Apple allergy, a cross-allergy with birch, peach or cherry (Siekierzynska et al., 2021), is caused by Mal d1 and its isoallergens, which are present in varying amounts in apples of different taxonomic varieties (Kaeswurm et al., 2023).

To authenticate apples with regard to their geographical origin, production method and taxonomic variety, various analytical approaches have been developed. For example, a differentiation between apple varieties was conducted by genome sequencing (Duan et al., 2017) and NMR spectroscopy on pulp and peel extracts (Eisenmann et al., 2016). Furthermore, classification of apples regarding the geographical origin was performed using elemental profiling (J. Zhang et al., 2019, 2023) and differentiation between organically and conventionally grown apples with diffraction grating and image processing (Jiang et al., 2018).

The analytical approaches presented so far, each focus on one authentication question, for example variety through genetic analyses. In order to target several of these questions using a single method, approaches analyzing the metabolome are promising. The metabolome is influenced by many factors, e.g. the genome of the analyzed sample, as

* Corresponding author. Grindelallee 117, 20146, Hamburg, Germany.
** Corresponding author. Max-Eyth-Str. 5, 24537, Neumünster, Germany.
E-mail addresses: rene.bachmann@lsh.landsh.de (R. Bachmann), stephan.seifert@uni-hamburg.de (S. Seifert).

https://doi.org/10.1016/j.foodcont.2024.110817
Received 3 May 2024; Received in revised form 30 July 2024; Accepted 13 August 2024
Available online 23 August 2024

well as external influences such as sun exposure, proximity to large bodies of water, humidity, fertilization, availability of water and soil composition (Fiehn, 2002). Thus, the metabolome can be understood as the closest representation of an organism's phenotype (Fiehn, 2002). The total metabolome is made up of thousands of metabolites, and so far, no technique has been able to investigate all metabolites at once. Different analysis approaches are therefore used to investigate different fractions of the metabolome of foods. These are primarily approaches based on mass spectrometry (MS) (Creydt et al., 2018), near-infrared (NIR) spectroscopy (Shakiba et al., 2021) or nuclear magnetic resonance (NMR) spectroscopy (Bachmann et al., 2018; Shakiba et al., 2021).

NMR has the advantages that it is non-destructive, highly reproducible and allows non-relative quantification of metabolites (Bingol & Brüschweiler, 2014; Markley et al., 2017). However, the identification of metabolites is difficult since signals overlap and chemical shifts are altered by experimental conditions such as pH or ionic strength (Pandey et al., 2023). Because of these chemical shifts, NMR spectra are usually aligned and binned, i.e. neighboring variables are combined. (Emwas et al., 2019; Markley et al., 2017; Ravanbakhsh et al., 2015; Wishart, 2007).

As NMR spectra are high-dimensional data with many variables from comparatively few samples, multivariate approaches are applied for their analysis. The popular unsupervised approach principal component analysis (PCA) creates latent variables via linear combination of the original variables. These principal components highlight the main variances within the dataset and allow the identification of groups with similar patterns (Bro & Smilde, 2014; Wold et al., 1987; Worley & Powers, 2012). Supervised approaches incorporate the class information of the samples into the analysis, resulting in a model that focuses on specific differences. For the authentication of food based on the supervised analysis of NMR data, for example artificial neural networks (Mendez et al., 2019), support vector machines (Boser et al., 1992; Mix et al., 2023) and random forest (RF) (Wenck et al., 2023) have been successfully applied.

RF is a non-parametric supervised ensemble learning algorithm, that performs classification based on the generation of a large number of binary decision trees (Breiman, 2001). This approach offers many advantages for the application to high-dimensional data, as it can handle small sample sizes and a large number of uninformative variables particularly well (Biau & Scornet, 2016). In addition, an internal validation is implemented, which is based on the fact, that each decision tree is trained on a different fraction of the samples, called bootstrap samples. The remaining samples of each tree can therefore be used to generate independent out-of-bag (oob) errors. Since it is not absolutely necessary to tune the RF parameters (Probst et al., 2019), samples can be utilized for both training and testing, which is particularly advantageous with very small data sets.

RF can also be used to generate variable importance scores. These scores are, for instance, based on the decrease of accuracy obtained by permutation of each variable or on the decrease of impurity, a variable is contributing to the RF (Breiman, 2001). Variable selection approaches use the importance scores to distinguish between relevant and irrelevant variables by defining a threshold. The Boruta approach randomly permutes variables to create unimportant *shadow variables*, whose importance is compared with that of the actual variables and those generally showing higher scores are selected (Kursa & Rudnicki, 2010). In contrast to most methods that consider variables individually, surrogate minimal depth (SMD), which has recently been further developed to also analyze qualitative variables (Voges et al., 2023), incorporates variable relations into the selection process (Seifert et al., 2019). SMD has already been applied to various analytical data (Lösel et al., 2023; Seifert, 2020; Shakiba et al., 2021; Wenck et al., 2021; Živanović et al., 2019), including NMR data (Wenck et al., 2023).

In this study, we apply ^{1}H NMR spectroscopy in combination with RF to classify apples with respect to various authentication issues. These range from the origin of the apples in terms of country of origin, the question of whether they are from Germany or not and the distinction between different regions within Germany, to taxonomic variety and the distinction between organically and conventionally produced apples. In addition, using variable selection methods, we show that these classifications are mainly based on different parts of the NMR fingerprint. This demonstrates that ^{1}H NMR spectroscopy is a promising approach for investigating multiple authentication issues of apples with a single approach.

2. Materials and methods

2.1. Collection of samples and preparation

109, 66 and 42 samples were collected in 2020, 2021 and 2022, respectively, adding up to 217 samples in total. For each of these samples, the whole apples were homogenized using a standard household juicer (Sage Appliances, Krefeld Germany) and 1000 µL of the liquid received was transferred to a 2.0 mL reaction tube (Eppendorf, Germany) and centrifuged (18.138 rcf, 10 min). 900 µL of the supernatant was taken and 100 µL of potassium phosphate buffer (1M) was added. The samples were vortexed for 30 s and 600 µL of each sample was transferred to an NMR tube (Deutero, Castellaun Germany).

2.2. Reagents & chemicals

Deuterium oxide (99.9%) was purchased from Eurisotop (Saint-Aubin Cedex, France), sodium azide (99.5%), mono- and dipotassium phosphate dibasic anhydrous (>98%) from VWR (VWR International GmbH, Darmstadt, Germany). For preparation of the extraction buffer 4.2 g monopotassium phosphate (anhydrous) and 10.9 g dipotassium phosphate (anhydrous) were dissolved in 100 mL D_2O. The D_2O included TMSP in a concentration of about 100 mg/L 2 mg sodium azide was added to the extraction buffer and the pH value was adjusted to 7.0.

2.3. NMR data acquisition

All spectra were acquired on a Bruker Ascend 400 MHz spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at 400.13 MHz equipped with an Avance III console. The noesygppr1d (Bruker annotation) pulse sequence was used for acquisition of water suppressed ^{1}H NMR spectra applying the digitization mode baseopt (Bruker annotation). The spectra were recorded at 301.8 K, with 16 scans, 65536 complex data points, D1 of 4 s, and a spectral width of 8417.5 Hz. The receiver gain was set to 16 and the transmitter frequency offset was set to 1880.6 Hz. The pulse sequence hsccedetgpsisp2.3 was used for the acquisition of HSQC spectra. The spectra were recorded at 300 K, 32 Scans, 2048 (F2) and 256 (F1) data points, D1 of 1.5 s and a spectral width of 5197.5 (F2) and 16611.3 (F1) Hz. The RG was set to 207 and the transmitter frequency offset was set to 1880.6 (F2) and 7546.0 (F1) Hz. For acquisition of HMBC spectra the puls sequence hmbcetgpl3nd with NS = 64, 1024 (F2) and 256 (F1) data points, D1 of 1.5 s and a spectral width of 2403.8 (F2) and 22123.9 (F1) Hz and a RG of 207 was used. TOCSY spectra were acquired using the pulssequence mlevphpr.2 with NS = 32, 2048 (F2) and 256 (F1) data points, D1 of 2 s and a spectral width of 4000 Hz (F2+F1) and a RG of 207.

2.4. Data processing and analysis

For data processing the FIDs were Fourier transformed with a line broadening factor of 0.3 Hz, baseline corrected and phased with Topspin 3.6.2 (Bruker Biospin, Rheinstetten, Germany). All spectra were aligned to the signal of glucose (5.22 ppm) because the carbohydrate signals are least sensitive to changes and exhibit a constant chemical shift. For multivariate analysis the integrals and buckets were defined using Amix 3.9.15 (Bruker Biospin, Rheinstetten, Germany). The spectra were

divided into 490 buckets with a variable width of 0.02 ppm from 9.990 to 0.210 ppm. The selected bucket width is based on experience from previous studies and is chosen so that, on the one hand, not too many data points are obtained in proportion to the samples and, on the other hand, the sharpest possible differentiation of individual signals is achieved. The buckets were normalized to total intensity and the region of the water signal from 4.9 to 4.5 was excluded. The final bucket table was exported as a *.csv file and can be found in Table S2. Principal Component Analysis was performed with centered and non-scaled data. Random Forest classification and the subsequently conducted variable selection were carried out with the parameters listed in Table 1.

2.5. Software

The software R (version 3.6.3) and the R packages ranger (version 0.14.1, CRAN) for RF classification, mdatools (version 0.14.1, CRAN) for PCA, Pomona (version 1.0.1, https://github.com/silkeszy/Pomona, accessed on December 06, 2023) for Boruta variable selection and SurrogateMinimalDepth (version 0.2.0, https://github.com/Stephan Seifert/SurrogateMinimalDepth, accessed on January 15, 2024) for SMD variable selection were used. Figures were created with ggplot2 (version 3.4.1, CRAN) and UpSetR (version 1.4.0, CRAN).

2.6. Analyzed research questions and utilized data

The data of the 217 samples were utilized differently for the different research questions. For the separation of German and non-German samples, all available samples meaning 130 German and 87 non-German samples were used. The latter were comprised of 31, 19, 13, 9, 9, 3, 2 and 1 samples from Italy, New Zealand, South Africa, Chile, France, Argentina, Austria and Poland. Since data from a sufficiently high number of samples must be available, we only used groups of at least 8 for the classification of the individual countries and, thus, the samples from Argentina, Austria and Poland were excluded for this analysis. The same minimum group size was also used to analyze the varieties, resulting in a data set with 26, 24, 22,12,11,9 and 8 samples from *Gala*, *Cripps Pink*, *Elstar*, *Boskoop*, *Braeburn*, *Holsteiner Cox* and *Jonagold*, respectively, giving a total of 112 samples and 7 varieties. For the analysis regarding the local origin within Germany, two growing regions "North" (32, 11 and 3 samples from Schleswig-Holstein, Lower Saxony and Hamburg, respectively) and "South" (36 samples from Baden-Wuerttemberg and 1 from Bavaria) were defined. Only samples of varieties found in both regions, namely *Boskoop*, *Braeburn*, *Delcorf*, *Elstar*, *Gala*, *Glockenapfel*, *Gloster*, *Jonagold*, *Santana*, *Topaz* and *Wellant*, were used. Also, for the differentiation between organically and conventionally grown apples, only the 73 samples of varieties that occur in both groups, namely *Boskoop*, *Elstar*, *Gala*, *Holsteiner Cox*, *Jonagold* and *Topaz*, were used. More detailed information about the samples and their assignment to the corresponding classes can be found in Table S1.

Table 1
Parameters used for RF-based approaches with p representing the total number of variables.

Approach	Parameter	Description	Value
Random Forest	ntree	number of trees	50,000
	min.node.size	number of samples in terminal node	1
	mtry	number of candidate variables	105 ($\sqrt[3]{p^4}$)[a]
	case.weights	weights for sampling of training observations	chosen according to the size of the respective class
Boruta	importance	applied importance measure	impurity_corrected
	pValue	confidence level	0.01
	maxRuns	maximum number of importance source runs	100
Surrogate Minimal Depth	s	predefined number of surrogate variables	25 (p • 0.05)

[a] Motivated by Ishwaran et al. (Ishwaran et al., 2011).

2.7. Assignment of NMR signals to metabolites

NMR signals were assigned to the molecules 1-butanol, isoleucine, valine, 1,3-butandiole, ethanol, lactic acid, citramalic acid, alanine, quinic acid, malic acid, asparagine, glucose, fructose, sucrose, xylose, chlorogenic acid, epicatechin, fumaric acid, formic acid and acetaldehyde using databases. The results were subsequently checked for plausibility using HSQC, HMBC, and TOCSY spectra according to recommended standard procedures (Schönberger et al., 2023). The respective 2D spectra are shown in the Supplementary Figures S1-S4 and Table S3 lists the buckets together with the assigned metabolites.

3. Results and discussion

3.1. H NMR data

In Fig. 1, a representative NMR spectrum of an apple sample is depicted. Many of the signals could be assigned to known apple metabolites, mostly amino acids, sugars, as well as organic acids and their anions. The signals with the largest intensities are located in the spectral regions between 3.3 and 4.3 ppm, in which a large number of sugar signals are superimposed. However, some signals could be assigned to fructose (4.09–4.12 ppm), glucose (4.64–4.66 ppm and 5.24–5.26 ppm), xylose (5.20–5.22 ppm), as well as sucrose (5.40–5.42 ppm). Sugars are known to be relevant apple components and main contributors to their sweet taste (Aprea et al., 2017).

The organic acid with the most intense signals in the spectrum is malic acid showing two doublets at 2.35–2.50 ppm and 2.65–2.75 ppm. Malic acid is known to be present in apples in higher concentrations than other organic acids such as lactic, succinic and fumaric acid, and is mainly responsible for the sour flavor. This has been demonstrated in *Fuji* apples, for example (Yan et al., 2018) and is confirmed by our analysis as the lactic acid signal at 1.31–1.34 ppm is much less intense than the signals of malic acid. The metabolites quinic acid, epicatechin and chlorogenic acid could be assigned to signals in the range of 1.8–2.3 ppm, at 6.15 ppm and around 7.1 ppm, and at 6.4 ppm, 7.2 ppm and 7.6 ppm, respectively. Quinic acid has antibacterial properties after digestion, making it an important compound for human nutrition (Bai et al., 2022).

Epicatechin has been shown to have beneficial effects on brachial artery flow-mediated vasodilation in individuals with elevated blood pressure, emphasizing its importance in nutrition (Saarenhovi et al., 2017). In addition, it has been shown in *Fuji* apples, that high levels of epicatechin in apples increase their resistance to *Botrytis cinerea*, a common mold responsible for apple spoilage (M. Zhang et al., 2020), which may indicate that this metabolite counteracts exogeneous stress in apples. Chlorogenic acid is one of the compounds responsible for the tangy flavor and enzymatic browning of apples (Siebert et al., 2019) and is one of the most abundant polyphenolic compounds in human diet (Liao et al., 2021). It has proven anti-bacterial and anti-inflammatory properties, acts hepato- and cardio-protective, and has been identified as a beneficial compound in anti-diabetic and anti-obesity oriented nutrition (Cho et al., 2010; Maalik et al., 2016; Santana-Gálvez et al., 2017).

It is important to emphasize, that signals in the ^{1}H NMR spectra are often superimposed, meaning that although we assigned signals to metabolites in a particular spectral region, these are usually not the only metabolites that show signals there. The spectrum can therefore be considered as a fingerprint, which is analyzed in the following sections using multivariate methods.

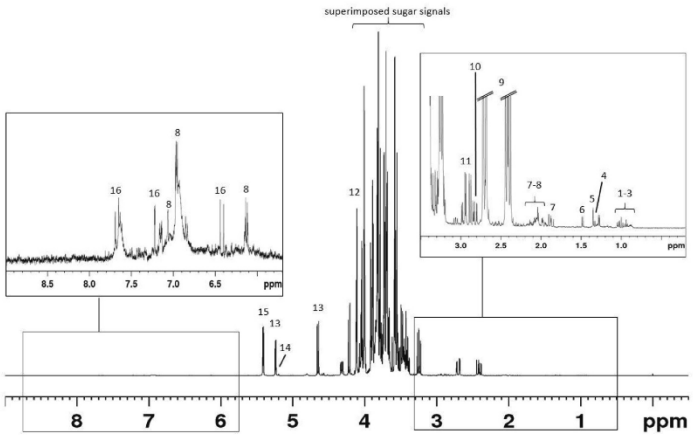


Fig. 1. ^{1}H NMR spectrum of a conventionally grown German *Jonagold* sample from Lower Saxony representing spectra of the apple samples in this study. Many of the NMR signals could be assigned to metabolites by the procedure explained in section 2: 1 – Leucine, 2 – Isoleucine, 3 – Valine, 4 – Lactic acid, 5 – Citramalic acid, 6 – Alanine, 7 – Quinic Acid, 8 – Epicatechin, 9 – Malic acid, 10 – Aspartic acid, 11 – Asparagine, 12 – Fructose, 13 – Glucose, 14 – Xylose, 15 – Sucrose, 16 – Chlorogenic acid.

3.2. Principal component analysis

In order to analyze the largest variance among the samples, PCA was applied. The scores of the first and second principal component (PC) colored according to the geographical origin are depicted in Fig. 2. No grouping of samples from individual countries can be observed. A clear grouping is also not obtained if the PCA is performed and colored according to the other authentication issues examined (see Figs. S5–8). The loadings (Fig. S9) show, that the main variance within the data can be attributed to the signals of malic acid at 2.60–2.80 ppm the spectral region between 3.10 and 4.50 ppm characterized by sugar signals, as well as the glucose signal between 5.24 and 5.26 ppm and the sucrose signal at 5.40–5.42 ppm. The variance of the most distinctive signals in the spectra, mainly attributed to the sugar composition of the apples (see previous section and Fig. 1), are therefore obviously not decisive for the authentication issues investigated. Additionally, hierarchical clustering was performed and also in the dendrograms no clear clustering or explainable substructures according to the authentication issues could be observed (Figs. S10–S14).

In order to specifically investigate these questions, the supervised approach random forest was applied in the following sections.

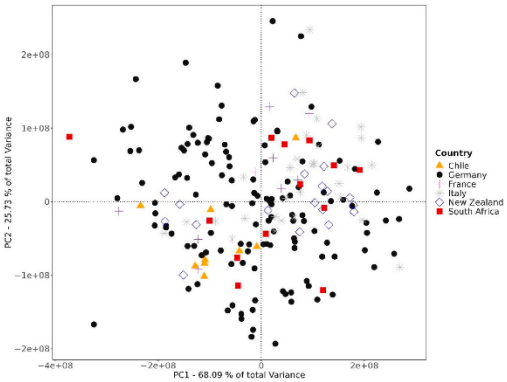


Fig. 2. Results of the PCA: Scores of the first and second PC are shown and labelled according to the country of origin of the respective samples. The PCA for the other authentication issues examined, as well as the loadings, can be found in Supplementary Figs. S5–9.

3.3. Random forest analyses

In order to train models on specific differences in the NMR spectra, random forest was applied. In the following, models for the classification of the apple samples with regard to their country of origin (3.3.1), specifically for the question of whether they originate from Germany or not (3.3.2), with regards to their regional origin within Germany (3.3.3) and with regards to their method of production (3.3.4) and taxonomic variety (3.3.5) are presented.

3.3.1. Differentiation by country of origin

The results for the differentiation of the apples by country of origin are shown in Table 2. The highest sensitivity was obtained for samples from Germany with 90 %. The lowest values for sensitivity were obtained for the samples from France, Chile and South Africa with 0 %, 33.3 % and 30.8 %, respectively. In general, there are comparatively many misclassifications. On the one hand, this could be due to the geographical proximity and the associated similar environmental conditions of the samples within Europe, as French samples are often misclassified as German or Italian. Similarities between the samples in terms of other properties could also be an obstacle to successful classification. The French samples, for example, are all of different taxonomic varieties and variety-based similarities with samples from other countries could have a negative impact on the classification of the country of origin. The most likely reason for the misclassifications is that the high variance of the corresponding classes, e.g. caused by different taxonomic varieties, is not sufficiently represented by the comparatively few samples. An example of the comparatively high variance is the group of South African samples. South Africa's apple growing regions are located in the Eastern and Western Cape provinces and these regions have very different climates, altitude and growing conditions (Vink, Tregurtha, 2005), which should also result in the metabolome of the samples from these two groups being very different (Fiehn, 2002). The assumption

Table 2
Results of the apple authentication regarding the country of origin reaching a total accuracy of 75.8 %, corresponding to an oob error of 24.2 %. The following abbreviations are used: CL – Chile, DE – Germany, FR – France, IT – Italy, NZ – New Zealand, ZA – South Africa.

	Predicted							
True		CL	DE	FR	IT	NZ	ZA	Sensitivity [%]
	CL (9)	**3**	1	2	2	0	1	33.3
	DE (130)	0	**117**	1	5	6	1	90.0
	FR (9)	2	4	**0**	3	0	0	0
	IT (31)	0	6	1	**22**	2	0	71.0
	NZ (19)	0	1	0	4	**14**	0	73.7
	ZA (13)	2	3	0	1	3	**4**	30.8
	Specificity [%]	97.5	74.1	97.6	90.2	93.0	98.7	

that an insufficient number of samples to express the variance of the group is the main reason for the misclassifications is confirmed by the fact that larger groups such as New Zealand, Italy and, especially, Germany show better accuracies than the smaller groups. It has been shown that the determination of the geographical origin of honey based on mass spectrometric data could be improved by increasing the sample size and thus better representing the variance of the corresponding classes in the random forest model (Hansen et al., 2024). When analyzing the origin of apples using NMR spectroscopy as shown here, similar effects of increasing sample size on the classification model can be expected.

3.3.2. *Differentiation between German and non-German samples*

Regionality, i.e. whether a sample originates from the country in which it is sold, is often the most relevant factor in determining origin. Therefore, a two-class classification between German and non-German samples was carried out and the results are shown in Table 3. A total accuracy of 88.48 % was obtained demonstrating that it is possible to differentiate between these two groups with a comparatively high accuracy. The misclassified non-German samples primarily originate from other European countries, namely France, Italy, Austria and Poland (see Table S1). This suggests that similar environmental conditions due to geographic proximity are the main reason for these comparatively few misclassifications.

The successful differentiation between German and non-German samples shows that classification with high accuracy is possible if data from many samples are available for training in order to adequately represent the variance of the groups in the model. This was not always the case for the groups of individual countries in the previous section.

3.3.3. *Differentiation of regional origin within Germany*

The regionality of the apple samples does not only concern the country of origin. For consumers, it is often just as relevant that it comes from a specific region within a country e.g. to ensure short transport distances. To analyze the regional distinctiveness of apple samples using NMR spectroscopy, a differentiation of northern and southern German samples was conducted only using samples from northern and southern states and omitting the samples from the center of Germany. The results are shown in Table 4. A total accuracy of around 81 % was reached demonstrating that also regional differences of apple composition are reflected in the NMR fingerprint. The misclassification could be due to various causes. It is known that, for example salinity and soil composition influence apples on a molecular level (X. Li et al., 2023), and specifically clay in the soil of apples can cause significant changes (Schimmel et al., 2024). Since clay can be contained in high amounts in both northern and southern Germany soils, which was demonstrated in a study of the German Federal Department of Environment from 2015 (UBA, 2015), the misclassifications could be due to similar soil conditions of the respective samples.

Table 3
Results of the apple authentication for the differentiation of German and non-German samples. A total accuracy of 88.5 % was obtained, corresponding to an oob error of 11.5 %.

	Predicted			
True		German	Non-German	Sensitivity [%]
	German (130)	**122**	8	93.8
	Non-German (87)	17	**70**	80.5
	Specificity [%]	80.5	93.8	

Table 4
Results of the apple authentication regarding the differentiation of regional origin within Germany. A total accuracy of 80.7 % was obtained corresponding to an oob error of 19.3 %.

	Predicted			
True		North	South	Sensitivity [%]
	North (46)	**40**	6	86.9
	South (37)	9	**28**	75.7
	Specificity [%]	75.7	86.9	

3.3.4. *Differentiation of biologically and conventionally produced apples*

The fact that apples are grown organically, without the use of herbicides, insecticides, pesticides, fungicides or synthetic fertilizers, is another feature that is becoming increasingly important to consumers. The classification of the NMR spectra with regard to this property was also analyzed and the results are shown in Table 5. The differentiation between conventionally grown and organically grown apples is possible with an accuracy of 79.5 % according to these results. Similar to the analysis of the samples regarding country of origin in section 3.3.1., the smaller group of organic apples shows a lower correct classification with around 53% compared to around 88 % of the larger group of conventionally produced apples. Since, as before, the statistical bias that arises when using different sample sizes was compensated by oversampling when training the models, the reason for this different performance is probably also due to a poorer representation of the group variance in the data set.

3.3.5. *Differentiation by taxonomic variety*

The results for the classification of seven varieties, which achieve an accuracy of 73 %, are shown in Table 6. Again, there are differences in performance between the larger groups *Gala* and *Cripps Pink*, which achieve very high sensitivities of around 96 % and 92 %, respectively, and the smaller groups *Holsteiner Cox* and *Jonagold*, which show sensitivities below 50%. The medium sized groups *Braeburn* and *Elstar* have sensitivities in between, namely 54.5 % and 63.6 %. As with the

Table 5
Results of the apple authentication for the differentiation of biologically and conventionally produced apples. A total accuracy of 79.5 % was obtained, corresponding to an oob error of 20.5 %.

	Predicted			
True		Organic	Conventional	Sensitivity [%]
	Organic (17)	**9**	8	52.9
	Conventional (56)	7	**49**	87.5
	Specificity [%]	87.5	52.9	

Table 6
Results of the apple authentication regarding the taxonomic variety reaching a total accuracy of 73.2 %, corresponding to an oob error of 26.8 %. The following abbreviations are used: Bo – *Boskoop*, Br – *Braeburn*, CP – *Cripps Pink*, Els – *Elstar*, Ga – *Gala*, HC – *Holsteiner Cox*, Jo – *Jonagold*.

	Predicted								
True		Bo	Br	CP	Els	Ga	HC	Jo	Sensitivity [%]
	Bo (12)	**9**	0	0	**1**	0	2	0	75.0
	Br (11)	0	**6**	0	2	2	0	1	54.5
	CP (24)	1	1	**22**	0	0	0	0	91.7
	Els (22)	0	1	2	**14**	0	4	1	63.6
	Ga (26)	0	1	0	0	**25**	0	0	96.2
	HC (9)	1	1	0	2	1	**4**	0	44.4
	Jo (8)	1	1	0	**1**	3	0	**2**	25.0
	Specificity [%]	96.1	93.8	96.8	91.9	90.5	92.9	97.6	

classifications discussed previously, the size of the group, and thus the ability to represent the variance of the groups in the model, appears to be critical to the success of the differentiation by taxonomic variety. However, there are several additional possible causes for misclassification among the varieties, for example similar soil composition as mentioned in 3.3.3. (Schimmel et al., 2024). Another explanation could be a molecular similarity due to a close genetic relationship between the varieties, since many of the modern apple varieties share common ancestry (Bannier, 2011). The relatively high number of misclassifications between *Holsteiner Cox* and *Elstar*, for example, could be explained by the fact that both have the variety *Cox Orange* as an ancestor. (Baric et al., 2020). In addition, the three *Jonagold* samples incorrectly predicted as *Gala* could be due to the fact that both varieties have *Golden Delicious* as a parent. (Baric et al., 2020).

3.4. Analyzing the overlap of the spectral information used

Fig. 3 and Fig. S15 show the overlap of the selected variables for the analyzed authentication issues using Boruta and SMD as selection approaches, respectively.

Boruta mainly selects variables that are exclusive for the respective authentication question. These are labelled in the NMR spectrum in Fig. 4.

For the differentiation by taxonomic variety, variables located between 6.74 and 7.06 ppm, as well as between 2.3 and 2.5 ppm and 2.6 and 2.8 ppm are exclusively selected (see Fig. 4 and Table S4). The first region can be assigned to epicatechin signals and the second two regions to malic acid signals. These results are consistent with previous findings, since taxonomic variety and catechin or malic acid concentration have previously been shown to be related (Y. Li, Sun, Li, Qin, Niu, et al., 2021; Y. Li, Sun, Li, Qin, Yang, et al., 2021). These findings are also confirmed by the fact that apples of different taxonomic varieties taste differently sour. For the differentiation by country of origin, mainly variables in the lower ppm range, e.g. between 1.38 and 1.42 ppm assigned to citramalic acid are selected exclusively (see Fig. 4 and Table S4). For both, the differentiation of the growing method and the region within Germany mainly variables assigned to sugars are exclusively selected, e.g. between 5.34 and 5.36 ppm and 3.10 and 3.16 ppm, respectively. However, there is an overlap between the selected variables for the differentiation by country of origin and taxonomic variety, which is even more pronounced due to the inclusion of variable relationships when using SMD (see Fig. S10). The jointly selected variables include the regions between 1.88 and 2.10 ppm, 5.22 and 5.24 ppm, as well as between 2.96 and 2.98 ppm (see Table S4), and can be assigned to quinic

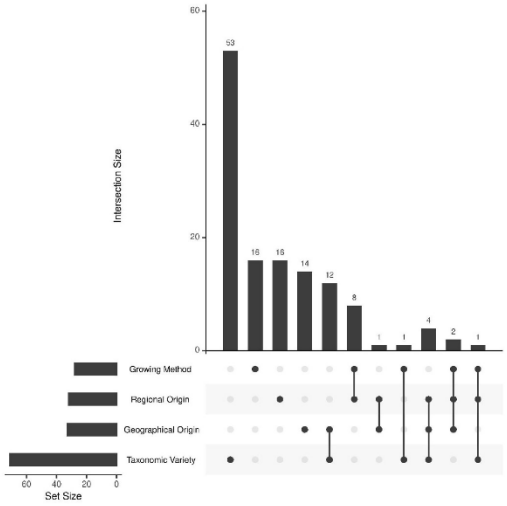


Fig. 3. UpSetPlot showing the overlap of selected variables for the differentiation of apples regarding their geographical origin, regional origin, growing method and taxonomic variety. Each column with a single dot shows the number of variables exclusively selected for a single authentication question, while columns with multiple dots indicate the number of variables selected for those multiple questions.

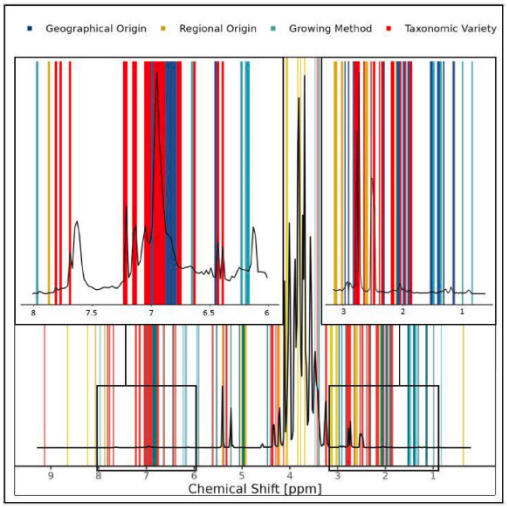


Fig. 4. Bucketed ^{1}H NMR spectrum of a conventionally grown German *Jonagold* sample from Lower Saxony colored with selected buckets that are exclusive for the respective authentication questions.

acid, glucose and asparagine, respectively. This overlap can presumably be explained by the fact that the samples of different taxonomic varieties are not homogeneously distributed across the different countries and thus the existing taxonomic variety also represents a country-specific characteristic in the data set. Examples for this are the variety *Holsteiner Cox*, a regional variety, which is almost exclusively grown in northern Germany, the varieties *Jonagold* and *Boskoop*, which only originate from Germany and *Cripps Pink*, which do not come from Germany at all.

The variables jointly selected for the differentiation by regional origin within Germany and growing method are mostly located in the region between 3.1 and 5.0 ppm and can therefore be assigned primarily to sugar signals. In summary, different parts of the complex NMR spectrum are used for different authentication issues, making this approach promising for a comprehensive simultaneous application for apple authentication.

4. Conclusion and outlook

In this study, 217 apple samples were analyzed with ^{1}H NMR spectroscopy to investigate different authentication issues. To this end, the NMR data were analyzed using random forest, which was used not only as a black box approach, but also to understand the underlying relationships by identifying relevant variables and metabolites through variable selection. We demonstrated that ^{1}H NMR spectroscopy is a promising tool for the simultaneous investigation of different authentication issues, as it is possible to identify regions of the spectrum that are exclusively relevant to individual issues. However, for successful classification, it is important that the variance of each group in the model is adequately represented by a sufficient sample size in the training data. This was the case in this study, for example, for the differentiation between German and non-German samples, which is why a high accuracy could be achieved for this authentication question. Further experiments should therefore be carried out on the smaller groups in this study to confirm the potential of NMR spectroscopy for apple authentication. In addition, other analytical methods that generate complex data, e.g. based on mass spectrometry, should be tested for their suitability for classification according to different authentication issues on the basis of a single data set.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Soeren Wenck: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **René Bachmann:** Writing – review & editing, Resources, Investigation, Funding acquisition, Data curation, Conceptualization. **Sarah-Marie Barmbold:** Resources, Investigation. **Anna Lena Horns:** Resources, Investigation. **Nele Paasch:** Resources, Investigation. **Stephan Seifert:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Deepl in order to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

The authors thank Lyn Christiansen, Frederik Lütjen, Eileen Bahnsen, Johanna Greive, Melike Kaya, Isabel-Monique Moreau and Annik Krohn for their support in sample preparation and Julia Kaeswurm and Maria Buchweitz for helpful discussions. The authors would also like to thank all farms, groups and individuals who provided samples and metadata. We acknowledge financial support from the Open Access Publication Fund of Universität Hamburg.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2024.110817.

References

Aprea, E., Charles, M., Endrizzi, I., Laura Corollaro, M., Betta, E., Biasioli, F., & Gasperi, F. (2017). Sweet taste in apple: The role of sorbitol, individual sugars, organic acids and volatile compounds. *Scientific Reports, 7*(1), Article 44950. https://doi.org/10.1038/srep44950

Bachmann, R., Klockmann, S., Haerdter, J., Fischer, M., & Hackl, T. (2018). ^{1}H-NMR spectroscopy for determination of the geographical origin of hazelnuts. *Journal of Agricultural and Food Chemistry, 66*(44), Article 44. https://doi.org/10.1021/acs.jafc.8b03724

Bai, J., Wu, Y., Bu, Q., Zhong, K., & Gao, H. (2022). Comparative study on antibacterial mechanism of shikimic acid and quinic acid against Staphylococcus aureus through transcriptomic and metabolomic approaches. *Lebensmittel-Wissenschaft & Technologie, 153*, Article 112441. https://doi.org/10.1016/j.lwt.2021.112441

Bannier, H.-J. (2011). Moderne Apfelzüchtung: Genetische Verarmung und Tendenzen zur Inzucht: Vitalitätsverluste erst bei Verzicht auf Fungizideinsatz sichtbar. *Erwerbs-obstbau, 52*(3–4), 85–110. https://doi.org/10.1007/s10341-010-0113-4

Baric, S., Storti, A., Hofer, M., Guerra, W., & Dalla Via, J. (2020). Molecular genetic identification of apple cultivars based on microsatellite DNA analysis. I. The database of 600 validated profiles. *Erwerbs-obstbau, 62*(2), 117–154. https://doi.org/10.1007/s10341-020-00483-0

Becker, S., Becker, S., Chebib, S., Schwab, W., Dierend, W., Zuberbier, T., & Bergmann, K.-C. (2021). Die Testung von Äpfeln auf ihre Allergenität. *Erwerbs-obstbau, 63*(4), 409–415. https://doi.org/10.1007/s10341-021-00600-7

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bingol, K., & Brüschweiler, R. (2014). Multidimensional approaches to NMR-based metabolomics. *Analytical Chemistry, 86*(1), 47–57. https://doi.org/10.1021/ac403520j

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152). https://doi.org/10.1145/130385.130401

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods, 6*(9), 2812–2831. https://doi.org/10.1039/C3AY41907J

Chitarrini, G., Dordevic, N., Guerra, W., Robatscher, P., & Lozano, L. (2020). Aroma investigation of New and standard apple varieties grown at two altitudes using gas chromatography-mass spectrometry combined with sensory analysis. *Molecules, 25*(13), 3007. https://doi.org/10.3390/molecules25133007

Cho, A.-S., Jeon, S.-M., Kim, M.-J., Yeo, J., Seo, K.-I., Choi, M.-S., & Lee, M.-K. (2010). Chlorogenic acid exhibits anti-obesity property and improves lipid metabolism in high-fat diet-induced-obese mice. *Food and Chemical Toxicology, 48*(3), 937–943. https://doi.org/10.1016/j.fct.2010.01.003

Creydt, M., Hudzik, D., Rurik, M., Kohlbacher, O., & Fischer, M. (2018). Food authentication: Small-molecule profiling as a tool for the geographic discrimination of German white Asparagus. *Journal of Agricultural and Food Chemistry, 66*(50), 13328–13339. https://doi.org/10.1021/acs.jafc.8b05791

Denver, S., & Jensen, J. D. (2014). Consumer preferences for organically and locally produced apples. *Food Quality and Preference, 31*, 129–134. https://doi.org/10.1016/j.foodqual.2013.08.014

Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., Wang, X., Jiao, C., Legall, N., Mao, L., Wan, S., Wang, K., He, T., Feng, S., Zhang, Z., Mao, Z., Shen, X., Chen, X., & Jiang, Y.

(2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nature Communications, 8*(1), 249. https://doi.org/10.1038/s41467-017-00336-7

Eisenmann, P., Ehlers, M., Weinert, C., Tzvetkova, P., Silber, M., Rist, M., Luy, B., & Muhle-Goll, C. (2016). Untargeted NMR spectroscopic analysis of the metabolic variety of New apple cultivars. *Metabolites, 6*(3), 29. https://doi.org/10.3390/metabo6030029

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., & Wishart, D. S. (2019). NMR spectroscopy for metabolomics research. *Metabolites, 9*(7), 123. https://doi.org/10.3390/metabo9070123

Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology, 48*(1/2), Article 1. https://doi.org/10.1023/A:1013713905833

Hansen, J., Kunert, C., Münstermann, H., Raezke, K.-P. R., & Seifert, S. (2024). Application of untargeted liquid chromatography-mass spectrometry to routine analysis of food using three-dimensional bucketing and machine learning. *Scientific Reports, 14*, Article 16594. https://doi.org/10.1038/s41598-024-67459-y

Ishwaran, H., Kogalur, U. B., Chen, X., & Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 4*(1), 115–132. https://doi.org/10.1002/sam.10103

Jiang, N., Song, W., Wang, H., Guo, G., & Liu, Y. (2018). Differentiation between organic and non-organic apples using diffraction grating and image processing—a cost-effective approach. *Sensors, 18*(6), 1667. https://doi.org/10.3390/s18061667

Kaeswurm, J. A. H., Straub, L. V., Siegele, A., Brockmeyer, J., & Buchweitz, M. (2023). Characterization and quantification of mal d 1 isoallergen profiles and contents in traditional and commercial apple varieties by mass spectrometry. *Journal of Agricultural and Food Chemistry, 71*(5), 2554–2565. https://doi.org/10.1021/acs.jafc.2c05801

Kaur, C., & Kapoor, H. C. (2001). Antioxidants in fruits and vegetables – the millennium's health. *International Journal of Food Science and Technology, 36*(7), 703–725. https://doi.org/10.1111/j.1365-2621.2001.00513.x

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), Article 11. https://doi.org/10.18637/jss.v036.i11

Li, X., Ma, Z., Song, Y., Shen, W., Yue, Q., Khan, A., Tahir, M. M., Wang, X., Malnoy, M., Ma, F., Bus, V., Zhou, S., & Guan, Q. (2023). Insights into the molecular mechanisms underlying responses of apple trees to abiotic stresses. *Horticulture Research, 10*(8), uhad144. https://doi.org/10.1093/hr/uhad144

Li, Y., Sun, H., Li, J., Qin, S., Niu, Z., Qiao, X., & Yang, B. (2021). Influence of genetic background, growth latitude and bagging treatment on phenolic compounds in fruits of commercial cultivars and wild types of apples (Malus sp.). *European Food Research and Technology, 247*(5), 1149–1165. https://doi.org/10.1007/s00217-021-03695-0

Li, Y., Sun, H., Li, J., Qin, S., Yang, W., Ma, X., Qiao, X., & Yang, B. (2021). Effects of genetic background and altitude on sugars, malic acid and ascorbic acid in fruits of wild and cultivated apples (Malus sp.). *Foods, 10*(12), 2950. https://doi.org/10.3390/foods10122950

Liao, L., Zhang, W., Zhang, B., Cai, Y., Gao, L., Ogutu, C., Sun, J., Zheng, B., Wang, L., Li, L., & Han, Y. (2021). Evaluation of chlorogenic acid accumulation in cultivated and wild apples. *Journal of Food Composition and Analysis, 104*, 104156. https://doi.org/10.1016/j.jfca.2021.104156

Lösel, H., Brockelt, J., Gärber, F., Teipel, J., Kuballa, T., Seifert, S., & Fischer, M. (2023). Comparative analysis of LC-ESI-IM-qToF-MS and FT-NIR spectroscopy approaches for the authentication of organic and conventional eggs. *Metabolites, 13*(8), 882. https://doi.org/10.3390/metabo13080882

Maalik, A., Bukhari, S. M., Zaidi, A., Shah, K. H., & Khan, F. A. (2016). Chlorogenic acid: A pharmacologically potent molecule. *Acta Poloniae Pharmaceutica, 73*(4), 851–854.

Markley, J. L., Brüschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., & Wishart, D. S. (2017). The future of NMR-based metabolomics. *Current Opinion in Biotechnology, 43*, 34–40. https://doi.org/10.1016/j.copbio.2016.08.001

Mendez, K. M., Broadhurst, D. I., & Reinke, S. N. (2019). The application of artificial neural networks in metabolomics: A historical perspective. *Metabolomics, 15*(11), 142. https://doi.org/10.1007/s11306-019-1608-0

Mix, T., Janneschütz, J., Ludwig, R., Eichbaum, J., Fischer, M., & Hackl, T. (2023). From nontargeted to targeted analysis: Feature selection in the differentiation of truffle species (*tuber* spp.) using ^{1}H NMR spectroscopy and support vector machine. *Journal of Agricultural and Food Chemistry, 71*(46), 18074–18084. https://doi.org/10.1021/acs.jafc.3c05786

Pandey, A. K., Buchholz, C. R., Nathan Kochen, N., Pomerantz, W. C. K., Braun, A. R., & Sachs, J. N. (2023). pH effects can dominate chemical shift perturbations in ^{1}H, ^{15}N-HSQC NMR spectroscopy for studies of small molecule/α-synuclein interactions. *ACS Chemical Neuroscience, 14*(4), 800–808. https://doi.org/10.1021/acschemneuro.2c00782

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research, 20*(53), 1–32. https://doi.org/10.48550/arXiv.1802.09596

Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R., & Wishart, D. S. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One, 10*(5), Article e0124219. https://doi.org/10.1371/journal.pone.0124219

Saarenhovi, M., Salo, P., Scheinin, M., Lehto, J., Lovró, Z., Tiihonen, K., Lehtinen, M. J., Junnila, J., Hasselwander, O., Tarpila, A., & Raitakari, O. T. (2017). The effect of an apple polyphenol extract rich in epicatechin and flavan-3-ol oligomers on brachial artery flow-mediated vasodilatory function in volunteers with elevated blood pressure. *Nutrition Journal, 16*(1), 73. https://doi.org/10.1186/s12937-017-0291-0

Santana-Gálvez, J., Cisneros-Zevallos, L., & Jacobo-Velázquez, D. (2017). Chlorogenic acid: Recent advances on its dual role as a food additive and a nutraceutical against metabolic syndrome. *Molecules, 22*(3), 358. https://doi.org/10.3390/molecules22030358

Schimmel, J., Gentsch, N., Boy, J., Uteau, D., Rohr, A.-D., Winkelmann, T., Busnena, B., Liu, B., Krueger, J., Kaufhold, S., Rammlmair, D., Dultz, S., Maurischat, P., Beerhues, P., & Guggenberger, G. (2024). Alleviation of apple replant disease in sandy soils by clay amendments. *Silicon*. https://doi.org/10.1007/s12633-024-03002-y

Schönberger, T., Bachmann, R., Gerhardt, N., Panzer, J., Meyer, K., Romoth, M., Teipel, J., Scharinger, A., Weber, M., Kuballa, T., Maiwald, M., Esslinger, S., Fauhl-Hassek, C., Horn, B., Riedl, J., Becker, R., Annweiler, E., Meier, M., & Ohmenhäuser, M.. Guide to NMR method development and validation – Part I: Identification and quantification (update 2023). https://doi.org/10.13140/RG.2.2.30200.83208.

Seifert, S. (2020). Application of random forest based approaches to surface-enhanced Raman scattering data. *Scientific Reports, 10*(1), Article 1. https://doi.org/10.1038/s41598-020-62338-8

Seifert, S., Gundlach, S., & Szymczak, S. (2019). Surrogate minimal depth as an importance measure for variables in random forests. *Bioinformatics, 35*(19), Article 19. https://doi.org/10.1093/bioinformatics/btz149

Shakiba, N., Gerdes, A., Holz, N., Wenck, S., Bachmann, R., Schneider, T., Seifert, S., Fischer, M., & Hackl, T.. Determination of the geographical origin of hazelnuts (Corylus avellana L.) by near-infrared spectroscopy (NIR) and a low-level fusion with nuclear magnetic resonance (NMR) [Preprint]. https://doi.org/10.33774/chemrxiv-2021-1cr4g.

Siebert, M., Berger, R. G., & Pfeiffer, F. (2019). Hydrolysis of chlorogenic acid in apple juice using a *p* -coumaryl esterase of *Rhizoctonia solani*. *Journal of the Science of Food and Agriculture, 99*(14), 6644–6648. https://doi.org/10.1002/jsfa.9940

Siekierzynska, A., Piasecka-Kwiatkowska, D., Myszka, A., Burzynska, M., Sozanska, B., & Sozanski, T. (2021). Apple allergy: Causes and factors influencing fruits allergenic properties–Review. *Clinical and Translational Allergy, 11*(4), Article e12032. https://doi.org/10.1002/clt2.12032

Voges, L. F., Jarren, L. C., & Seifert, S. (2023). Exploitation of surrogate variables in random forests for unbiased analysis of mutual impact and importance of features. *Bioinformatics, 39*(8). https://doi.org/10.1093/bioinformatics/btad471

Wenck, S., Creydt, M., Hansen, J., Gärber, F., Fischer, M., & Seifert, S. (2021). Opening the random forest black box of the metabolome by the application of surrogate minimal depth. *Metabolites, 12*(1), 5. https://doi.org/10.3390/metabo12010005

Wenck, S., Mix, T., Fischer, M., Hackl, T., & Seifert, S. (2023). Opening the random forest black box of 1H NMR metabolomics data by the exploitation of surrogate variables. *Metabolites, 13*(10), 1075. https://doi.org/10.3390/metabo13101075

Wishart, D. S. (2007). Current progress in computational metabolomics. *Briefings in Bioinformatics, 8*(5), Article 5. https://doi.org/10.1093/bib/bbm030

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Worley, B., & Powers, R. (2012). Multivariate analysis in metabolomics. *Current Metabolomics, 1*(1), Article 1. https://doi.org/10.2174/2213235X11301010092

Yan, Z., Zheng, L., Nie, J., Li, Z., & Cheng, Y. (2018). Evaluation indices of sour flavor for apple fruit and grading standards. *Journal of Integrative Agriculture, 17*(5), 994–1002. https://doi.org/10.1016/S2095-3119(17)61795-7

Zhang, J., Nie, J., Kuang, L., Shen, Y., Zheng, H., Zhang, H., Farooq, S., & Asim, S. (2019). Geographical origin of Chinese apples based on multiple element analysis. *Journal of the Science of Food and Agriculture, 99*(14), 6182–6190. https://doi.org/10.1002/jsfa.9890

Zhang, J., Shen, Y., Ma, N., & Xu, G. (2023). Authentication of apples from the Loess Plateau in China based on interannual element fingerprints and multidimensional modelling. *Food Chemistry X, 20*, Article 100948. https://doi.org/10.1016/j.fochx.2023.100948

Zhang, M., Wang, D., Gao, X., Yue, Z., & Zhou, H. (2020). Exogenous caffeic acid and epicatechin enhance resistance against Botrytis cinerea through activation of the phenylpropanoid pathway in apples. *Scientia Horticulturae, 268*, Article 109348. https://doi.org/10.1016/j.scienta.2020.109348

Živanović, V., Seifert, S., Drescher, D., Schrade, P., Werner, S., Guttmann, P., Szekeres, G. P., Bachmann, S., Schneider, G., Arenz, C., & Kneipp, J. (2019). Optical nanosensing of lipid accumulation due to enzyme inhibition in live cells. *ACS Nano, 13*(8), 8. https://doi.org/10.1021/acsnano.9b04001

Website References

FAO Statista – Food and Agriculture Statistics. (2023). Global fruit production in 2022, by selected variety (in million metric tons). Retrieved from https://www.statista.com/statistics/264001/worldwide-production-of-fruit-by-variety/. (Accessed 26 July 2024).

Statista. (2023). Das Lieblingsobst der Deutschen. Retrieved from https://de.statista.com/infografik/12081/lieblingsobst/. (Accessed 26 July 2024).

Ble. (2020). *Bundesanstalt für Landwirtschaft und Ernährung*. Äpfel in Kürze. Retrieved from https://www.ble.de/SharedDocs/Downloads/DE/Ernaehrung-Lebens

mittel/Vermarktungsnormen/VermarktungsnormenObstGemuese/Flyer/Aepfel.pdf?_blob=publicationFile&v=1. (Accessed 26 July 2024).
UBA – Umweltbundesamt. (2015). Bodenzustand in Deutschland zum "Internationalen Jahr des Bodens., Page 7, Retrieved from https://www.umweltbundesamt.de/sites/default/files/medien/378/publikationen/bodenzustand_in_deutschland_0.pdf. (Accessed 26 July 2024).
Vink, T. (2005). Western Cape agricultural sector: Structure, performance and future prospects. Retrieved from https://www.westerncape.gov.za/other/2005/10/final_first_paper_overview_agriculture.pdf. (Accessed 26 July 2024).

5.2. Klassifizierung und Charakterisierung von Speisetrüffeln (*Tuber* sp.) mit ^{1}H Kernresonanz-Spektroskopie

In der zweiten Publikation wurden die ^{1}H NMR-Daten von 80 authentischen Proben der fünf relevantesten Speisetrüffelarten mit RF-Methoden untersucht. Die Klassifikation erreichte eine Vorhersagegenauigkeit von 100 %, was mit den Ergebnissen von Mix et al. übereinstimmt, die durch Anwendung von *support vector machines* erreicht wurde.[63] Durch die Variablenselektion mit Boruta wurden 341 Variablen und mit SMD 210 Variablen selektiert, wobei 209 Variablen von beiden Verfahren selektiert wurden. Da Boruta im Gegensatz zu SMD Variablen lediglich einzeln hinsichtlich ihrer Relevanz bewertet, ist davon auszugehen, dass die 132 ausschließlich von Boruta selektierten Variablen vergleichsweise individuelle Informationen für die Klassifizierung liefern, was in der Publikation bestätigt werden konnte.[116,118] Der gemeinsame Einfluss der selektierten Variablen wurde anschließend mit dem SMD Parameter MAA analysiert und es wurden verschiedene Ebenen bezüglich der Variablenbeziehungen identifiziert, welche im Folgenden ausgeführt werden sollen.

Die stärksten Variablenbeziehungen zeigten jeweils benachbarte oder recht nah beieinander liegende Variablen, die jeweils auf das gleiche Multiplet Signal im Spektrum, welches bei der Präprozessierung in verschiedene Buckets geteilt wurde, zurückgeführt werden konnte. Darüber hinaus konnten intramolekulare Zusammenhänge aufgedeckt und mittels ^{1}H-^{1}H TOCSY-NMR-Spektren und *spike-in*-Experimenten bestätigt und somit zur Metabolit-Identifizierung genutzt werden. Darüber hinaus konnten Gruppen von Metaboliten identifiziert werden, die erhöhte Werte für den Beziehungsparameter aufwiesen und damit einen ähnlichen Einfluss auf das Klassifizierungsmodell haben. Diese Gruppen enthielten chemisch ähnliche Moleküle, z.B. verschiedene ähnliche Aminosäuren aber auch Metabolite, deren Wechselwirkungen in Stoffwechselwegen bekannt sind, z.b. im Citratzyclus oder Kohlenhydratmetabolismus.

Allgemein konnten wir in dieser Publikation zeigen, dass SMD ein leistungsstarkes Werkzeug für die Untersuchung von NMR-*metabolomics*-Daten darstellt, welches

sowohl für die Identifizierung der Metabolite als auch für die umfassende Analyse deren Wechselwirkung angewendet werden kann.

metabolites

MDPI

Article

Opening the Random Forest Black Box of ^{1}H NMR Metabolomics Data by the Exploitation of Surrogate Variables

Soeren Wenck [1,†], Thorsten Mix [2,†], Markus Fischer [1], Thomas Hackl [1,2] and Stephan Seifert [1,*]

[1] Institute of Food Chemistry, Hamburg School of Food Science, University of Hamburg, Grindelallee 117, 20146 Hamburg, Germany; markus.fischer@uni-hamburg.de (M.F.); thomas.hackl@uni-hamburg.de (T.H.)
[2] Institute of Organic Chemistry, University of Hamburg, Martin-Luther-King-Platz 6, 20146 Hamburg, Germany; thorsten.mix@uni-hamburg.de
* Correspondence: stephan.seifert@uni-hamburg.de; Tel.: +49-40-42838-8818
† These authors contributed equally to this work.

Abstract: The untargeted metabolomics analysis of biological samples with nuclear magnetic resonance (NMR) provides highly complex data containing various signals from different molecules. To use these data for classification, e.g., in the context of food authentication, machine learning methods are used. These methods are usually applied as a black box, which means that no information about the complex relationships between the variables and the outcome is obtained. In this study, we show that the random forest-based approach surrogate minimal depth (SMD) can be applied for a comprehensive analysis of class-specific differences by selecting relevant variables and analyzing their mutual impact on the classification model of different truffle species. SMD allows the assignment of variables from the same metabolites as well as the detection of interactions between different metabolites that can be attributed to known biological relationships.

Keywords: classification; characterization; nuclear magnetic resonance spectroscopy; random forest; variable selection; variable relations; machine learning; chemometrics; surrogate minimal depth; truffles

Citation: Wenck, S.; Mix, T.; Fischer, M.; Hackl, T.; Seifert, S. Opening the Random Forest Black Box of ^{1}H NMR Metabolomics Data by the Exploitation of Surrogate Variables. *Metabolites* **2023**, *13*, 1075. https://doi.org/10.3390/metabo13101075

Academic Editor: Junsong Wang

Received: 18 September 2023
Revised: 5 October 2023
Accepted: 10 October 2023
Published: 13 October 2023

1. Introduction

Metabolomics is the research field that aims at the comprehensive analysis of metabolites, which are small molecules (<1500 Da) within biological organisms. Metabolites take part in cellular regulatory processes and are influenced by both endogenous factors such as the genotype and exogeneous factors such as climate, soil composition, distance to large bodies of waters, and fertilization [1]. Thus, the metabolome is the best representation of the phenotype [2]. Since there is no approach that can capture the entire metabolome, various combinations of extraction and measurement techniques have been introduced through which different parts of the metabolome can be analyzed [3]. Many of these analytical methods are based on nuclear magnetic resonance (NMR) and mass spectrometry (MS) platforms [4–7].

^{1}H NMR combines highly repeatable and reproducible non-destructive data acquisition, simultaneous structural elucidation and quantitative analysis of compounds. However, interpreting NMR spectra of biological samples is difficult, since they contain hundreds of signals from several dozens of metabolites [7–11]. For this reason, assigning signals to specific molecules is not straightforward and usually requires individual strategies. A number of databases and tools are available, such as the Human Metabolome Database (HMDB), the Biological Magnetic Resonance Database (BMRB) or the Chenomx software [12]. In addition to some inherent errors that can occur in any database, experimental conditions such as solvent, pH, or ionic strength have a huge impact on chemical shifts and make the exclusive use of databases difficult, leading to unreliable assignments. Besides the standard 2D NMR methods, such as TOCSY or HSQC, some classical experiments, such as *J*-resolved NMR or 1D methods, such as selective TOCSY or NOESY have gained new popularity [13].

The combination of different experiments increases the likelihood of identifying additional metabolites, and the combination of NMR and MS is a promising approach to identifying compounds of interest because these analytical techniques offer complementary information. Recently published cheminformatics combinations of NMR and MS are the NMR/MS translator [14] and the SUMMIT MS/NMR method [15]. The correlation between NMR and MS data can be established when these techniques are used in combination with liquid chromatography, which has been demonstrated through approaches such as parallel dynamic NMR/LC-MS spectroscopy (NMR/LC-MS PDS) [16] or the Semi-automatic COrrelation analysis for REliable metabolite IDentification (SCORE-metabolite-ID) [17]. Typically, spike-in experiments with either purchased or synthesized reference compounds are performed on the mixture samples to verify the proposed structures.

NMR data can be analyzed using a technique called Statistical Total Correlation Spectroscopy (STOCSY) to detect correlated NMR signals based on structural connectivity or intermolecular correlations resulting from the connectivity of metabolic pathways in biological systems [18]. However, STOCSY and other statistics-based approaches require large sample sets for analysis and cannot distinguish between different types of correlation. Statistical heterospectroscopy (SHY) is another approach that is based on STOCSY but uses a combination of NMR and MS data [19].

The analysis of NMR metabolomics data is usually performed by either fitting patterns of signals from expected metabolites to spectral regions within the data or binning [1,7,20]. The latter is usually applied to aligned spectra to reduce the chemical shift variety and to achieve comparability among different spectra [21]. Since NMR data sets are high-dimensional, meaning that they contain many variables from comparatively few samples, multivariate approaches have to be applied for data analysis [22]. The popular unsupervised approach principal component analysis (PCA) creates latent variables by linear combinations of the original variables. These principal components are focused on the main variances of the data and can enable the identification of groups with similar patterns [23–25]. In contrast to unsupervised approaches, supervised machine learning algorithms such as support vector machines (SVM) [26], artificial neural networks (ANN) [27], and random forests (RF) include the group affiliation of samples in the analysis and train classification models based on specific class differences.

RF is a non-parametric ensemble learning algorithm based on multiple binary decision trees that offers many advantages for application to high-dimensional data, such as the inherent independent validation [28,29]. This validation is based on the fact that each of the decision trees is trained on a different fraction of the samples, the so-called bootstrap samples, while the respective remaining samples are used to generate independent out-of-bag errors. Another advantage of RF is that it can also be used to generate variable importance scores. These scores are, for example, based on the decrease of accuracy obtained by the permutation of a variable or on the decrease of Gini impurity calculated by the summarized Gini gains, a variable is contributing to the RF. Variable selection methods use these importance scores to separate important from unimportant variables, and various approaches that differ in the way in which they define the threshold between important and unimportant variables have been developed. Boruta creates shadow variables by random permutation and evaluates whether the real variables generally show higher importance scores than the highest scores of the shadow variables [30]. Surrogate Minimal Depth (SMD) is a variable importance score and selection approach that incorporates variable relations into the selection process [31]. This is achieved by the combination of minimal depth [32], an importance measure based on the first appearance of variables in decision trees, with surrogate variables, which were originally introduced by Breimann et al. [28] for the compensation of missing variables. SMD thus determines the variable importance measure not only by considering primary split variables but also surrogate variables. In addition to variable selection, SMD can also be applied to calculate the relation parameter *mean adjusted agreement*, analyzing the mutual impact of the variables on the random forest model. This relation parameter, which has recently been further developed to also

analyze qualitative variables [33], enables a comprehensive analysis of the interplay of the relevant variables. It has been successfully applied in various fields and to different types of data, including gene expression [31], surface-enhanced Raman scattering [34,35], FT-NIR [5], and LC MS data [36], as well as to analyze relations across the latter two analytical techniques [37].

Here, we apply SMD to ^{1}H NMR metabolomics data for the first time and show that it can reveal various relationships between predictor variables and outcome, as well as between predictor variables. More precisely, buckets containing information from the same signals and molecules can be identified, and meaningful biological relations between different metabolites can be determined and utilized for the investigation of specific class differences. As a model data set, we use data from truffle samples as the truffle species show a clear distinction and, thus, a comparatively simple interpretation of the selected markers and observed differences is possible [38]. Due to limited harvest periods, difficult cultivation, and their unique aromatic properties, truffles are one of the most expensive foods and, hence, prone to food fraud [39,40].

2. Materials and Methods

2.1. Samples and Data Acquisition

The ^{1}H NMR data set used in this study contained 80 samples from five different *Tuber* species (see Table 1) and is provided in Table S1. For detailed information about the measurement and preprocessing of the data, please refer to Mix et al. [38]. However, the data utilized here adopted a bucket width of 0.01 ppm, whereas Mix et al. opted for a width of 0.03 ppm. In addition to the ^{1}H NMR measurement, every sample was analyzed with ^{1}H-^{1}H TOCSY. The measurement was conducted with the *dipsi2esgpph* (Bruker notation) pulse sequence. Homonuclear Hartman-Hahn transfer using *DIPSI2* (Bruker notation) sequence for mixing was performed. The data were collected with a spectral width of 4401.4 Hz. The spin-locking field of 8.3 KHz was generated with a 30 µs pulse at a power of −2.5 dB. Eight scans per increment in a matrix of 2048 × 256 were obtained with a mixing time of 60 ms, and the data were zero-filled to 2048 × 512. To generate phase-sensitive data, the States-*TPPI* phase cycling was used. The data were processed with a *QSINE* function in both dimensions and a Sine Bell Shift (SSB) of 2. The parameter set *dipsi2esgpph* (Bruker notation) was applied in accordance with Shaka et al. for water suppression [41].

Table 1. Overview of the truffle samples used in this study.

	T. aestivum	*T. borchii*	*T. indicum*	*T. magnatum*	*T. melanosporum*
Amount	28	7	12	21	12
Color	black	white	black	white	black

2.2. Identification of Truffle Metabolites

The identification of metabolites was carried out according to Mix et al. [38] by column chromatographic fractionation of the mixture and subsequent analysis of the fractions by NMR and MS techniques. The NMR and MS signals were correlated manually or using the SCORE-metabolite-ID app [17]. For the verification of proposed structures, spike-in experiments were performed in which 10 to 200 µg of a specific metabolite was added to one of the sample fractions containing the corresponding metabolite. The mixtures were remeasured with the pulse program noesygppr1d (Bruker notation) at 300 K. For visual clarity, the measurements were conducted at 400 MHz or 600 MHz (Ribonate) and with 32 or 64 scans with TMSP as an internal standard. An increase in the signal intensity confirmed the spiked metabolite in the spectrum [42].

2.3. Software and Data Analysis

Data acquisition was performed with Topspin (version 4.0.94) and bucketing with Aurelia Amix (version 3.9.15). The software R (version 3.6.3) and the R packages ranger

(version 0.14.1, CRAN) for RF classification [43], mdatools (version 0.12.0, CRAN) for PCA [44], Pomona (version 1.0.1, https://github.com/silkeszy/Pomona, accessed on 11 October 2023) for Boruta variable selection [45], and SurrogateMinimalDepth (version 0.2.0, https://github.com/StephanSeifert/SurrogateMinimalDepth, accessed on 11 October 2023) for SMD variable selection and relation analysis were used [31]. Figures were created with ggplot2 (version 3.4.0, CRAN) [46] and heatmaps with pheatmap (version 1.0.12, CRAN, https://CRAN.R-project.org/package=pheatmap, accessed on 11 October 2023) [47].

The RF approaches were applied in classification mode with the parameters listed in Table 2. Due to the imbalance of the classes, the samples were weighted accordingly using the parameter case.weights. The variable relation analysis was performed on variables selected by Boruta and SMD, analyzing relationships that were assigned to the same signal and those that corresponded to different signals and metabolites. For the latter, a hierarchical cluster analysis with Euclidean distance measure and Ward's algorithm [48] was applied. For the clarity of this analysis, the variables of the same signals covering multiple buckets were reduced to one representative each, which was chosen by the lowest surrogate minimal depth value, i.e., the highest importance. In addition, the variables that could not be identified clearly were also removed from this analysis.

Table 2. Parameters used for RF-based approaches with p representing the total number of variables.

Approach	Parameter	Description	Value
RF	ntree	number of trees	10,000
	min.node.size	number of samples in terminal node	1
	mtry	number of candidate variables	157 ($p^{3/4}$) [1]
	case. weights	weights for sampling of training observations	chosen according to the size of the respective class
SMD	s	Predefined number of surrogate splits	42 ($p \cdot 0.05$)
Boruta	pValue	applied importance measure	impurity_corrected
	importance	confidence level	0.01
	maxRuns	maximum number of importance source runs	157 ($p^{3/4}$) [1]

[1] Motivated by [32].

3. Results and Discussion

3.1. Classification of Truffle Samples

The main objective of this study was to open the black box of the ^{1}H NMR metabolome by the application of random forest-based approaches. For this, a data set with clear distinction between classes was needed and we applied random forest on the truffle data containing 80 samples from five different species to verify whether this was the case. The confusion matrix of the classification results is shown in Table 3, showing an accuracy of 100% confirming the prerequisites formulated above and the previous classification results that were obtained by support vector machines [38]. These clear differences between the truffle species are only partially evident from the results of the unsupervised principal component analysis, demonstrating that supervised approaches should be applied for classification (see Figures 1 and S1).

Table 3. Result of the random forest classification of truffle samples. An out-of-bag error of 0% corresponding with a classification accuracy of 100% was obtained.

	T. aestivum	*T. borchii*	*T. indicum*	*T. magnatum*	*T. melanosporum*	**Sensitivity [%]**
T. aestivum	28	0	0	0	0	100
T. borchii	0	7	0	0	0	100
T. indicum	0	0	12	0	0	100
T. magnatum	0	0	0	21	0	100
T. melanosporum	0	0	0	0	12	100
Specificity [%]	100	100	100	100	100	

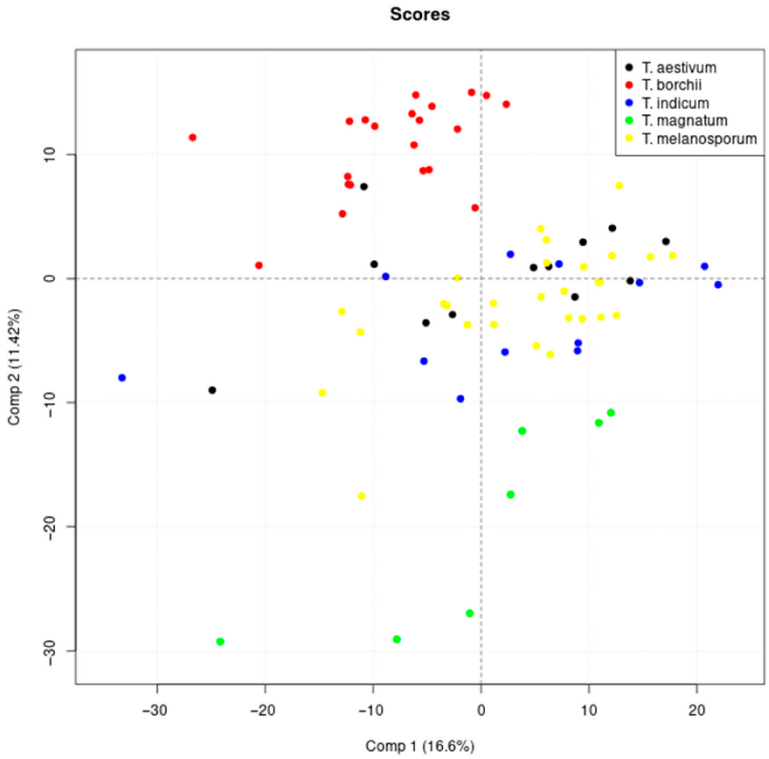


Figure 1. Results of the principal component analysis: Scores of the first and second principal components are shown.

3.2. Bucket Assignment for Truffle Metabolites

In principle, knowledge of the underlying metabolites is not necessary for classification. However, it is essential for biological interpretation. We used a metabolite identification procedure described in [38]. Identification was carried out both independently of the SMD results, in particular by using the SCORE-metabolite-ID app and further NMR experiments, and especially when relationships between different buckets resulted from the SMD analysis. A total of 35 metabolites were identified. Based on fractionation by LC-MS-NMR correlation, the identities of all metabolites could be verified by spike-in experiments of the single fractions. Furthermore, as data from total extracts were used for classification and SMD analysis, spike-in experiments were also performed on the total extracts to clearly assign the corresponding buckets. The NMR spectra from these spike-in experiments are shown in Figures S10–S34. 23 of these metabolites were considered in the SMD analysis. They included amino acids (aspartic acid, asparagine, arginine, isoleucine, glutamic acid, glutamine, histidine, leucine, lysine, proline, threonine, tryptophan, and valine), carbohydrates (trehalose and ribonate), organic acids (citric, fumaric, and malic acid), uridine 5′-diphosphate-*N*-acetylglucoseamine (UDP-GlcNAc), betaine, choline-O-sulfate, and glycerophosphorylcholine (GPC).

3.3. Variable Selection

The first step on the way from black box classification to the comprehensive characterization of the metabolites involved is the selection of relevant variables by variable selection

approaches. For this, the two approaches SMD and Boruta were applied, selecting 210 and 341 variables, respectively. The selected variables are listed in Table S2. Many variables with high importance could be assigned to organic or amino acids and carbohydrates, e.g., fumaric acid, lysine, and trehalose. The latter is a major fungal carbohydrate in ectomycorrhizal fungi such as truffles that are, in addition to their role in carbohydrate storage, involved in various cellular processes not directly related to carbohydrate metabolism [49]. Figure 2 shows the overlap of the selected variables of the two approaches: SMD selected only one variable that was not selected by Boruta, while Boruta selected additional 132 variables. In principle, the two selection approaches have very different objectives: Boruta evaluates the importance of a variable individually, while SMD includes variable relations into the selection process analyzing their mutual impact. Hence, the variables that were selected only by Boruta should show comparatively low relations to other variables. This is confirmed when comparing the variable relations of both methods in Figures S2 and S3, because the variables selected only by Boruta show almost no relation to other variables. To further investigate the variables that contribute mutual information, the relationship parameter mean adjusted agreement generated by SMD is examined in more detail in the following section.

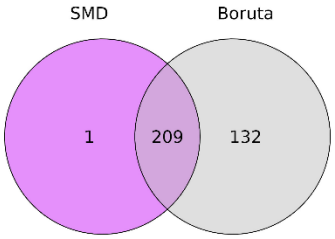


Figure 2. Venn diagram showing the overlap of variables selected by SMD and Boruta.

3.4. Analysis of Variable Relations

The obtained relations between the selected variables could be attributed to different causes. For clarity, these are discussed separately in the following sections.

3.4.1. Relations of Variables Containing the Same Signals

We frequently observed neighboring buckets with very high *mean adjusted agreement* values, often above 0.9. In Figure 3, this is shown exemplarily for the two spectral regions between 5.13 and 5.19 ppm and between 5.93 and 5.99 ppm, which were assigned to trehalose (see Figure S26) and UDP-GlcNAc (see Figure S12), respectively. It is obvious that the high mean adjusted agreement values are caused by the same respective multiplet signal that is present in multiple buckets. The linewidth of NMR signals is approximately between 0.7 and 3 Hz. A bucket size of 0.01 ppm corresponds exactly to 4 Hz. Thus, a single line can either lie exactly in one bucket or cross the bucket boundary into two adjacent buckets. Coupling constants range from 0 to 18 Hz. Thus, two lines belonging to the same signal may be separated by one to two buckets. Trehalose shows a doublet between 5.16 and 5.19 ppm and a coupling constant of 3.9 Hz (Figure 3b). As both lines are exactly on the bucket boundaries, the doublet extends over three buckets, which are highly related to each other and provide similar information to the classification model (Figure 3a), while the other buckets between 5.13 and 5.16 ppm mainly contain noise and show comparatively low relations. Similarly, the doublet of UDP-GlcNAc between 5.94 and 5.98 ppm, with a coupling constant of 8.1 Hz (Figure 3d), causes very strong relations of the respective buckets with each other (Figure 3c), while comparatively low relations occur to the buckets between 5.93 and 5.94 ppm as well as 5.98 and 5.99 ppm.

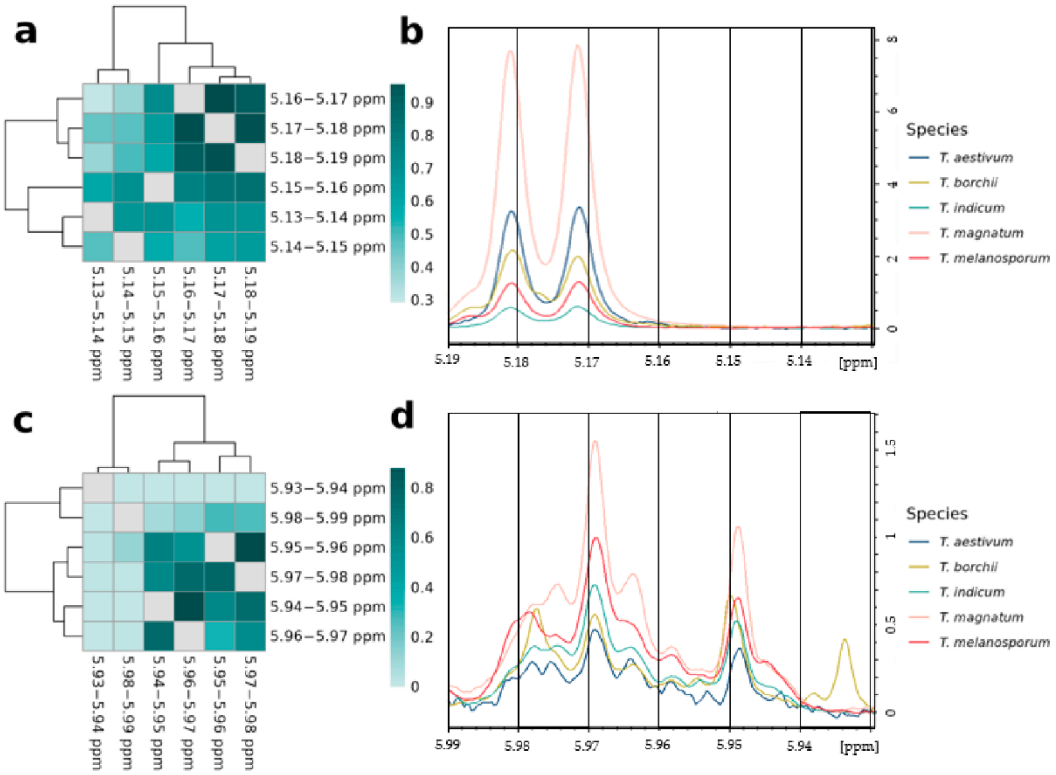


Figure 3. Analysis of adjacent variables from the same signals: Shown are heatmaps of mean adjusted agreement values and parts of the NMR spectra for the spectral regions between 5.13 and 5.19 ppm (**a**,**b**) and between 5.93 and 5.99 ppm (**c**,**d**). For the latter, one representative spectrum for each truffle species is shown and the black vertical lines show the limits of the buckets. For the heatmaps, cluster analysis with Euclidean distance measure and Ward's algorithm was applied.

We also observed variables with high mean adjusted agreement values that were not directly next to each other, but still very close together. This is shown by the two spectral regions between 7.95 and 8.00 ppm and 2.33 and 2.37 ppm in Figure 4. In the first region, there is a strong relation between the buckets at 7.98–7.99 ppm and 7.96–7.97 ppm, while the relation with the other variables in this area, including the variable between them at 7.97–7.98 ppm, is much weaker (Figure 4a). The reason for this is that the two subpeaks of a doublet assigned to UDP-GlcNAc (see Figure S12) populate exactly one bucket and are separated by a coupling constant of 8 Hz. The variable at 7.97–7.98 ppm does not contain any signal intensity from this doublet (Figure 4b).

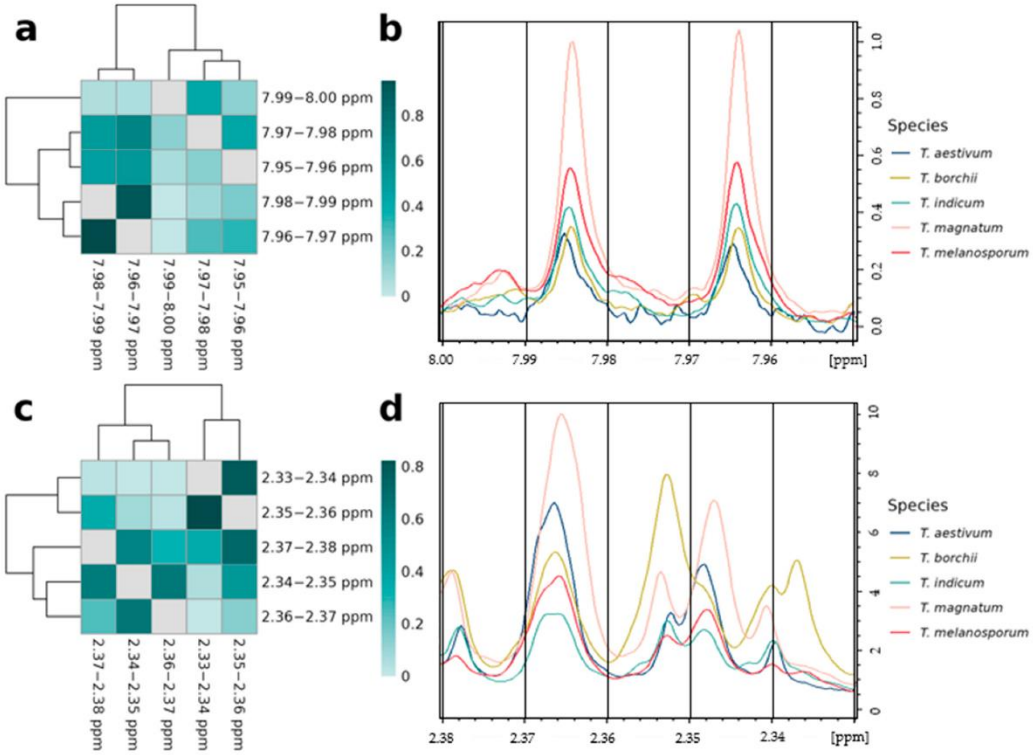


Figure 4. Analysis of close-by variables from the same signals: Heatmaps of mean adjusted agreement values and parts of the NMR spectra for the spectral regions between 7.95 and 8.00 ppm (**a**,**b**) and between 2.33 and 2.37 ppm (**c**,**d**). For the latter, one representative spectrum for each truffle species is shown and the black vertical lines show the limits of the buckets. For the heatmaps, cluster analysis with Euclidean distance measure and Ward's algorithm was applied.

For the spectral region between 2.33 and 2.37 ppm, two different clusters are built: the variables at 2.34–2.35 ppm and 2.36–2.37 ppm, assigned to a doublet of glutamic acid (see Figure S33), are strongly related to each other, while the other variables in the other cluster at 2.36–2.37 ppm, 2.34–2.35 ppm, and 2.37–2.38 ppm show slightly lower values for the relation parameter (Figure 4c). Hence, the glutamic acid doublet is overlapping with a second doublet, which is most pronounced at the buckets at 2.36–2.37 ppm and 2.34–2.35 ppm. That signals of two different metabolites are present here is also evident from the fact that the intensities of the truffle species are different: in the buckets 2.34–2.35 ppm and 2.36–2.37 ppm, the spectrum of *T. magnatum* is most intense, while *T. borchii* shows the most intensive peaks at 2.33–2.34 ppm and 2.35–2.36 ppm (Figure 4d).

3.4.2. Relations of Variables from the Same Metabolites

For the following relation analysis, the examined variables were reduced to the variables that could clearly be assigned to metabolites by the above-explained procedure. Furthermore, since the highly related variables of neighboring and close-by variables could be assigned to the same signals in the previous section, for clarity, only the respective most important variable was used for the analysis. Figure 5 shows the results of the relation analysis. In addition to four larger clusters, which are discussed in the following section,

it is apparent that small groups of variables with very high values for the relation parameter *mean adjusted agreement* (often above 0.9) are built. These relations can be attributed to intramolecular structural relationships and, hence, are assigned to the same metabolite. Specifically, the variables at 2.52–2.53 ppm and 2.68–2.69 ppm, 1.72–1.73 ppm and 1.94–1.95 ppm, 3.27–3.28 ppm and 3.87–3.88 ppm, as well as 5.97–5.98 ppm, 4.34–4.35 ppm, 7.98–7.99 ppm, and 5.51–5.51 ppm, are assigned to citric acid, arginine, betaine, and UDP-GlcNAc, respectively. We confirmed this finding by comparison to the ^{1}H-^{1}H TOCSY spectra, which are displayed in Figure 6. They show the coupling between the variables of citric acid (Figure 6a), arginine (Figure 6b), and UDP-GlcNAc (Figure 6d). The variables of betaine at 3.27–3.28 and 3.87–3.88 ppm (Figure 6c), however, do not show any coupling since the two signals are not part of the same spin system. The conducted spike-in experiments confirmed the presence of signals from these metabolites in the mentioned spectral regions (Figures S12, S15, S21 and S34). We can therefore conclude that the relationship analyses performed by SMD are consistent with the ^{1}H-^{1}H TOCSY experiment and are able to reveal chemical structure-based relationships. While ^{1}H-^{1}H TOCSY reveals chemical correlations within individual spin systems, the example of betaine shows that intramolecular relationships between different spin systems can also be made visible by the application of SMD.

The assignment of various variables to the same metabolite based on the SMD relation analysis is largely in agreement with the results of correlation analysis, which is usually applied for this purpose in STOCSY experiments (see Figure S4). However, the mean adjusted agreement values of variables of the same metabolite differ much more from those of different metabolites, which simplifies the assignment considerably.

Since the signals from multiple metabolites can be superimposed in individual buckets, it can be difficult to determine which molecules provide the relevant information for classification when only variable selection is performed. SMD relation analysis, however, can be applied to analyze these buckets in more detail: the variable at 3.23–3.24 ppm, for example, was associated with choline-O-sulfate, glycerophosphorylcholine (GPC), and arginine. While this variable shows high values of the relation parameter for another selected variable assigned to choline-O-sulfate at 4.49–4.50 ppm, the additional variables associated with GPC or arginine are characterized by relation values around zero. We can therefore assume that the classification-relevant information contained in the variable at 3.23–3.24 ppm originates from choline-O-sulfate. In contrast, the variables at 3.82–3.83 ppm and 3.41–3.42 ppm, which were assigned to trehalose and ribonate, and trehalose and proline, respectively, show relationships with both other variables assigned to trehalose, e.g., at 5.18–5.19 ppm, and variables at 4.13–4.14 ppm and 4.08–4.09 ppm assigned to ribonate and proline, respectively. Thus, in both cases, both metabolites are relevant for the classification. In summary, the parameter mean adjusted agreement for the analysis of variable relationships is a useful additional element to complement the toolbox for the identification of metabolites in authentication experiments.

3.4.3. Relations of Variables from Different Metabolites

In Figure 5, four clusters are built based on the mutual information the respective metabolites contribute for classification. This information can be examined in more detail in Figure 7, in which boxplots of exemplary variables of each cluster are displayed, and in Figures S4–S8, showing boxplots of all variables contained in the respective clusters.

Cluster I contains various variables with high intensities for *T. magnatum* (Figures 7I and S5). The high values for the *mean adjusted agreement* of UDP-GlcNAc and trehalose could be explained by the biosynthesis of chitin, in which both molecules are involved in [50], indicating a different cell wall composition of *T. magnatum*. The relations between signals from arginine, proline, and lysine could be explained by structural similarities because they are all amino acids with nitrogenous side chains. Since these variables also show strong relations to asparagine and aspartic acid, which are important nitrogen carriers in plants [51,52], this could indicate differences in amino acid metabolism, nitrogen assimilation, and growth of *T. magnatum*.

Figure 5. Result of the relation analysis of the identified variables. For the hierarchical cluster analysis, Euclidean distances and Ward's algorithm were applied and the clusters are labeled with I—IVa/b. The variables are labelled with the assigned metabolites, whereby the assignments, which play a rather minor role for the classification due to the relationship analysis, are shown in brackets (see discussion in Section 3.4.2). Abbreviations: GPC—Glycerophosphorylcholine; UDP-GlcNAc—uridine 5′-diphosphate-*N*-acetylglucoseamine.

Cluster II contains variables with specific classification information for *T. borchii* (see Figures 7II and S6). The variables assigned to malic and fumaric acid show very high values for the relation parameter, thus building a small subcluster. Since fumaric acid is converted to malic acid in the tricarboxylic acid cycle (TCA), this could indicate principal differences in the energy metabolism of *T. borchii*. In the fungus *Rhizopus arrhizus*, the

accumulation of malic and fumaric acid could be traced back to the TCA and glyoxylic acid pathway, which could also be the source of the enrichment in *T. borchii* [53,54]. However, the specific difference of *T. borchii* is not apparent from all selected variables of the TCA, and variables that are associated with citric acid are grouped in cluster I, providing vastly different information for the classification model (see Figures S5 and S6). This could be explained by the fact that citric acid acts as an intermediate, while both fumaric and malic acid act as main products. A variable at 4.13–4.14 ppm assigned to ribonate is also grouped in Cluster II. This is in accordance with our previous study because this metabolite, which is also related to energy metabolism, was identified as an exclusive marker for *T. borchii* [38]. In our analysis, it becomes apparent that high concentrations of ribonate are highly related to low concentrations of histidine in *T. borchii*. This could be explained by the presence of *Pseudomonas*, which are known to populate *T. borchii* [55], because they use histidine as a carbon source [56]. In summary, the metabolites of Cluster II show differences in the energy metabolism of *T. borchii*, which can be used to uniquely identify this species.

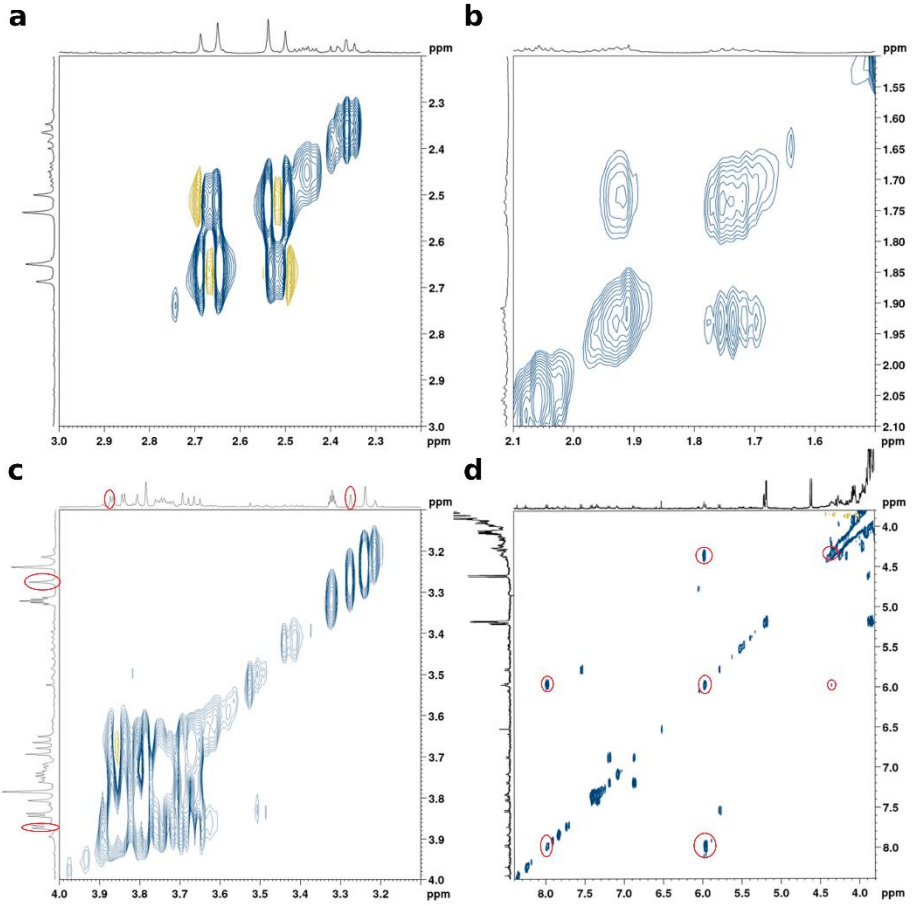


Figure 6. ^{1}H-^{1}H TOCSY spectra showing the spectral regions between 2.20 and 3.00 ppm of *T. magnatum* (**a**), 1.50 and 2.10 ppm of *T. magnatum* (**b**) 3.10 and 4.00 ppm of *T. melanosporum* (**c**) and 8.50 and 3.80 ppm of *T. magnatum* (**d**) assigned to the variables of citric acid, arginine, betaine and UDP-GlcNAc respectively.

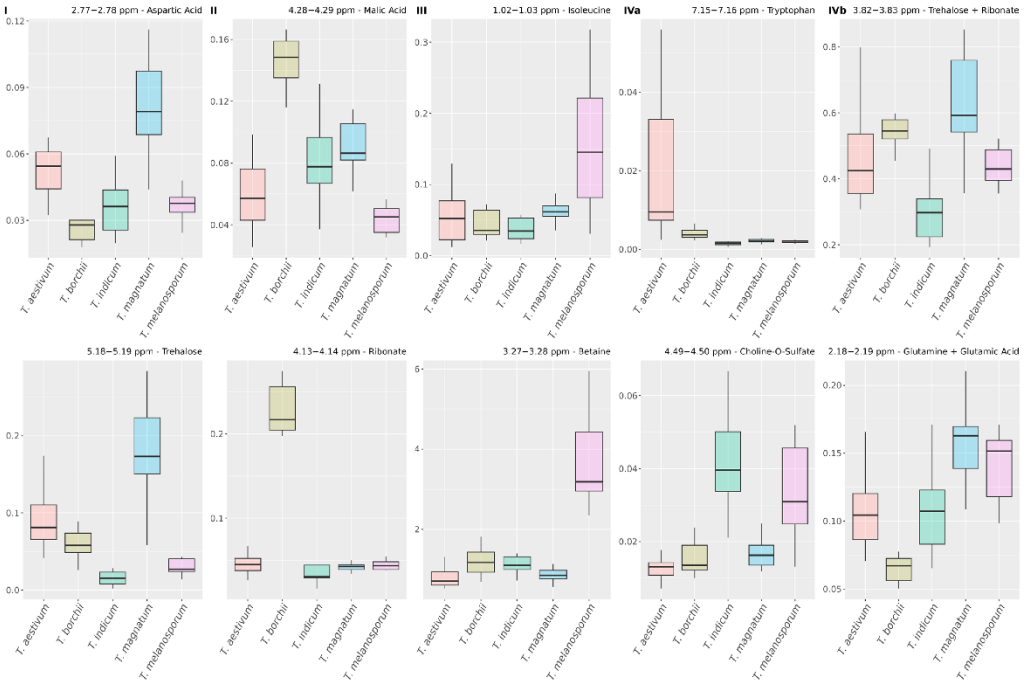


Figure 7. Boxplots of two representative variables for each cluster in Figure 5. The boxplots of the respective other variables of the clusters are shown in Figures S4–S8.

Cluster III is specific for the identification of *T. melanosporum* and contains five variables with comparatively high concentrations for this species (Figures 7III and S7). Two of these variables were assigned to betaine and the other three to isoleucine, leucine, and valine. The high values of the relationship parameter for the latter three can be explained by the fact that these metabolites are structurally and functionally very similar amino acids, called branched chain amino acids (BCAAs). Since they show specific classification information for *T. melanosporum*, differences in the synthesis and usage of BCAAs, which are well studied for fungi, can be assumed [57]. Betaine is known to be built in plants as a widespread response against environmental stress [58]. Hence, *T. melanosporum* could have a different stress tolerance or react differently to it than the other analyzed species.

Cluster IV contains variables with inhomogeneous classification information and we split them into two subclusters. Cluster IVa, the first subcluster (Figures 7IVa and S8), contains a variable at 7.15–7.16 ppm that has a very high concentration for *T. aestivum* and thus provides very specific classification information for this species. In *Cryptococcus neoformans*, tryptophan uptake and biosynthesis is essential for the survival of the organism at lower temperatures or when non-preferred nitrogen sources are available [59]. Higher tryptophan concentrations in *T. aestivum* could indicate that this species reacts differently to such external influences than the other species. The variables assigned to choline-O-sulfate show specific classification information to separate *T. indicum* and *T. melanosporum* from the other truffle species. Since it has been shown that fungi use this metabolite as a source of sulfur, this could demonstrate that the *Tuber* species have different sulfur metabolism [60].

Cluster IVb contains four variables (Figures 7IVb and S9). Two of them, which are assigned to glutamic acid and glutamine, are specific for the identification of *T. borchii* with very low levels for this class. They are therefore related to Cluster II, confirm-

ing the conclusion that this species could differ in energy metabolism. The variable at 3.82–3.83 ppm provides specific information for the classification of *T. indicum* and is assigned to ribonate and trehalose. The comparison of the classification of truffle species based on variables containing only ribonate or trehalose (see Figure 7I,II) shows that this bucket is indeed characterized by an overlap of the contributions of both metabolites. This is confirmed by the strong relations to the other variables of these metabolites, which were also discussed previously (see Section 3.4.2). However, since the increased concentration of *T. melanosporum* is not caused by one of the two metabolites, a third, unfortunately unidentified metabolite probably influences the variable at 3.82–3.83 ppm. The variable at 4.22–4.23 ppm associated with threonine shows unique classification information for *T. aestivum*. It is therefore strongly related to the other variable contributing this information at 7.15–7.16 ppm, which is assigned to tryptophan and was discussed in the previous paragraph. Threonine has been identified as a common residue from dephosphorylation reactions of proteins within *Saccharomyces cerevisiae* and other fungi, suggesting a different protein metabolism of *T. aestivum* [61].

In summary, the relationship analysis with SMD identified groups of variables with similar classification information that can be used to interpret class differences. Since these relationships are not apparent in the correlation analysis (see Figure S4), our analysis shows the benefit of including classification information in the relationship analysis of variables from NMR data.

4. Conclusions

In this study, using the classification of different truffle species, we demonstrate that the random forest black box for ^{1}H NMR metabolomics data can be opened by the application of SMD. We show this by the selection of important variables and the comprehensive analysis of variable relations based on their mutual impact on the random forest model. Groups of metabolites characteristic of specific species could be identified and linked to meaningful biological relationships. In addition, based on the SMD relation parameter, variables assigned to the same signals and metabolites could be identified and buckets with superimposed information could be unraveled. In summary, this analysis shows the potential of SMD for the comprehensive analysis of complex ^{1}H NMR metabolomics data to select and characterize the variables involved and support the identification and interpretation of the corresponding metabolites.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/metabo13101075/s1, Figure S1: Results of the principal component analysis; Figure S2: Relations of variables selected by SMD; Figure S3: Relations of variables selected by Boruta; Figure S4: Correlations of variables selected by SMD; Figures S5–S9: Boxplots of selected variables; Figures S10–S34: Results of spike-in experiments; Table S1: Data used for the RF analyses; Table S2: List of the variables selected by SMD and Boruta.

Author Contributions: Conceptualization, S.W., T.M., T.H. and S.S.; methodology, S.W. and S.S.; validation, S.W. and T.M.; formal analysis, S.W. and T.M.; investigation, S.W. and T.M.; resources, S.S., T.H. and M.F.; data curation, T.M. and S.W.; writing—original draft preparation, S.W.; writing—review and editing, S.W., T.M., T.H. and S.S.; visualization, S.W.; supervision, T.H. and S.S.; project administration, T.H. and S.S.; funding acquisition, S.S., T.H. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This study was performed within the project "Food Profiling—Development of Analytical Tools for the Experimental Verification of the Origin and Identity of Food". This project (funding reference number 2816500914) was supported by means of the Federal Ministry of Food and Agriculture (BMEL) by a decision of the German Bundestag (parliament). Project support was provided by the Federal Institute for Agriculture and Food (BLE) within the scope of the program for promoting innovation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are provided in the supplement.

Acknowledgments: We thank Frederic Saive for proofreading.

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

References

1. Wishart, D.S. Current Progress in Computational Metabolomics. *Brief. Bioinform.* **2007**, *8*, 279–293. [CrossRef]
2. Fiehn, O. Metabolomics—The Link between Genotypes and Phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171. [CrossRef]
3. Mushtaq, M.Y.; Choi, Y.H.; Verpoorte, R.; Wilson, E.G. Extraction for Metabolomics: Access to the Metabolome. *Phytochem. Anal.* **2014**, *25*, 291–306. [CrossRef]
4. Bachmann, R.; Klockmann, S.; Haerdter, J.; Fischer, M.; Hackl, T. ^{1}H-NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts. *J. Agric. Food Chem.* **2018**, *66*, 11873–11879. [CrossRef]
5. Shakiba, N.; Gerdes, A.; Holz, N.; Wenck, S.; Bachmann, R.; Schneider, T.; Seifert, S.; Fischer, M.; Hackl, T. Determination of the Geographical Origin of Hazelnuts (*Corylus avellana* L.) by Near-Infrared Spectroscopy (NIR) and a Low-Level Fusion with Nuclear Magnetic Resonance (NMR). *Microchem. J.* **2022**, *174*, 107066. [CrossRef]
6. Creydt, M.; Hudzik, D.; Rurik, M.; Kohlbacher, O.; Fischer, M. Food Authentication: Small-Molecule Profiling as a Tool for the Geographic Discrimination of German White Asparagus. *J. Agric. Food Chem.* **2018**, *66*, 13328–13339. [CrossRef]
7. Markley, J.L.; Brüschweiler, R.; Edison, A.S.; Eghbalnia, H.R.; Powers, R.; Raftery, D.; Wishart, D.S. The Future of NMR-Based Metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34–40. [CrossRef]
8. Bingol, K. Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods. *High-Throughput* **2018**, *7*, 9. [CrossRef]
9. Nagana Gowda, G.A.; Raftery, D. Can NMR Solve Some Significant Challenges in Metabolomics? *J. Magn. Reson.* **2015**, *260*, 144–160. [CrossRef]
10. Fan, T.W.-M.; Lane, A.N. Applications of NMR Spectroscopy to Systems Biochemistry. *Prog. Nucl. Magn. Reson. Spectrosc.* **2016**, *92–93*, 18–53. [CrossRef]
11. Takis, P.G.; Ghini, V.; Tenori, L.; Turano, P.; Luchinat, C. Uniqueness of the NMR Approach to Metabolomics. *TrAC Trends Anal. Chem.* **2019**, *120*, 115300. [CrossRef]
12. Hoch, J.C.; Baskaran, K.; Burr, H.; Chin, J.; Eghbalnia, H.R.; Fujiwara, T.; Gryk, M.R.; Iwata, T.; Kojima, C.; Kurisu, G.; et al. Biological Magnetic Resonance Data Bank. *Nucleic Acids Res.* **2023**, *51*, D368–D376. [CrossRef] [PubMed]
13. Garcia-Perez, I.; Posma, J.M.; Serrano-Contreras, J.I.; Boulangé, C.L.; Chan, Q.; Frost, G.; Stamler, J.; Elliott, P.; Lindon, J.C.; Holmes, E.; et al. Identifying Unknown Metabolites Using NMR-Based Metabolic Profiling Techniques. *Nat. Protoc.* **2020**, *15*, 2538–2567. [CrossRef]
14. Bingol, K.; Brüschweiler, R. NMR/MS Translator for the Enhanced Simultaneous Analysis of Metabolomics Mixtures by NMR Spectroscopy and Mass Spectrometry: Application to Human Urine. *J. Proteome Res.* **2015**, *14*, 2642–2648. [CrossRef] [PubMed]
15. Bingol, K.; Bruschweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Brüschweiler, R. Metabolomics Beyond Spectroscopic Databases: A Combined MS/NMR Strategy for the Rapid Identification of New Metabolites in Complex Mixtures. *Anal. Chem.* **2015**, *87*, 3864–3870. [CrossRef]
16. Dai, D.; He, J.; Sun, R.; Zhang, R.; Aisa, H.A.; Abliz, Z. Nuclear Magnetic Resonance and Liquid Chromatography–Mass Spectrometry Combined with an Incompleted Separation Strategy for Identifying the Natural Products in Crude Extract. *Anal. Chim. Acta* **2009**, *632*, 221–228. [CrossRef] [PubMed]
17. Watermann, S.; Bode, M.-C.; Hackl, T. Identification of Metabolites from Complex Mixtures by 3D Correlation of 1H NMR, MS and LC Data Using the SCORE-Metabolite-ID Approach. *Sci. Rep.* **2023**, *13*, 15834. [CrossRef]
18. Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R.H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J.C.; Holmes, E.; et al. Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic ^{1}H NMR Data Sets. *Anal. Chem.* **2005**, *77*, 1282–1289. [CrossRef] [PubMed]
19. Crockford, D.J.; Holmes, E.; Lindon, J.C.; Plumb, R.S.; Zirah, S.; Bruce, S.J.; Rainville, P.; Stumpf, C.L.; Nicholson, J.K. Statistical Heterospectroscopy, an Approach to the Integrated Analysis of NMR and UPLC-MS Data Sets: Application in Metabonomic Toxicology Studies. *Anal. Chem.* **2006**, *78*, 363–371. [CrossRef]
20. Ravanbakhsh, S.; Liu, P.; Bjordahl, T.C.; Mandal, R.; Grant, J.R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; et al. Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS ONE* **2015**, *10*, e0124219. [CrossRef]
21. Emwas, A.-H.; Saccenti, E.; Gao, X.; McKay, R.T.; Dos Santos, V.A.P.M.; Roy, R.; Wishart, D.S. Recommended Strategies for Spectral Processing and Post-Processing of 1D ^{1}H NMR Data of Biofluids with a Particular Focus on Urine. *Metabolomics* **2018**, *14*, 31. [CrossRef] [PubMed]
22. Debik, J.; Sangermani, M.; Wang, F.; Madssen, T.S.; Giskeødegård, G.F. Multivariate Analysis of NMR-based Metabolomic Data. *NMR Biomed.* **2022**, *35*, e4638. [CrossRef] [PubMed]
23. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
24. Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr. Metabolomics* **2012**, *1*, 92–107. [CrossRef]
25. Bro, R.; Smilde, A.K. Principal Component Analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]

26. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
27. Mendez, K.M.; Broadhurst, D.I.; Reinke, S.N. The Application of Artificial Neural Networks in Metabolomics: A Historical Perspective. *Metabolomics* **2019**, *15*, 142. [CrossRef]
28. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*, 1st ed.; Routledge: London, UK, 2017; ISBN 978-1-315-13947-0.
29. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
30. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
31. Seifert, S.; Gundlach, S.; Szymczak, S. Surrogate Minimal Depth as an Importance Measure for Variables in Random Forests. *Bioinformatics* **2019**, *35*, 3663–3671. [CrossRef]
32. Ishwaran, H.; Kogalur, U.B.; Chen, X.; Minn, A.J. Random Survival Forests for High-Dimensional Data: Random Survival Forests for High-Dimensional Data. *Stat. Anal. Data Min. ASA Data Sci. J.* **2011**, *4*, 115–132. [CrossRef]
33. Voges, L.F.; Jarren, L.C.; Seifert, S. Exploitation of Surrogate Variables in Random Forests for Unbiased Analysis of Mutual Impact and Importance of Features. *Bioinformatics* **2023**, *39*, btad471. [CrossRef]
34. Seifert, S. Application of Random Forest Based Approaches to Surface-Enhanced Raman Scattering Data. *Sci. Rep.* **2020**, *10*, 5436. [CrossRef] [PubMed]
35. Živanović, V.; Seifert, S.; Drescher, D.; Schrade, P.; Werner, S.; Guttmann, P.; Szekeres, G.P.; Bachmann, S.; Schneider, G.; Arenz, C.; et al. Optical Nanosensing of Lipid Accumulation Due to Enzyme Inhibition in Live Cells. *ACS Nano* **2019**, *13*, 9363–9375. [CrossRef] [PubMed]
36. Wenck, S.; Creydt, M.; Hansen, J.; Gärber, F.; Fischer, M.; Seifert, S. Opening the Random Forest Black Box of the Metabolome by the Application of Surrogate Minimal Depth. *Metabolites* **2022**, *12*, 5. [CrossRef] [PubMed]
37. Lösel, H.; Brockelt, J.; Gärber, F.; Teipel, J.; Kuballa, T.; Seifert, S.; Fischer, M. Comparative Analysis of LC-ESI-IM-qToF-MS and FT-NIR Spectroscopy Approaches for the Authentication of Organic and Conventional Eggs. *Metabolites* **2023**, *13*, 882. [CrossRef]
38. Mix, T.; Janneschütz, J.; Fischer, M.; Hackl, T. Differentiation of Truffle Species (*Tuber* spp.) by ^{1}H NMR Spectroscopy and support vector machine. *ChemRxiv* **2023**. *preprint*. [CrossRef]
39. Mannina, L.; Sobolev, A.P.; Capitani, D. Applications of NMR Metabolomics to the Study of Foodstuffs: Truffle, Kiwifruit, Lettuce, and Sea Bass: General. *Electrophoresis* **2012**, *33*, 2290–2313. [CrossRef]
40. Li, X.; Zhang, X.; Ye, L.; Kang, Z.; Jia, D.; Yang, L.; Zhang, B. LC-MS-Based Metabolomic Approach Revealed the Significantly Different Metabolic Profiles of Five Commercial Truffle Species. *Front. Microbiol.* **2019**, *10*, 2227. [CrossRef]
41. Shaka, A.J.; Lee, C.J.; Pines, A. Iterative Schemes for Bilinear Operators; Application to Spin Decoupling. *J. Magn. Reson. 1969* **1988**, *77*, 274–293. [CrossRef]
42. Dona, A.C.; Kyriakides, M.; Scott, F.; Shephard, E.A.; Varshavi, D.; Veselkov, K.; Everett, J.R. A Guide to the Identification of Metabolites in NMR-Based Metabonomics/Metabolomics Experiments. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 135–153. [CrossRef]
43. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C^{++} and *R. J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
44. Kucheryavskiy, S. Mdatools—R Package for Chemometrics. *Chemom. Intell. Lab. Syst.* **2020**, *198*, 103937. [CrossRef]
45. Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets. *Brief. Bioinform.* **2019**, *20*, 492–503. [CrossRef]
46. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer International Publishing: Cham, Switzerland, 2016; ISBN 978-3-319-24277-4.
47. Kolde, R. Pheatmap: Pretty Heatmaps. 2019. Available online: https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf (accessed on 11 October 2023).
48. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
49. Martin, F.; Canet, D.; Marchal, J.P. ^{13}C Nuclear Magnetic Resonance Study of Mannitol Cycle and Trehalose Synthesis during Glucose Utilization by the Ectomycorrhizal Ascomycete *Cenococcum graniforme*. *Plant Physiol.* **1985**, *77*, 499–502. [CrossRef] [PubMed]
50. Merzendorfer, H. The Cellular Basis of Chitin Synthesis in Fungi and Insects: Common Principles and Differences. *Eur. J. Cell Biol.* **2011**, *90*, 759–769. [CrossRef]
51. Genetet, I.; Martin, F.; Stewart, G.R. Nitrogen Assimilation in Mycorrhizas: Ammonium Assimilation in the N-Starved Ectomycorrhizal Fungus *Cenococcum Graniforme*. *Plant Physiol.* **1984**, *76*, 395–399. [CrossRef] [PubMed]
52. Lam, H.-M.; Coschigano, K.T.; Oliveira, I.C.; Melo-Oliveira, R.; Coruzzi, G.M. The Molecular-Genetics of Nitrogen Assimilation into Amino Acids in Higher Plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **1996**, *47*, 569–593. [CrossRef]
53. Kenealy, W.; Zaady, E.; Du Preez, J.C.; Stieglitz, B.; Goldberg, I. Biochemical Aspects of Fumaric Acid Accumulation by *Rhizopus arrhizus*. *Appl. Environ. Microbiol.* **1986**, *52*, 128–133. [CrossRef] [PubMed]
54. Roa Engel, C.A.; Straathof, A.J.J.; Zijlmans, T.W.; Van Gulik, W.M.; Van Der Wielen, L.A.M. Fumaric Acid Production by Fermentation. *Appl. Microbiol. Biotechnol.* **2008**, *78*, 379–389. [CrossRef]
55. Citterio, B.; Malatesta, M.; Battistelli, S.; Marcheggiani, F.; Baffone, W.; Saltarelli, R.; Stocchi, V.; Gazzanelli, G. Possible Involvement of *Pseudomonas fluorescens* and *Bacillaceae* in Structural Modifications of *Tuber borchii* Fruit Bodies. *Can. J. Microbiol.* **2001**, *47*, 264–268. [CrossRef] [PubMed]

56. Zhang, X.-X.; Rainey, P.B. Dual Involvement of CbrAB and NtrBC in the Regulation of Histidine Utilization in *Pseudomonas fluorescens* SBW25. *Genetics* **2008**, *178*, 185–195. [CrossRef]
57. Gross, S.R. Genetic Regulatory Mechanisms in the Fungi. *Annu. Rev. Genet.* **1969**, *3*, 395–424. [CrossRef]
58. Chen, T.H.H.; Murata, N. Enhancement of Tolerance of Abiotic Stress by Metabolic Engineering of Betaines and Other Compatible Solutes. *Curr. Opin. Plant Biol.* **2002**, *5*, 250–257. [CrossRef]
59. Fernandes, J.D.S.; Martho, K.; Tofik, V.; Vallim, M.A.; Pascon, R.C. The Role of Amino Acid Permeases and Tryptophan Biosynthesis in *Cryptococcus neoformans* Survival. *PLoS ONE* **2015**, *10*, e0132369. [CrossRef] [PubMed]
60. Spencer, B.; Hussey, E.C.; Orsi, B.A.; Scott, J.M. Mechanism of Choline *O*-Sulphate Utilization in Fungi. *Biochem. J.* **1968**, *106*, 461–469. [CrossRef]
61. Ariño, J.; Velázquez, D.; Casamayor, A. Ser/Thr Protein Phosphatases in Fungi: Structure, Regulation and Function. *Microb. Cell* **2019**, *6*, 217–256. [CrossRef] [PubMed]

5.3. Klassifizierung und Charakterisierung von Daten aus gekoppelter Flüssigchromatographie mit Massenspektrometrie von weißem Spargel (*Asparagus officinalis*)

In den beiden bisher beschriebenen Publikationen wurde SMD für die Untersuchung von ^{1}H NMR-*metabolomics*-Daten angewendet. Hier erfolgte nun eine Anwendung auf LC-MS Daten und es wurden 317 authentische Proben weißen Spargels aus Deutschland, Polen, Griechenland, Peru und den Niederlanden hinsichtlich ihrer Herkunft mit RF klassifiziert. Die Vorhersagegenauigkeit betrug 87.1 %. Die deutschen Proben, bei denen die entsprechende Information vorlag, wurden anschließend zusätzlich bezüglich deren Sorte (Backlim, Cumulus, Gijnlim, Grolim) klassifiziert, wobei die Vorhersagegenauigkeit 70 % betrug.

Beide Klassifizierungsfragestellungen wurden zusätzlich mit *probability machines*[112] untersucht, um Ähnlichkeiten der Klassen zu untersuchen. Dabei konnte gezeigt werden, dass sich die Proben aus den verschiedenen Ländern i.A. stärker unterscheiden als die Proben verschiedener Sorten, was die Klassifizierungsergebnisse bestätigt. Zusätzlich wurde eine Variablenselektion mit SMD durchgeführt und die Ergebnisse mit der Selektion mit Boruta verglichen. Für die Klassifizierung nach Herkunft und Sorte selektierte SMD 98 bzw. 60 Variablen, während Boruta 165 bzw. 48 selektierte. Auch hier zeigte sich, dass für die ausschließlich von Boruta selektierten Variablen vergleichsweise wenige andere Variablen mit ähnlichem Einfluss auf das Modell vorlagen.

Die Analyse der Variablenbeziehungen der selektierten Variablen ergab auch hier, dass Werte nahe 1 meisten auf Variablen desselben Metaboliten zurückzuführen sind, die durch Fragmentierung oder Bildung von Proton-, Natrium- oder Ammonium-Addukten entstanden. Für die Herkunftsanalyse konnte dies insbesondere für die im Spargel identifizierten Phytosterolester und für verschiedene Di- und Triglyceride beobachtet werden. Analog zu den NMR Daten konnten auch hier basierend auf der SMD Variablenbeziehungsanalyse Gruppen von Metaboliten mit ähnlichem Einfluss auf das

Modell gebildet werden. Diese konnten auf bekannte Stoffwechselwege im Lipidmetabolismus, der Sterol-Biosynthese oder auch verschiedene Coenzyme zurückgeführt werden.

metabolites

MDPI

Article

Opening the Random Forest Black Box of the Metabolome by the Application of Surrogate Minimal Depth

Soeren Wenck, Marina Creydt, Jule Hansen, Florian Gärber, Markus Fischer and Stephan Seifert *

Institute of Food Chemistry, Hamburg School of Food Science, University of Hamburg, Grindelallee 117, 20146 Hamburg, Germany; soeren.wenck@chemie.uni-hamburg.de (S.W.); marina.creydt@chemie.uni-hamburg.de (M.C.); jule.hansen@chemie.uni-hamburg.de (J.H.); florian.gaerber@studium.uni-hamburg.de (F.G.); markus.fischer@chemie.uni-hamburg.de (M.F.)
* Correspondence: stephan.seifert@chemie.uni-hamburg.de; Tel.: +49-40-42838-8818

Abstract: For the untargeted analysis of the metabolome of biological samples with liquid chromatography–mass spectrometry (LC-MS), high-dimensional data sets containing many different metabolites are obtained. Since the utilization of these complex data is challenging, different machine learning approaches have been developed. Those methods are usually applied as black box classification tools, and detailed information about class differences that result from the complex interplay of the metabolites are not obtained. Here, we demonstrate that this information is accessible by the application of random forest (RF) approaches and especially by surrogate minimal depth (SMD) that is applied to metabolomics data for the first time. We show this by the selection of important features and the evaluation of their mutual impact on the multi-level classification of white asparagus regarding provenance and biological identity. SMD enables the identification of multiple features from the same metabolites and reveals meaningful biological relations, proving its high potential for the comprehensive utilization of high-dimensional metabolomics data.

Keywords: classification; characterization; white asparagus; LC-MS; metabolomics; random forest; feature selection; feature relations; machine learning; chemometrics; surrogate minimal depth

Citation: Wenck, S.; Creydt, M.; Hansen, J.; Gärber, F.; Fischer, M.; Seifert, S. Opening the Random Forest Black Box of the Metabolome by the Application of Surrogate Minimal Depth. *Metabolites* **2022**, *12*, 5. https://doi.org/10.3390/metabo12010005

Academic Editors: J. Rafael Montenegro-Burke and Xavier Domingo-Almenara

Received: 17 November 2021
Accepted: 18 December 2021
Published: 21 December 2021

1. Introduction

Metabolomics is the in-depth analysis of metabolites, small molecules within biological systems (<1500 Da), that are products of cellular regulatory processes [1]. The composition of the metabolome is directly influenced by environmental factors such as fertilization, soil, climate, or proximity to large bodies of water, with genotype also having a dominant influence [2]. Accordingly, metabolites vary in their chemical structures and concentrations, corresponding to the environmental influences and include, e.g., lipids, sugars, or amino acids. Due to the diversity of compound classes, various technologies and methods have been developed for metabolome analysis, many of which are based on nuclear magnetic resonance (NMR) and mass spectrometry (MS) [3,4]. These techniques are complementary, as NMR technologies probe different substances than MS-based platforms [5]. However, untargeted NMR- and MS data sets are characterized by high numbers of features and small number of samples, which makes the data analysis challenging and prevents the application of classical statistical methods [6].

The popular unsupervised multivariate approach principal component analysis (PCA) reduces the dimensions of data by creating latent variables, which are linear combinations of the original variables. The generated principal components can be used to detect correlations and identify groups with a similar pattern. However, PCA illustrates the main variances in the data set that do not necessarily correspond to the differences that the researcher is interested in [7]. To obtain a model that focuses on these differences, supervised machine learning techniques such as support vector machines (SVM) [8], artificial neural networks (ANN) [9], partial least squares-discriminant analysis (PLS-DA) [10], and random

forest (RF) [11] are applied to metabolomics data. RF is particularly suitable because it considers feature interactions and is well suited for high-dimensional data [12].

RF consists of numerous binary decision trees [13]. Each of these decision trees uses a different bootstrap sample containing approximately 63% of the samples (some of them multiple times). For this reason, for each decision tree about 37% of the samples, called out-of-bag (OOB) samples, are left for evaluation, and RF provides independent error estimates (OOB errors) that do not require external data. RF is quite flexible in terms of input and output variables, so it is applied to different data sets, e.g., for classification and regression. To obtain an optimal classification model, the Gini index is applied at each node of each tree to identify the optimal partition among the candidate features. The candidates, whose quantity *mtry* is an important parameter in RF, are randomly selected from all features in each node. However, when RF is applied in a classification setting for prediction, only the assigned class is given and no information about the clarity of the class assignment is provided. In order to close this gap, probability machines have been developed [14]. As their name implies, they generate probabilities that can be used to evaluate the possibility of membership for each class. For this purpose, RF is applied in regression mode providing probabilities for each class that are averaged over all decision trees.

In addition to the application to build classification and regression models, RF can also be used to analyze the importance of individual features and rank them according to their impact on the outcome. This importance measure is either based on the decrease in accuracy of the model when the feature is permuted or the decrease of (Gini) impurity at the nodes that use the respective feature. Since the latter is biased, e.g., in favor of features with many possible split points, an unbiased adaptation called actual impurity reduction (AIR) has recently been introduced [15]. Based on the importance measures, various selection techniques have been developed that separate important from unimportant features. In a comprehensive comparison study, we recently identified Boruta as the best performing approach [16]. Boruta selects features with higher importance than the maximum value of so-called shadow variables, which are obtained by the permutation of the data across observations. For this comparison, a statistical test is applied, assigning significantly larger and smaller importance values. The generation of shadow variables and the comparison is repeated until all variables are labeled or a given number of runs (*maxRuns*) is reached [17].

An alternative approach for feature selection is Surrogate Minimal Depth (SMD), which incorporates relationships into the selection process and does not treat features individually, but as collaborating groups [18]. SMD exploits surrogate variables that have been developed to compensate for missing values in the data [19]. Furthermore, surrogate variables are also used to analyze the relationships between features based on a specific relation parameter that is called mean adjusted agreement. This relation parameter considers the mutual impact of the features to the outcome, and hence goes beyond the analysis of ordinary correlation coefficients. This is why SMD has been successfully applied in various fields and to data sets from different analytical techniques, for example to breast cancer gene expression data [18], FT-NIR food profiling data of hazelnuts [20], and to data from surface-enhanced Raman scattering [21], e.g., for the analysis of drugs in living cells [22].

In this study, we show that RF approaches and, in particular, SMD can be applied to LC-MS data sets for comprehensive multi-level classification and characterization. For this, we use data from asparagus authentication experiments [23–28].

2. Results and Discussion

2.1. Multi-Level Classification

An asparagus LC-MS data set was used for the classification with RF regarding the geographical origin and the botanical variety. In Table 1, the results for the determination of the geographical origin of 213 German, 25 Greek, 31 Dutch, 13 Peruvian, and 35 Polish samples are summarized. An OOB error rate of 12.9% was reached corresponding to a total accuracy of 87.1%. German as well as Greek samples and German as well as Peruvian

samples show values above 90% for sensitivity and specificity, respectively, while the Polish samples have values below 70% for both parameters. A detailed evaluation of the performance of the single samples shows that samples from all classes are misclassified as Polish and that the misclassification of German, Dutch, and Polish samples frequently happen among these classes. The reason for these misclassifications probably is the geographical proximity of the North-European samples that were evaluated. Since Germany borders the Netherlands and Poland, less pronounced differences between metabolomes are to be expected here than, for example, in the distinction between Schleswig-Holstein and Bavaria (distance approximately 850 km). Overall, the classification performance regarding the determination of the geographical origin is similar to previous work using LC-MS [26] and other analytical techniques [23–25].

Table 1. Confusion matrix for the classification of the geographical origin.

	Germany	**Greece**	**Netherlands**	**Peru**	**Poland**	**Sensitivity [%]**
Germany	196	4	5	0	8	92.0
Greece	0	23	0	0	2	92.0
Netherlands	5	0	23	0	3	74.2
Peru	1	0	0	11	1	84.6
Poland	7	2	3	0	23	65.7
Specificity [%]	93.8	79.3	74.2	100	62.2	

The results for botanical diversity classification of 56 Backlim, 23 Cumulus, 42 Gijnlim, and 29 Grolim samples are shown in Table 2. The OOB error and over-all accuracy are 30.0 and 70.0%, respectively, and the values of specificity and sensitivity for the single classes range between 60 and 75%. No misclassification patterns can be identified and the misclassification is evenly distributed among the classes. An exception is the Gijnlim samples, which are frequently misclassified as the variety Backlim. Overall, the classification results are worse than for the determination of the geographical origin. However, previous food profiling techniques that analyze the metabolome of asparagus did not focus on the determination of the variety [23–26], for which usually the genome is evaluated [29,30]. Hence, with this novel approach, we established a new level for the classification of asparagus LC-MS data showing classification accuracies that are substantially higher than the random distribution of 25%.

Table 2. Confusion matrix for the classification of the botanical variety.

	Backlim	**Cumulus**	**Gijnlim**	**Grolim**	**Sensitivity [%]**
Backlim	41	2	6	7	73.2
Cumulus	3	14	3	3	60.9
Gijnlim	9	1	31	1	73.8
Grolim	1	4	3	19	65.5
Specificity [%]	73.2	66.7	72.1	63.3	

The different varieties can be harvested at different times and provide asparagus with different characteristics. For example, the variety Gijnlim provides high yields and rather thin stems right at the beginning of the harvest period, while from the variety Backlim rather thick stems are obtained that can be harvested at the end of the harvest period. For this reason, farmers typically grow several varieties to ensure that there is sufficient product available throughout the harvest season. Samples for this study were all taken in the middle of the harvest season to ensure that, as far as possible, all varieties were present in sufficient numbers. Our results show that the LC-MS metabolome can be utilized to reflect genetic variances caused by different taxonomic varieties of asparagus. However, these are apparently not very large or probably do not affect the composition of the metabolome very much. However, the analysis shows that the metabolome contains

comprehensive information that can be used for the detailed RF classification of biological samples on multiple levels.

2.2. Probability Machines

A disadvantage of classification models regarding their application for unknown samples is that only the final class assignments are reported, and the user does not obtain any information about the clarity of this decision. This is why probability machines have been developed that can be conducted utilizing different machine learning techniques, including RF [14]. For samples that are analyzed by those machines, probabilities for each class are provided that enable a more detailed analysis of the assignment and, if necessary, also the manual labelling as unknown class when the probabilities are too similar. Tables S1 and S2 show the probabilities for each of the samples for the determination of the geographical origin and botanical variety. From these probabilities, boxplots for each class were generated in order to analyze the overall clarity of the decisions. (Figure 1) The assignment for German, Greek, Dutch, and Peruvian samples is quite clear with values mostly above 50% for the respective correct geographical origin (Figure 1a). However, for German and Dutch samples, the probabilities for the respective other North-European samples are slightly increased. This is in agreement with the misclassification patterns that were observed in the confusion matrix in the previous section (see Table 1). It is remarkable that the samples from Poland mainly show comparatively lower probabilities for the correct class assignment of ca. 25% to 50% and the other class assignment have similar probabilities between 10% and 20%. Therefore, it can be concluded that the misclassification of Polish samples is generally caused by a less accurate representation of this group in the classification model and not only by individual samples that come from a region close to the German border.

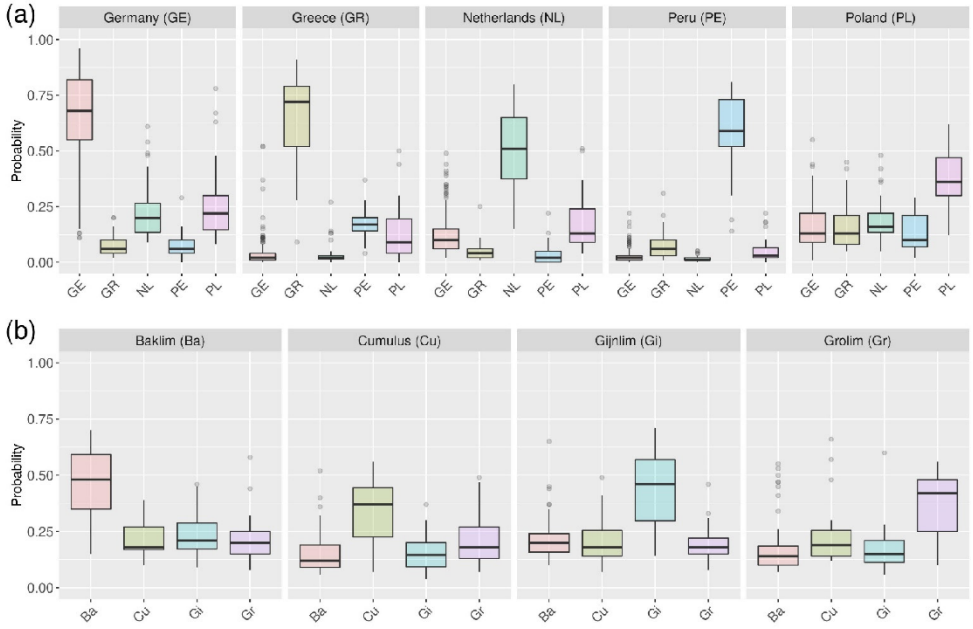


Figure 1. Results of the probability machines for the determination of the geographical origin (**a**) and the botanical variety (**b**). For the samples of each group (title), boxplots are shown that summarize the probabilities for the respective classes.

The probabilities of the correct class assignment for the biological varieties are generally less clear and mainly range between 25% and 60% (Figure 1b). However, they are still higher than the probabilities of the other respective classes that mainly show values between 10% and 30%. Interestingly, the probabilities of the Gijnlim class are slightly higher for the Backlim samples and the probabilities of the Grolim class for the Cumulus class and vice versa. Hence, the LC-MS metabolomes of Backlim and Gijnlim and Cumulus and Grolim are slightly more similar to each other, which was not apparent from the confusion matrix in the previous section (see Table 2). This demonstrates that the analysis of probabilities can give valuable additional insights about the similarities of the analyzed classes.

2.3. Feature Selection

In order to characterize the differences between the classes of asparagus samples, we applied the feature selection approaches SMD and Boruta. For the determination of the geographical origin 98 features with SMD and 165 features with Boruta and for the botanical variety 60 features with SMD and 48 features with Boruta were selected. Lists of the selected features are given in Tables S3 and S4. The overlap of the different approaches and the classification levels is visualized separately in Figure 2 and together in Figure S1. The Venn diagrams show that many features are selected by both approaches (Figure 2a,b). The differences are caused by the fact that the mutual importance of multiple metabolites is evaluated by SMD, while Boruta evaluates the metabolites individually. The small overlap between different classification levels that are depicted in Figure 2c,d show that mainly different metabolites are responsible for the classification regarding provenance and botanical variety. This means that mostly different subsets of the metabolites are utilized to build the different classification models.

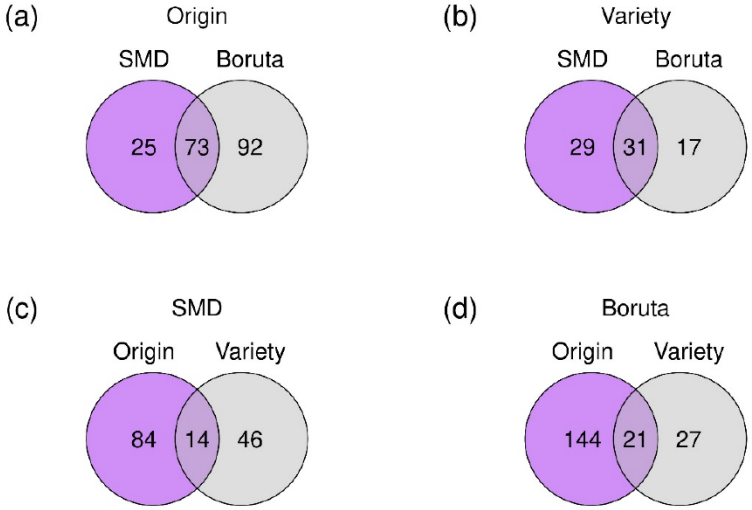


Figure 2. Venn diagrams of selected features for the classification on the different levels of geographical origin (**a**) and botanical variety (**b**) utilizing the approaches SMD (**c**) and Boruta (**d**). Detailed lists of the selected features can be found in Tables S3 and S4.

2.4. Analysis of the Relations of Selected Features

We applied SMD in order to analyze the relations between the selected features. As a result, for each pairwise feature combination, the relation parameter mean adjusted agreement is obtained, which represents the mutual impact on the outcome (see Tables

S5 and S6). For a comprehensive characterization of the class differences, we depicted the results in a heatmap that was obtained by the application of cluster analysis. We generated heatmaps for both, the selected features by Boruta and SMD. The comparison of the heatmaps of the different approaches confirms the assumption from the previous section that Boruta mainly selected important individual features, while SMD selected important groups (compare Figure 3 with Figures S2 and S3).

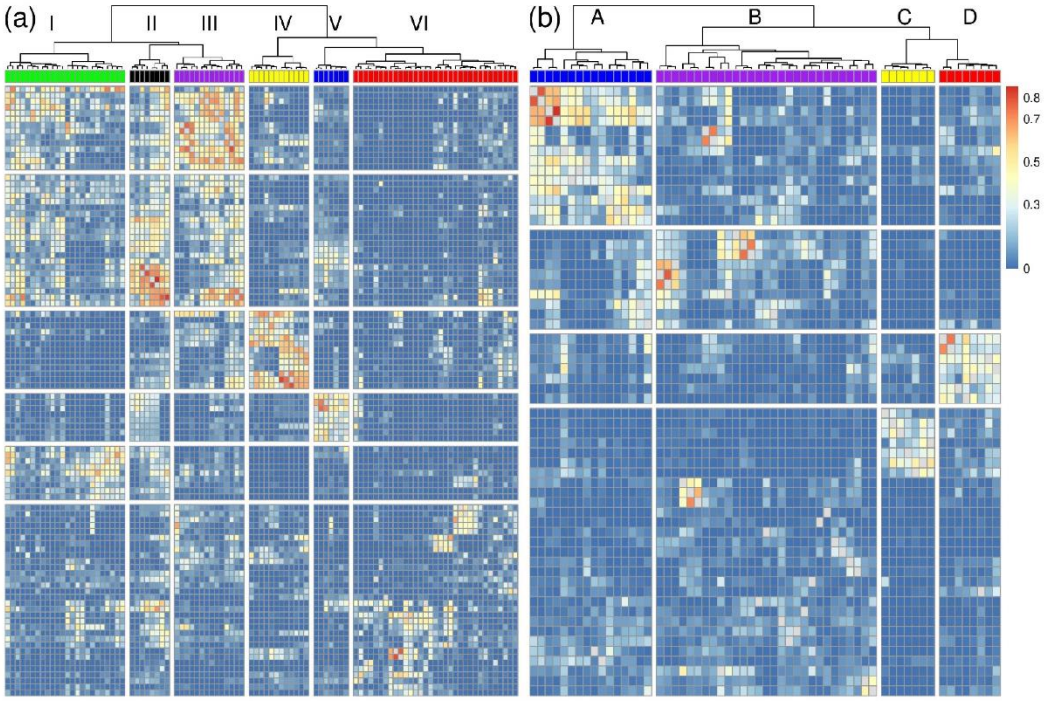


Figure 3. Results of the relation analysis of selected features with SMD for the determination of geographical origin (**a**) and botanical variety (**b**). For the hierarchical cluster analyses, Euclidean distances and Ward algorithm were applied and the clusters are labeled with I–VI and A–D. The clusters were assigned to the following molecular groups: I: Triaclyglycerols, II: Oxylipins, III: acylated monogalactosyldiacylglycerols, IV: Phytosterol esters, V: Triacylglycerols, VI: Various, A: Cycloartenol derivates, B: Coenzymes Q9 and Q10, C: Stigmasterol derivates, D: Diacylglycerols.

Figure 3a shows the results for the 98 features selected by SMD for the determination of the geographical origin. Six distinct clusters I-VI that contain metabolites with mainly similar information for the classification are obtained. In Figure 4, the intensities of two representative metabolites are shown for each cluster, while boxplots for all selected features are depicted in Figures S4–S9. The clusters I–III contain metabolites with similar classification patterns to separate North-European samples with higher intensities from the Peruvian and Greek samples that show lower intensities. Each of these clusters are mainly associated with one specific molecular class. Cluster I contains triacylglycerols with mono or double unsaturated fatty acids (Figure 4I). The degree of saturated and unsaturated fatty acids of plants is related to biotic and abiotic stress induced by external influences, which is why fatty acids have already been utilized to distinguish different geographical origins of food [26,31–33]. Cluster II consists of triacylglycerols with epoxy-fatty acids like triver-

nolin (Figure 4II). Cluster III contains different acylated monogalactosyldiacylglycerols (acMGDG) that also have previously been related to biotic and abiotic stress [32,34–36] and were identified as markers for the identification of German asparagus samples [26,37]. (Figure 4III) The high relations between the clusters I–III can be explained by the fact that all, triacylglycerols, triacylglycerols with epoxy-fatty acids, and MGDGs, are involved in the lipid metabolism of plants to build oxylipins as an environmental response [36]. Oxylipins are major actors in plant defense and have been shown to counteract bacterial and fungal infestation of plants [36,38]. Hence, it is very plausible that these metabolites in those clusters are useful interacting markers for the determination of the geographical origin of asparagus.

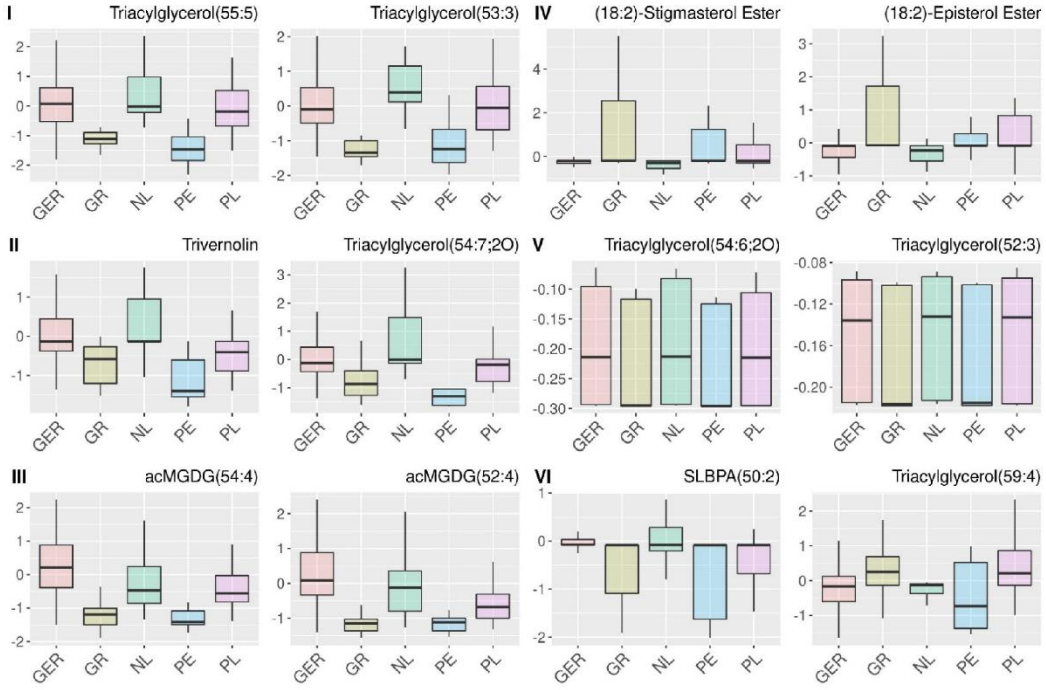


Figure 4. Boxplots of the autoscaled intensities for two exemplary metabolites of the clusters (**I–VI**) from the relation analysis Figure 3. A. Detailed information about the metabolites can be obtained from Table S7. Abbreviations: acMGDG: acylated monogalactosyldiacylglycerols; SLBPA: semi-lyso-bis-phosphatidic acid.

Arguably the most interesting cluster for interpretation is cluster IV, which can be applied to distinguish Dutch and German from the other samples that have higher intensities. This cluster contains several (18:2)-Phytosterol esters, e.g., (18:2)-Stigmasterol ester and (18:2)-Episterol ester, but also (18:2)-Campesterol ester and (18:2)-Sitosterol ester. (Figure 4IV). Also, the molecules of this class that interact in the sterol biosynthesis pathway have previously been identified as important for the determination of German asparagus samples because those compounds are affected by environmental changes [26,39,40].

Cluster V shows a similar but slightly different grouping of the classes as the clusters I to III and consists of triacylglycerols with multiple double bonds and their epoxides (Figure 4V). However, since those metabolites are arranged in a separate cluster, it can be concluded that they contribute to the classification model in essentially other ways.

Cluster VI is quite diverse, including various metabolites such as triacylglycerols and phospholipids like semi-lyso-bis-phosphatidic acids (SLBPAs), which contribute differently to the separation of the geographical origins. (Figure 4VI).

Similar as for the classification of the geographical origin, the selected features for the determination of the botanical variety were evaluated. Figure 3b shows the results of the relation analysis, while Figure 5 shows boxplots of two representative features for each cluster, and all features are depicted in Figures S10–S13. Also, here the clusters mainly contain metabolites that carry similar information for the classification. Cluster A is characterized by cycloartenol derivates that differentiate between low intensities of the varieties Backlim and Cumulus and high intensities of Gijnlim and Grolim (Figure 5A). Cycloartenol has been identified as the precursor for phytosterols influencing membrane fluidity, which is relevant in colder regions [41]. Cluster B consists of different metabolites that also contain various information for the variety classification. (Figure 5B) These metabolites include coenzymes Q9 and Q10, both of which are ubiquinones and act as carriers of electrons in mitochondrial membranes [42] and are involved in the abiotic stress response of plants [43]. Cluster C and D contain stigmasterol derivates and diacylglycerols and both show higher intensities for the varieties Backlim and Gijnlim. Hence, the metabolites in these clusters are responsible for the similarities of those botanical varieties that we observed by the analysis of class similarities by probability machines in Section 3.2.

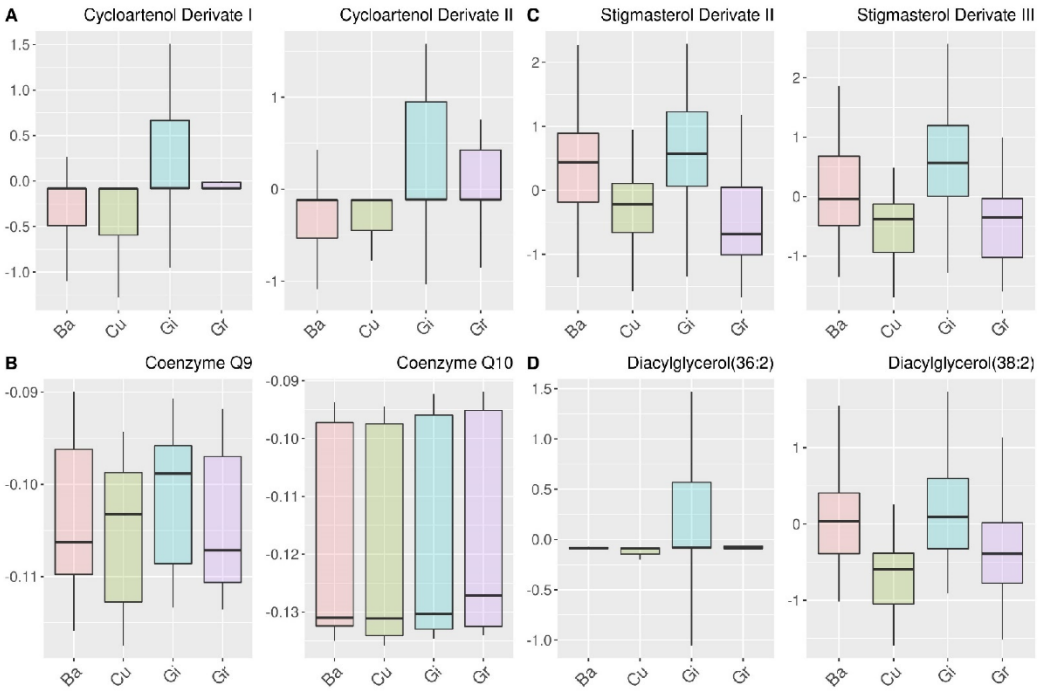


Figure 5. Boxplots of the autoscaled intensities for two exemplary features of the clusters (**A**–**D**) from the relation analysis for the determination of botanical variety in Figure 3B. Detailed information about the metabolites can be obtained from Table S7.

To summarize: The relation analysis that is accessible by the application of SMD enables the comprehensive characterization of the LC-MS metabolome of biological samples. This analysis goes beyond the analysis of pairwise correlation coefficients, which is demon-

strated by the comparison of the SMD relation analysis results (Figure 3) to heatmaps that were generated based on pairwise Pearson correlation coefficients (Figures S14 and S15). The correlation coefficients do not show the biological context of the metabolites that can be revealed by SMD, which is apparent by the clear grouping of molecule classes, e.g., of the phytosterol esters (Figure 3a, cluster IV) and the Coenzymes (Figure 3b, cluster B). In future applications, metabolites with mutual impact identified by SMD could be simultaneously utilized to directly test for specific characteristics, e.g., by the application of pathways-guided random forest approaches [44].

In all the clusters that we comprehensively analyzed in this section, small groups of highly related features could be identified. Those groups that consist of two to three elements contain very similar information for the classification, and we could assign those features to the same metabolites. In Figure 6, this is exemplarily shown for campesterol ester (a) and coenzyme Q9 (b). Since we always retained the $[M - NH_4]^+$ adduct in preprocessing, it is not surprising that the respective features can be identified in addition to the $[M + Na]^+$ adduct that is commonly detected. In addition, we also found typical molecular fragments formed from these molecules during the mass spectrometric measurement, which could be merged based on their mean adjusted agreement. Hence, SMD relation analysis in combination with an appropriate threshold could be applied in addition to correlation coefficients in LC-MS data processing workflows in order to merge features from the same metabolites.

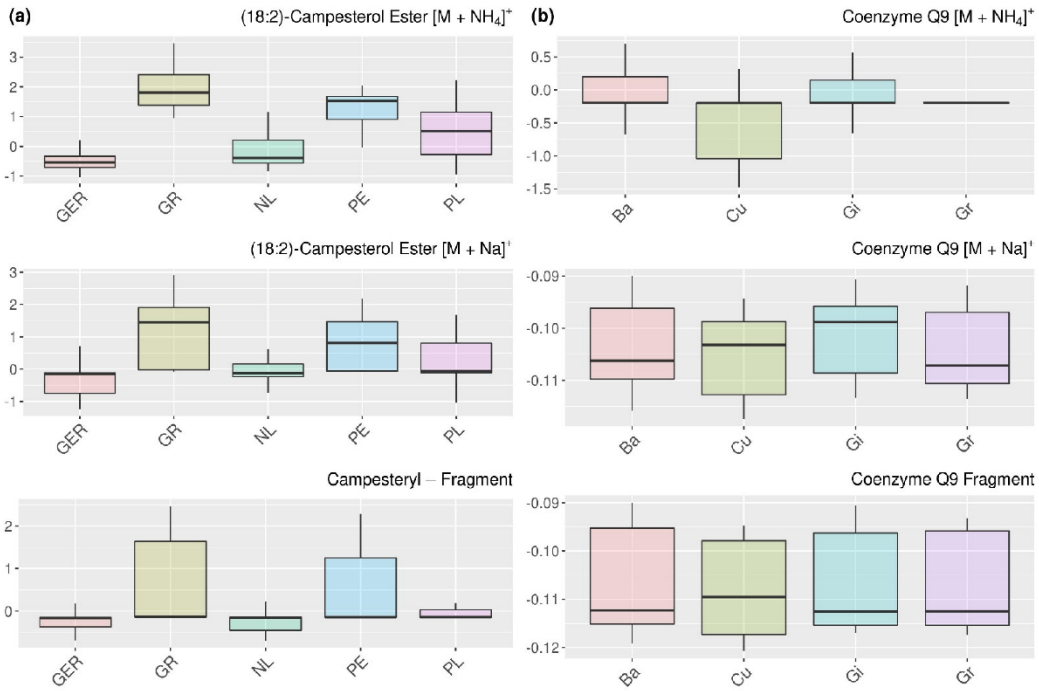


Figure 6. Comparison of features with very high pairwise mean adjusted agreement values that could be assigned to the same metabolites. Features of campesterol ester for the relation analysis regarding the determination of geographical origin (**a**) and features of coenzyme Q9 for the botanical variety (**b**) are shown. Detailed information about the metabolites can be obtained from Table S7.

3. Materials and Methods

3.1. Data Acquisition and Preprocessing

An LC-MS data set containing 317 samples obtained in the years 2016, 2017, and 2018 was used. For detailed information about the LC-MS measurement, please refer to [37]. The raw data set comprise the following positively charged adducts: $[M + Na]^+$; $[M + K]^+$; $[M - H_2O + H]^+$; $[M - CO_2 + H]^+$; $[M - NH_3 - H]^+$. In order to reduce the features from the same compounds, only the one with the highest intensity was used. Potential ammonium adducts were taken into account separately, since in a previous study, in some cases incorrect binning was observed [37]. Subsequently, the data set was further processed by excluding features that are present in less than 80% of the samples resulting in 718 features. In addition, missing data points were imputed with missForest [45] and, since the samples were measured in three batches, autoscaling was conducted separately for each batch [46]. In Figure S16, it can be observed that autoscaling significantly reduced the batch effects. For the identification of the important metabolites, MS/MS experiments were carried out for the features that were selected by SMD (see Table S7).

For the determination of the geographical origin, all 317 samples assigned to the five classes, Germany, Greece, Netherlands, Peru, and Poland were used (Table 3). In addition, some samples were used to differentiate between varieties. The focus was on the four most commonly grown varieties in Germany: Backlim, Cumulus, Gijnlim, and Grolim (Table 4).

Table 3. Overview of the analyzed samples regarding geographical origin.

Origin	2016	2017	2018
Germany	105	77	31
Greece	14	7	4
Netherlands	10	10	11
Peru	7	4	2
Poland	16	11	8

Table 4. Overview of the analyzed samples regarding botanical variety.

Variety	2016	2017
Backlim	33	23
Cumulus	12	11
Gijnlim	22	20
Grolim	18	11

3.2. Software and Analyses

The software R (version 3.6.3) and the R packages missForest (version 1.4, CRAN) [45] for missing value imputation, Pomona (Version 1.0.1, https://github.com/silkeszy/Pomona, accessed on 20 December 2021) [16] for Boruta feature selection, SurrogateMinimalDepth (version 0.2.0, https://github.com/StephanSeifert/Surrogate MinimalDepth, accessed on 20 December 2021) [18] for SMD feature selection and relation analysis, ranger (version 0.12.1, CRAN) [47] for RF analysis (classification and probability forest) and mdatools (version 0.12.0, CRAN) for PCA were used [48].

The RF approaches were applied in classification and probability mode with the parameters summarized in Table 5. In order to compensate for the class imbalance, the parameter *case.weights* was chosen accordingly. This means that the samples from rare classes were sampled more frequently for the bootstrap samples to train the RF. For visualization of the results of variable relation analysis, heatmaps of the mean adjusted agreement values of important features were depicted by the R package pheatmap using hierarchical cluster analysis with Euclidean distance and Ward algorithm [49]. For comparison, heatmaps of the Pearson correlation coefficients were generated accordingly.

Table 5. Parameters used for the RF-based approaches with p representing the total number of features.

Approach	Parameter	Description	Value
RF	ntree	number of trees	10,000
	min.node.size	number of samples in terminal node	1
	mtry	number of candidate features	138 ($p^{3/4}$) [1]
	case.weights	weights for sampling of training observations	chosen according to the size of the respective class
Boruta	importance	applied importance measure	impurity_corrected
	pValue	confidence level	0.01
	maxRuns	maximum number of importance source runs	100
SMD	*s*	predefined number of surrogate splits	35 ($p \cdot 0.05$)

[1] Motivated by [50].

4. Conclusions

In this study, we demonstrate the enormous potential that RF approaches, and SMD in particular, provide for the extensive exploitation of high-dimensional LC-MS data. We do this through their application to the data of biological samples, which goes far beyond black box classification. To be more precise, the classification of an asparagus data set regarding provenance and botanical varieties is complemented by the detailed evaluation of the class similarities obtained by the application of RF probability machines and the characterization by feature selection and relation analysis. For the relation analysis, we investigate the mutual impact of the features on the outcome that is accessible by SMD. This approach is very promising to get a comprehensive picture of the complex impact of metabolites on the outcome, as it reveals specific molecular groups and biomolecules known to interact in biological pathways.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/metabo12010005/s1, Figure S1: Venn diagram of all selected features, Figures S2 and S3: Results of the relation analysis of features selected by Boruta for the determination of geographical origin (S1) and botanical variety (S2), Figures S4–S9: Boxplots of the autoscaled intensities for the features of the cluster I-VI from the relation analysis for the determination of geographical origin in Figure 3a, Figures S10–S13: Boxplots of the autoscaled intensities for the features of the clusters A–D from the relation analysis for the determination of botanical variety in Figure 3b, Figures S14 and S15: Results of the correlation analysis of features selected by SMD for the determination of geographical origin (S14) and botanical variety (S15), Figure S16: Results of the PCA showing the sample distribution before and after autoscaling, Tables S1 and S2: Results of the probability machines for the determination of geographical origin (S1) and botanical variety (S2), Tables S3 and S4: Results of variable selection for the determination of geographical origin (S3) and botanical variety (S4), Tables S5 and S6: Results of variable relation analysis for the determination of geographical origin (S5) and botanical variety (S6), Table S7: Identified key metabolites for the separation of asparagus samples, Table S8: Feature table.

Author Contributions: Conceptualization, S.W. and S.S.; Methodology, S.W. and S.S.; Validation, F.G.; Formal analysis, S.W.; Investigation, S.W.; Resources, M.F. and S.S.; Data curation, M.C., S.S. and S.W.; Writing original draft preparation, S.W., S.S. and J.H.; Writing, review, and editing, S.S., M.F. and M.C.; Visualization, S.W.; Supervision, S.S.; Project administration, S.S.; Funding acquisition, S.S. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This study was performed within the Project "Asparagus Monitoring: Metabolomics-Based Methods for the Determination of the Geographical Origin of Asparagus (*Asparagus officinalis*) using NMR and LC−MS/MS together with Bioinformatics Analysis". This Industrial Collective Research (IGF) Project (18349 N) of FEI is supported via AiF within the program for promoting the IGF of the German Ministry of Economic Affairs and Energy (BMWi) based on a resolution of the German Parliament.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The unnormalized feature table is included in the Supplementary Materials.

Acknowledgments: We want to thank Silke Szymczak for helpful discussions and Lucas Voges for technical support.

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

References

1. Wishart, D.S. Current Progress in Computational Metabolomics. *Brief. Bioinform.* **2007**, *8*, 279–293. [CrossRef] [PubMed]
2. Fiehn, O. Metabolomics—The Link between Genotypes and Phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171. [CrossRef]
3. Dettmer, K.; Aronov, P.A.; Hammock, B.D. Mass Spectrometry-Based Metabolomics. *Mass Spectrom. Rev.* **2007**, *26*, 51–78. [CrossRef]
4. Bachmann, R.; Klockmann, S.; Haerdter, J.; Fischer, M.; Hackl, T. H-NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts. *J. Agric. Food Chem.* **2018**, *66*, 11873–11879. [CrossRef]
5. Ernst, M.; Silva, D.B.; Silva, R.R.; Vêncio, R.Z.N.; Lopes, N.P. Mass Spectrometry in Plant Metabolomics Strategies: From Analytical Platforms to Data Acquisition and Processing. *Nat. Prod. Rep.* **2014**, *31*, 784. [CrossRef]
6. Johnstone, I.M.; Titterington, D.M. Statistical Challenges of High-Dimensional Data. *Philos. Trans. Royal Soc.* **2009**, *367*, 4237–4253. [CrossRef]
7. Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr. Metabolomics* **2012**, *1*, 92–107. [CrossRef]
8. Klockmann, S.; Reiner, E.; Cain, N.; Fischer, M. Food Targeting: Geographical Origin Determination of Hazelnuts (*Corylus Avellana*) by LC-QqQ-MS/MS-Based Targeted Metabolomics Application. *J. Agric. Food Chem.* **2017**, *65*, 1456–1465. [CrossRef]
9. Long, N.P.; Lim, D.K.; Mo, C.; Kim, G.; Kwon, S.W. Development and Assessment of a Lysophospholipid-Based Deep Learning Model to Discriminate Geographical Origins of White Rice. *Sci. Rep.* **2017**, *7*, 8552. [CrossRef] [PubMed]
10. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis—a Marriage of Convenience or a Shotgun Wedding. *Anal. Chim. Acta* **2015**, *879*, 10–23. [CrossRef] [PubMed]
11. Erban, A.; Fehrle, I.; Martinez-Seidel, F.; Brigante, F.; Más, A.L.; Baroni, V.; Wunderlin, D.; Kopka, J. Discovery of Food Identity Markers by Metabolomics and Machine Learning Technology. *Sci. Rep.* **2019**, *9*, 9697. [CrossRef] [PubMed]
12. Qi, Y. Random Forest for Bioinformatics. In *Ensemble Machine Learning*; Zhang, C., Ma, Y., Eds.; Springer: Boston, MA, USA, 2012; pp. 307–323; ISBN 978-1-4419-9325-0.
13. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
14. Malley, J.D.; Kruppa, J.; Dasgupta, A.; Malley, K.G.; Ziegler, A. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. *Methods Inf. Med.* **2012**, *51*, 74–81. [CrossRef] [PubMed]
15. Nembrini, S.; König, I.R.; Wright, M.N. The Revival of the Gini Importance? *Bioinformatics* **2018**, *34*, 3711–3718. [CrossRef]
16. Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets. *Brief. Bioinform.* **2019**, *20*, 492–503. [CrossRef]
17. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
18. Seifert, S.; Gundlach, S.; Szymczak, S. Surrogate Minimal Depth as an Importance Measure for Variables in Random Forests. *Bioinformatics* **2019**, *35*, 3663–3671. [CrossRef]
19. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*, 1st ed.; Routledge: Abingdon, UK, 2017; ISBN 978-1-315-13947-0.
20. Shakiba, N.; Gerdes, A.; Holz, N.; Wenck, S.; Bachmann, R.; Schneider, T.; Seifert, S.; Fischer, M.; Hackl, T. Determination of the Geographical Origin of Hazelnuts (*Corylus Avellana* L.) by Near-Infrared Spectroscopy (NIR) and a Low-Level Fusion with Nuclear Magnetic Resonance (NMR). *Microchem. J.* **2022**, *174*, 107066. [CrossRef]
21. Seifert, S. Application of Random Forest Based Approaches to Surface-Enhanced Raman Scattering Data. *Sci. Rep.* **2020**, *10*, 5436. [CrossRef]
22. Živanović, V.; Seifert, S.; Drescher, D.; Schrade, P.; Werner, S.; Guttmann, P.; Szekeres, G.P.; Bachmann, S.; Schneider, G.; Arenz, C.; et al. Optical Nanosensing of Lipid Accumulation Due to Enzyme Inhibition in Live Cells. *ACS Nano* **2019**, *13*, 9363–9375. [CrossRef] [PubMed]

23. Richter, B.; Gurk, S.; Wagner, D.; Bockmayr, M.; Fischer, M. Food Authentication: Multi-Elemental Analysis of White Asparagus for Provenance Discrimination. *Food Chem.* **2019**, *286*, 475–482. [CrossRef] [PubMed]
24. Richter, B.; Rurik, M.; Gurk, S.; Kohlbacher, O.; Fischer, M. Food Monitoring: Screening of the Geographical Origin of White Asparagus Using FT-NIR and Machine Learning. *Food Control* **2019**, *104*, 318–325. [CrossRef]
25. Klare, J.; Rurik, M.; Rottmann, E.; Bollen, A.; Kohlbacher, O.; Fischer, M.; Hackl, T. Determination of the Geographical Origin of *Asparagus Officinalis* L. by ^{1} H-NMR Spectroscopy. *J. Agric. Food Chem.* **2020**, *68*, 14353–14363. [CrossRef] [PubMed]
26. Creydt, M.; Hudzik, D.; Rurik, M.; Kohlbacher, O.; Fischer, M. Food Authentication: Small-Molecule Profiling as a Tool for the Geographic Discrimination of German White Asparagus. *J. Agric. Food Chem.* **2018**, *66*, 13328–13339. [CrossRef]
27. Creydt, M.; Fischer, M. Metabolic Imaging: Analysis of Different Sections of White *Asparagus Officinalis* Shoots Using High-Resolution Mass Spectrometry. *J. Plant Physiol.* **2020**, *250*, 153179. [CrossRef] [PubMed]
28. Creydt, M.; Arndt, M.; Hudzik, D.; Fischer, M. Plant Metabolomics: Evaluation of Different Extraction Parameters for Nontargeted UPLC-ESI-QTOF-Mass Spectrometry at the Example of White *Asparagus Officinalis*. *J. Agric. Food Chem.* **2018**, *66*, 12876–12887. [CrossRef]
29. Zheng, Z.; Sun, Z.; Fang, Y.; Qi, F.; Liu, H.; Miao, L.; Du, P.; Shi, L.; Gao, W.; Han, S.; et al. Genetic Diversity, Population Structure, and Botanical Variety of 320 Global Peanut Accessions Revealed through Tunable Genotyping-by-Sequencing. *Sci. Rep.* **2018**, *8*, 14500. [CrossRef]
30. Scharf, A.; Lang, C.; Fischer, M. Genetic Authentication: Differentiation of Fine and Bulk Cocoa (*Theobroma Cacao L.*) by a New CRISPR/Cas9-Based in Vitro Method. *Food Control* **2020**, *114*, 107219. [CrossRef]
31. Torres-Moreno, M.; Torrescasana, E.; Salas-Salvadó, J.; Blanch, C. Nutritional Composition and Fatty Acids Profile in Cocoa Beans and Chocolates with Different Geographical Origin and Processing Conditions. *Food Chem.* **2015**, *166*, 125–132. [CrossRef]
32. Arena, E.; Campisi, S.; Fallico, B.; Maccarone, E. Distribution of Fatty Acids and Phytosterols as a Criterion to Discriminate Geographic Origin of Pistachio Seeds. *Food Chem.* **2007**, *104*, 403–408. [CrossRef]
33. Cossignani, L.; Blasi, F.; Simonetti, M.S.; Montesano, D. Fatty Acids and Phytosterols to Discriminate Geographic Origin of *Lycium Barbarum* Berry. *Food Anal. Methods* **2018**, *11*, 1180–1188. [CrossRef]
34. He, M.; Ding, N.-Z. Plant Unsaturated Fatty Acids: Multiple Roles in Stress Response. *Front. Plant Sci.* **2020**, *11*, 562785. [CrossRef]
35. Sauveplane, V.; Kandel, S.; Kastner, P.-E.; Ehlting, J.; Compagnon, V.; Werck-Reichhart, D.; Pinot, F. *Arabidopsis Thaliana* CYP77A4 Is the First Cytochrome P450 Able to Catalyze the Epoxidation of Free Fatty Acids in Plants: CYP77A4, an Epoxy Fatty Acid-Forming Enzyme. *FEBS J.* **2009**, *276*, 719–735. [CrossRef]
36. Cook, R.; Lupette, J.; Benning, C. The Role of Chloroplast Membrane Lipid Metabolism in Plant Environmental Responses. *Cells* **2021**, *10*, 706. [CrossRef] [PubMed]
37. Creydt, M.; Fischer, M. Mass-Spectrometry-Based Food Metabolomics in Routine Applications: A Basic Standardization Approach Using Housekeeping Metabolites for the Authentication of Asparagus. *J. Agric. Food Chem.* **2020**, *68*, 14343–14352. [CrossRef]
38. Rezzonico, E.; Moire, L.; Delessert, S.; Poirier, Y. Level of Accumulation of Epoxy Fatty Acid in Arabidopsis Thaliana Expressing a Linoleic Acid ?12-Epoxygenase Is Influenced by the Availability of the Substrate Linoleic Acid. *Theor. Appl. Genet.* **2004**, *109*, 1077–1082. [CrossRef]
39. Ferrer, A.; Altabella, T.; Arró, M.; Boronat, A. Emerging Roles for Conjugated Sterols in Plants. *Prog. Lipid Res.* **2017**, *67*, 27–37. [CrossRef] [PubMed]
40. Valitova, J.N.; Sulkarnayeva, A.G.; Minibayeva, F.V. Plant Sterols: Diversity, Biosynthesis, and Physiological Functions. *Biochemistry* **2016**, *81*, 819–834. [CrossRef]
41. Terletskaya, N.V.; Korbozova, N.K.; Kudrina, N.O.; Kobylina, T.N.; Kurmanbayeva, M.S.; Meduntseva, N.D.; Tolstikova, T.G. The Influence of Abiotic Stress Factors on the Morphophysiological and Phytochemical Aspects of the Acclimation of the Plant *Rhodiola Semenowii Boriss*. *Plants* **2021**, *10*, 1196. [CrossRef]
42. Swiezewska, E. Ubiquinone and Plastoquinone Metabolism in Plants. *Methods Enzymol.* **2004**, *378*, 124–131, ISBN 978-0-12-182782-3. [PubMed]
43. Liu, M.; Lu, S. Plastoquinone and Ubiquinone in Plants: Biosynthesis, Physiological Function and Metabolic Engineering. *Front. Plant Sci.* **2016**, *7*, 1898. [CrossRef] [PubMed]
44. Seifert, S.; Gundlach, S.; Junge, O.; Szymczak, S. Integrating Biological Knowledge and Gene Expression Data Using Pathway-Guided Random Forests: A Benchmarking Study. *Bioinformatics* **2020**, *36*, 4301–4308. [CrossRef]
45. Stekhoven, D.J.; Buhlmann, P. MissForest-Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
46. van den Berg, R.A.; Hoefsloot, H.C.; 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]
47. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and *R. J. Stat. Soft.* **2017**, *77*. [CrossRef]
48. Kucheryavskiy, S. Mdatools—R Package for Chemometrics. *Chemometrics Intell. Lab. Sys.* **2020**, *198*, 103937. [CrossRef]
49. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
50. Ishwaran, H.; Kogalur, U.B.; Chen, X.; Minn, A.J. Random Survival Forests for High-Dimensional Data: Random Survival Forests for High-Dimensional Data. *Stat. Anal. Data Min.* **2011**, *4*, 115–132. [CrossRef]

6. Diskussion

In diesem Kapitel sollen die Ergebnisse der vorliegenden Arbeit übergreifend diskutiert werden.

6.1 Einordnung der Arbeiten zu *metabolomics*-Analysen von Lebensmitteln

Metabolomics ist ein verhältnismäßig neues Forschungsfeld, das deshalb von besonderem Interesse ist, weil das Metabolom direkt von intrinsischen und extrinsischen Faktoren abhängt und somit auch durch Umgebungsbedingungen wie Temperatur, Wasser- und Nährstoffzufuhr sowie externen Stress beeinflusst wird.[53] Metabolomics-Untersuchungen, welche häufig auf den analytischen Verfahren der NMR-Spektroskopie und LC-MS basieren, werden für die Lebensmittelauthentifizierung eingesetzt, weil sie taxonomische oder geographische Unterschiede aufzeigen. Im Rahmen dieser Arbeit wurde dieser Ansatz für die Authentifizierung von Spargel, Trüffeln und Äpfeln erfolgreich angewendet. Die Unterscheidung verschiedener Arten von Trüffeln war dabei für alle untersuchten Proben möglich, während Unterscheidungen auf taxonomischer Ebene der Sorten, die für Apfel und Spargel durchgeführt wurde, geringere Genauigkeiten erzielte. Der Grund dafür liegt vermutlich darin, dass sich Proben auf dieser Ebene allgemein ähnlicher sind. Qi et al. haben bspw. gezeigt, dass *metabolomics*-Ansätze dafür geeignet sind, Pflanzenarten anhand spezifischer Metaboliten mit hoher Genauigkeit zu unterscheiden, aber auch Ähnlichkeiten zwischen Arten über ihre gemeinsamen Metabolite darzustellen.[123] Bei der Untersuchung von Sorten ist ein geringerer Unterschied zwischen den Metabolomen zu erwarten, weswegen mehr gemeinsame Metabolite in den Sorten vorhanden sind und die niedrigeren Klassifizierungsgenauigkeiten durch diesen Sachverhalt erklärt werden könnten. Auch Klassifizierungen basierend auf der geographischen Herkunft und der Produktionsweise sind allgemein mit geringeren Änderungen des Metaboloms verbunden, weshalb in dieser Arbeit für entsprechende Äpfel- und Spargelproben auf dieser Ebene keine perfekten Klassifizierungsergebnisse erreicht wurden. Die

Unterscheidung von biologisch und konventionell angebauten Äpfeln wurde bspw. von Xu et al. für Äpfel der Sorte *Fuji* durchgeführt. Sie erreichte eine akkuratere Vorhersage der Anbaubedingung mittels PCA und *orthogonal projection to latent structures discriminant analysis* von MS-Daten.[124] Ebenso wurden weitere Lebensmittel, wie bspw. Weißkohl[125] oder Kakao[126] hinsichtlich ihrer Anbaubedingung mit höheren Vorhersagegenauigkeiten klassifiziert. Es ist jedoch darauf hinzuweisen, dass die Klassen für die unter Abschnitt 5.1. präsentierte Studie selbst eine große innere Varianz aufweisen. So wurde Xus et al.'s Untersuchung von Äpfeln und auch die erwähnten Studien von Weißkohl und Kakao jeweils mit einer einzigen Sorte durchgeführt, während in den von uns präsentierten Untersuchungen Äpfel und Spargel von verschiedenen Sorten verwendet wurden. Die Annahme liegt also nahe, dass für eine genauere Klassifizierung der Anbaubedingungen und geographischer Herkunft Proben aus ein und derselben Sorte vorliegen sollten.

Äpfel werden aufgrund ihrer hohen Verfügbarkeit i.A. häufiger untersucht. So analysierten Bechynska et al. 274 authentische Apfelproben der Sorten *Idared*, *Golden Delicious* und *Gala* aus Polen und der Tschechischen Republik aus 2020 und 2022 mit LC-MS.[127] Ihre Klassifizierungsergebnisse ähneln den von uns präsentierten: Ihre Klassifizierungsmodelle mit *orthogonal partial least square discriminant analysis* erreichten Klassifizierungsgenauigkeiten innerhalb der einzelnen Sorten zwischen 65 und 88 %. Es ist aber zu betonen, dass sie ausreichend authentische Proben einer Sorte hatten, um diesen Ansatz zu verfolgen, was in unserer Studie nicht der Fall war.[127] Des Weiteren untersuchten Li et al. die NIR-Daten von Äpfeln der Sorten *Fuji, Red Star* und *Gala* hinsichtlich ihrer regionalen Herkunft innerhalb Chinas mit PCA, *backpropagation neural network*, SVM, *successive projections algorithms* und *extreme learning machines*. Sie erhielten eine breite Spanne an Vorhersagegenauigkeiten zwischen 43 und 97 %. Allerdings betrachteten auch Li et al. die Sorten jeweils immer nur einzeln, während uns Datensätze mit geringerer Homogenität vorlagen.

Trüffel werden ebenfalls häufiger untersucht, da durch die Verwendung von günstigeren Arten größere Gewinnmargen beim Lebensmittelbetrug erzielt werden können (Vergleich: Abschnitt 3.1.5.). Creydt et al. erreichten mit LC-MS sowie *collision cross section* MS/MS eine Vorhersagegenauigkeit von 100 %.[128] Šiškovič et al. erhielten mit einem *Headspace/solid space microextraction gas chromatography* Ansatz verschiedene Vorhersagegenauigkeiten für die Untersuchung von gefriergetrockneten Proben hinsichtlich verschiedener Trüffelarten und ihrer regionalen Herkunft innerhalb Sloweniens. Die von ihnen beschriebenen Ansätze erreichen Vorhersagewerte zwischen 60 und 85 % für *T. aestivum, T. brumale, T. magnatum* und *T. melanosporum*. Dem gegenüber erreichten Sibono et al. eine 95 %-ige Vorhersagegenauigkeit für die Unterscheidung zwischen spanischen und argentinischen *T. melanosporum* mit Gaschromatographie und gekoppelter MS und PCA-*linear discrimant analysis*.[129] Auch finden sich in der Literatur FT-NIR Ansätze: Segelke et al. erreichten mit einer *discriminant subspace analysis* eine 100 %-ige Unterscheidung innerhalb der weißen Trüffel und eine Vorhersagegenauigkeit von 99 % innerhalb der schwarzen Trüffel für *Tuber melanosporum* gegen *T. aestivum* und *T. indicum*. Dem gegenüber konnte *T. magnatum* von anderen schwarzen Trüffeln nur mit 83 % unterschieden werden. Es gibt desweiteren diverse Untersuchungen, die das Metabolom einzelner Trüffelarten untersuchen, wie Gao et al. für *T. indicum*[130] und Mannina et al. für *T. aestivum*. Diese Beispiele verfolgen allerdings keinerlei multivariate Klassifizierungsansätze, weswegen sie hier nur erwähnt werden.

Klassifizierungsfragen von Trüffel und Apfel werden in der Literatur häufig beschrieben, ihrer im Vergleich zu weißem Spargel größeren Marktrelevanz entsprechend. Da der weiße Spargel für den deutschen Markt besonders wichtig ist, stammen die meisten Studien tatsächlich aus Deutschland und zielen darauf, deutschen Spargel von anderen Herkünften zu unterscheiden. Da sich, wie in Abschnitt 3.1.7. beschrieben, Sorten im Prinzip lediglich bezüglich ihres Erntezeitraumes unterscheiden, wurde dieser Aspekt bisher kaum untersucht.

Klare et al. untersuchten weißen Spargel mit *isotope ratio* MS und NMR ausschließlich

hinsichtlich geographischen Herkunft mit SVM und erreichten Vorhersagegenauigkeiten von jeweils 70.9 % und 87.8 %.[131] In einem *isotopolomics* Ansatz beschrieben Richter et al. Messungen mit ICP-MS, womit sie eine Vorhersagegenauigkeit von 91.2 % mit SVM erreichten.[132] Creydt et al. untersuchten weißen Spargel in diversen MS-Ansätzen.[59,133,134] Da ihre Daten allerdings als Grundlage für die von uns untersuchten Variablenbeziehungen dienten, sollen sie hier nicht weiter vertieft werden.

Zusammenfassend ist zu sagen, dass wir mit unseren RF-Methoden detailliertere Analysen durchführten, die vielen bisherigen Methoden überlegen waren. Es ist aber immer zu beachten, dass nicht alle analytischen oder chemometrischen Methoden unmittelbar miteinander vergleichbar sind. So kann ein unterschiedlicher Ansatz in der Klassifizierung gleicher Daten zu unterschiedlichen Ergebnissen führen, und es ist überdies zu erwarten, dass bspw. bei der Unterscheidung der Apfel- und Spargelsorten mit größerer Anzahl an Proben jeder einzelnen Sorte bessere Vorhersagen für die Sorten- oder Herkunftsunterscheidung zu erzielen wären.

6.2 Verwendete multivariate Verfahren

Bei der Analyse aller dreier Lebensmittel zeigte sich, dass durch das unüberwachte Verfahren der PCA keine klare Unterscheidung bezüglich der untersuchten Fragestellungen, v.a. der taxonomischen Sorte oder der geographischen Herkunft, erreicht werden konnte. Somit sind die größten Unterschiede zwischen den jeweiligen Daten nicht allein auf diese Fragestellungen zurückzuführen, vermutlich überlagern sich verschiedene Parameter. Dies zeigte sich ebenfalls bei der Untersuchung von Shami et al., die genetisch veränderte Baumwollpflanzen anhand ihres Metaboloms kaum von unveränderter Baumwolle in einer PCA unterscheiden konnten.[135]

Wenn eine Überlagerung verschiedener Einflüsse vorliegt, sollten überwachte Methoden wie *artificial neural networks* (ANN), SVM, *partial least square discriminant analysis* (PLS-DA) und RF angewendet werden, denn diese trainieren ein Modell auf

spezifische, vorgegebene Klassenunterschiede. Um die Leistungsfähigkeit eines solchen Modells sinnvoll einzuschätzen, müssen allerdings Proben analysiert werden, welche nicht zum Training des Modells verwendet wurden. Deshalb teilt man üblicherweise die vorhandenen Daten in einen Trainings- und einen Testdatensatz, wobei auch der Trainingsdatensatz häufig noch zusätzlich, z.b. im Rahmen einer Kreuzvalidierung unterteilt wird, um die Parameter der überwachten Methode zu optimieren.[96]

Das am Häufigsten angewendete Verfahren zur Analyse von *metabolomics*-Daten ist die PLS-DA, wobei Gromski et al. die Frage aufwarfen, ob der Grund dafür wirklich in der Eignung des Verfahrens oder doch Bequemlichkeit liegt.[136] Sie wiesen darauf hin, dass hier auch andere Verfahren angewendet werden und analysiert werden sollte, welche Verfahren für die konkreten Datensätze geeignet sind. Bei RF wird deren Fähigkeit hervorgehoben, mit großen Datensätzen umzugehen, robust gegenüber *overfitting* zu sein und überdies eine *variable importance* berechnen zu können.[136]

Im Rahmen dieser Arbeit haben wir einen weiteren Vorteil von RF ausgenutzt, der bisher recht wenig hervorgehoben wurde. Wie in Abschnitt 3.3.3. ausgeführt, erhält man mit der auf *bagging* basierenden RF-Methode einen unabhängigen Vorhersagefehler der Trainingsdaten. Breiman führt aus , dass *bagging* bei Anwesenheit von starken Prädiktorvariablen der immanenten, hohen Korrelation der Bäume durch die zufällige Merkmalsauswahl in RFs entgegenwirkt.[106] Die damit entstehenden unkorrelierten Bäume ermöglichen so eine effektivere Varianzreduktion. In ökonomischen Untersuchungen zeigten Lee et al., dass *bagging* die Robustheit von RF erhöht.[137] Wenn zusätzlich darauf verzichtet wird, die Parameter des RF zu optimieren, was sowieso häufig zu keinen sehr großen Unterschieden in der Leistungsfähigkeit des Modells führt[110,111], wird ein OOB Fehler erhalten, der vergleichbar mit dem eines unabhängigen Testdatensatzes ist. Dies ist gerade dann von Vorteil, wenn ein Datensatz mit recht wenigen Proben, bzw. Klassen mit relativ wenigen Proben vorliegt, was im Rahmen von *metabolomics*-Untersuchungen häufig der Fall ist. Gerade bei diesen hochkomplexen Daten benötigt man jedoch hinreichende Probenzahlen, um extrinsische Einflüsse mit statistisch

robusten Datenpunkten abzudecken. Nyamundanda et al. leiteten bereits 2013 her, dass belastbare Aussagen von *metabolomics*-Analysen eine ausreichende Probenzahl erfordern.[138] Dieser Einfluss der Probenanzahl wurde auch durch Hansen et al. im Rahmen der Authentifizierung von Honig mittels LC-MS bestätigt, wo sie mit größerer Probenzahl höhere Klassifizierungsgenauigkeiten erreichten.

Bei der in dieser Arbeit durchgeführten NMR-basierten Klassifizierung von Äpfeln wurde festgestellt, dass kleinere Gruppen eine vergleichsweise geringe Vorhersagegenauigkeit aufwiesen, was darauf zurückgeführt wurde, dass die komplexe Varianz dieser Gruppen durch die geringe Anzahl an Proben im Modell nicht ausreichend repräsentiert war. Dies war hier der Fall, obwohl durch das vorteilhafte RF-Verfahren der Anteil der für das Training des Modells verwendeten Proben maximierte. Eine Nutzung eines anderen überwachten Verfahrens wäre für diese Fragestellung nicht sinnvoll möglich, weil zum einen dann der Trainingsdatensatz aus einer noch geringeren Zahl an Proben bestehen würde, vor allem aber, weil die ermittelte Genauigkeit dann nur auf wenigen Proben basierte. Der von uns gewählte Ansatz der RF Klassifizierung ohne Optimierung der Parameter ermöglicht somit die mit überwachten Verfahren durchgeführte Analyse von *metabolomics*-Daten, die sonst nicht sinnvoll hätten analysiert werden könnten. Interessanterweise war im Gegensatz zu Äpfeln die Klassifizierung der Trüffelart *T. borchii* erfolgreich, obwohl auch hier nur 7 Proben vorlagen. Hier zeigt sich, dass die Anzahl an Proben, die für eine aussagekräftige Analyse vorliegen müssen, sehr stark von der untersuchten Fragestellung, also von der Varianz der untersuchten Proben, aber auch davon abhängt, wie stark sich die verschiedenen Klassen unterscheiden.

6.3 Verwendete RF Parameter

Wie bereits ausgeführt, wurde im Rahmen dieser Arbeit keine Optimierung der RF Parameter *ntree*, *mtry* und *min.node.size* durchgeführt. Für den Parameter *ntree*, welcher die Anzahl trainierter Bäume festlegt, wurde jeweils ein Wert von 10000 verwendet. Dieser begründet sich darin, dass *ntree* ausreichend hoch sein sollte, um einen stabilen

OOB-Fehler zu erhalten, was bedeutet, dass jede Probe ausreichend häufig Teil der OOB-Proben ist. Aufgrund der asymptotischen Stabilisierung führt ein sehr viel höherer Wert von z.B. 100000 zu einer sehr viel höheren Rechenzeit, während das Ergebnis weitestgehend gleichbleibt.[106,139] Für den Parameter *min.node.size*, der bestimmt, wie tief die einzelnen Bäume trainiert werden, wurde der Wert von 1 verwendet, ein sehr häufiger Standardwert.[140] Für den Wert *mtry*, der die Anzahl der für den Split eines Knotenpunktes in Frage kommende Variablen bestimmt, wurde nicht der übliche Standardwert verwendet, der aus der Wurzel der Gesamtvariablen berechnet wird, sondern der Wert, der aus der Anzahl der Variablen hoch ¾ berechnet wird. Dieser Wert wurde deshalb gewählt, weil Ishwaran et al. für die Anwendung von MD diesen Wert in Simulationsstudien als besser geeignet identifizierten, und ein Schwerpunkt der Arbeit auf der Anwendung des darauf basierenden SMD Verfahrens liegt. Ein weiterer im Rahmen der RF Anwendungen festgelegter Parameter ist *case.weights*. Dieser legt die Wahrscheinlichkeit fest, mit der die entsprechenden Proben beim *bootstrapping* gezogen werden und ermöglicht so, ein *overfitting* des Modells zu verhindern. Dieser Parameter *case.weights* wurde somit entsprechend der Größe der jeweiligen Klassen festgelegt.

6.4 Selektion relevanter Variablen

Ein Vorteil von RF besteht darin, die Einschätzung der Wichtigkeit und Selektion relevanter Variablen zu ermöglichen. Diese bewerten jede Variable einzeln bezüglich ihrer Relevanz für das Modell. Im Rahmen dieser Arbeit wurde AIR in Kombination mit dem Selektionsverfahren Boruta angewendet. Zusätzlich dazu wurde SMD verwendet. SMD unterscheidet sich dadurch von etablierten Verfahren, dass Variablen nicht individuell, sondern im Zusammenhang mit anderen Variablen bewerten und selektiert werden. Als Erstes wurde dieses Verfahren auf Brustkrebs- sowie simulierte Genexpressions-Daten angewendet[118], anschließend aber auch auf komplexe Daten aus Zelluntersuchungen mit oberflächenverstärkter Raman-Spektroskopie zur Analyse der Wirkweise tricyclischer Antidepressiva, welche Lipidakkumulationen in den Lysosomen

hervorrufen.[141] Im Rahmen dieser Arbeit wurde SMD erstmalig auf *metabolomics*-Daten angewendet.

Der Vergleich der selektierten Variablen mit SMD und Boruta zeigte, dass jeweils ein großer Teil der Variablen von beiden Verfahren ausgewählt wurde, dass es aber auch Variablen gibt, die jeweils nur von einem Verfahren selektiert wurden. Die gemeinsam selektierten Variablen wurden meistens ebenfalls mit ANOVA selektiert, was dafürspricht, dass deren Relevanz für die jeweiligen Fragestellungen recht klar und der Einfluss vergleichsweise hoch ist.[59,63] Bei den Variablen, die nur von einem Verfahren selektiert wurden, ist dies nicht der Fall. Sie unterscheiden sich dadurch, dass die mit SMD zusätzlich selektierten Variablen mit anderen relevanten Variablen stark korrelieren, während die mit Boruta selektierten Variablen relativ individuelle Informationen enthalten.
Die Tatsache, dass SMD im Allgemeinen einzelne unkorrelierte relativ zu vielfach korrelierenden Variablen als unwichtiger einschätzt[118], ist eine bekannte Verzerrung dieser Methode, die auch hier der Grund dafür sein könnte, dass solche Variablen nur mit Boruta selektiert werden. SMD wiederrum selektiert im Vergleich zu Boruta eine größere Anzahl an Variablen mit ähnlichen Informationen, was sich auch bei der Analyse simulierter Raman Daten zeigte.[15] Wenn, wie im Falle dieser Arbeit, das Ziel der Selektion die Identifizierung aller relevanten Variablen zur Interpretation der Zusammenhänge ist (*all relevant*-Ansatz), lässt sich aus den Ergebnissen schlussfolgern, dass es sinnvoll ist, Boruta und SMD gemeinsam anzuwenden.

Die Auswahl und der Vergleich der Variablen, die für Authentifizierungsfragen wie z.B. die geographische Herkunft oder die taxonomische Sorte ausgewählt wurden, hat gezeigt, dass sowohl bei der Untersuchung von Äpfeln mit NMR als auch bei der Untersuchung von Spargel mit LC-MS größtenteils unterschiedliche Variablen relevant sind. Die Ergebnisse dieser Arbeit deuten also darauf hin, dass beide *metabolomics*-Verfahren unterschiedliche Aspekte der komplexen Daten für verschiedene

Authentifizierungsfragen nutzen und damit vielversprechend dafür sind, diese simultan zu untersuchen.

6.5 Analyse von Variablenbeziehungen mit SMD

Anders als die bereits etablierten Verfahren zur Selektion relevanter Variablen, wie z.B. Boruta, ermöglicht SMD zusätzlich eine Analyse von Variablenbeziehungen. Hierzu kann für jede paarweise Variablenbeziehung der Parameter MAA berechnet werden. Dieser wird einerseits durch die Korrelation der Variablen beeinflusst, aber auch, da er aus dem RF berechnet wird, durch die untersuchte Klassifizierungsfragestellung. Die MAA evaluiert also die Variablen bezüglich ihres gemeinsamen Einflusses auf das Klassifizierungsmodell, und es können Gruppen von Variablen mit ähnlichem Einfluss identifiziert werden. Im Rahmen dieser Arbeit wurde die SMD-Variablenbeziehungs-analyse auf NMR-Daten von Trüffeln und LC-MS-Daten von Spargel angewendet.

Das Ergebnis, eine zweidimensionale Ähnlichkeitsmatrix, wurde jeweils einer Clusteranalyse zugeführt und das Ergebnis in einer *heatmap* dargestellt. Um die Ergebnisse mit denen der Korrelationsanalyse zu vergleichen, wurden auch Korrelationskoeffizienten berechnet und auf ähnliche Weise, bzw. für NMR Daten, als zwei-dimensionale STOCSY Spektren dargestellt. Dieser Vergleich zeigte, dass einige der durch SMD identifizierten Variablenbeziehungen auch durch die Analyse der Korrelation ersichtlich sind, aber die Einbeziehung von Klasseninformationen in die Beziehungsanalyse und die Clusterbildung von Variablen mit homogenen Klassifizierungsmustern auch zusätzliche Erkenntnisse liefert, die für eine Interpretation der Zusammenhänge in RF Modellen und damit eine umfassende Analyse der Klassenunterschiede sehr hilfreich sind.

Allgemein zeigte sich sowohl für die Analyse von LC-MS- als auch von NMR-Datensätzen, dass Variablenbeziehungen auf verschiedenen Ebenen vorlagen und auch entsprechend interpretiert werden konnten. Bei NMR-Datensätzen wiesen sehr hohe MAA Werte bei

benachbarten Variablen darauf hin, dass hier einzelne Multiplet-Signale im Rahmen des angewendeten *bucketings* in mehreren Variablen aufgeteilt wurden. Die Analyse mit SMD kann hier gerade bei überlagernden Signalen dabei helfen, die entsprechenden Variablen sinnvoll zuzuordnen und zu interpretieren.

Bei der Analyse der Daten beider Analyseverfahren wurden kleine Gruppen von Variablen identifiziert und jeweils den gleichen Molekülen zugeordnet. Die LC-MS-Variablen konnten dabei auf verschiedene Addukte, bzw. bekannte Fragmente der Moleküle, zurückgeführt werden. Dass die NMR-Variablen tatsächlich Signale des jeweils gleichen Metaboliten sind, konnte durch 2-dimensionale NMR-Spektroskopie bestätigt werden. Allgemein zeigte sich hier, dass SMD ein hilfreiches Werkzeug zur Identifizierung von Molekülen mit überlagernden Informationen verschiedener Metabolite in komplexen Proben sein kann, und so wertvolle Beiträge bei der Analyse von Lebensmitteln liefert.

Der vermutlich jedoch wichtigste Aspekt der SMD Beziehungsanalyse ist die Analyse von Variablen verschiedener Metabolite und damit der Analyse intramolekularer Beziehungen. Hier wurden Gruppen von Variablen identifiziert, die ähnliche Informationen für die Klassifizierung liefern und deren Ähnlichkeiten sinnvoll interpretiert werden konnten. Beispielsweise wiesen bei der LC-MS-Analyse von Spargel zahlreiche Phytosterolester ähnliche Klassifizierungsmuster auf und erlaubten die Unterscheidung niederländischer und deutscher Proben von den Proben der anderen geographischen Herkünfte. Da diese Metabolite alle Bestandteile der Sterolsynthese sind, können die Ergebnisse dahingehend gedeutet werden, dass bei Proben der entsprechenden Länder prinzipielle Unterschiede dieses Stoffwechselwegs und nicht nur Unterschiede einzelner Metabolite vorliegen. Auch können die Unterschiede des Coenzym Q9 und Q10 Gehaltes dahingehend interpretiert werden, dass sich die Spargelsorten in dem Gehalt dieses Coenzyms unterscheiden. Des Weiteren wurden bei der NMR Klassifizierung von Trüffeln jeweils ähnliche Aminosäuren gruppiert, die auf generelle Unterschiede in der Synthese und Umsetzung dieser Metabolite hindeuten.

Auch konnten wir bspw. den Zusammenhang von Uridin-Diphosphat-*N*-Acetylglucosamin und Trehalose biochemisch sinnvoll auf Unterschiede in der Zellwandkomposition und des Zuckermetabolismus zurückführen.

Allgemein lässt sich zur Auswertung von *metabolomics*-Datensätzen mit SMD schließen, dass die *black box* Klassifizierung mit RF durch diese Analyse und die damit verbundene Integration von Variablenbeziehungen sinnvoll ergänzt werden kann, um eine umfangreiche Interpretation der Wirkungsweise der Variablen und Metabolite zu erreichen. Dies konnten wir für die Klassifizierung der geographischen Herkunft von Mandeln bestätigen, die u.a. durch verschiedene Fettsäuren beeinflusst wird.[142]

6.6 Grenzen und Ausblick der Anwendung von SMD

Ähnlich wie andere Verfahren zur Analyse der *variable importance* und der Selektion relevanter Variablen, ist SMD von bekannten Verzerrungen betroffen. Beispielsweise werden, wie bei der *Gini-Impurity*, Variablen mit vielen möglichen Splits systematisch bevorzugt. Ein ähnlicher Einfluss kann sowohl beim *importance* Maß SMD, als auch beim Beziehungsparameter MAA beobachtet werden.[143] Daher wurden die Parameter *mutual forest impact* (MFI) und *mutual impurity reduction* (MIR) entwickelt, welche auf dem gleichen Prinzip der Korrektur durch Permutation beruhen wie die AIR für die *Gini-Impurity*. Da im Rahmen dieser Arbeit *metabolomics*-Daten mit kontinuierlichen Variablen untersucht wurden, ist der Einfluss der Verzerrungen jedoch nicht sehr groß, sodass die MFI- und MIR-Analyse des NMR-Datensatzes von Trüffelproben sehr ähnliche Ergebnisse wie die entsprechende SMD-Untersuchung liefert (hier nicht gezeigt).

Da mittels einzelner analytischer Verfahren, wie auch der hier angewendeten Verfahren zur Metabolom-Analyse, jeweils nur ein geringer Teil der komplexen Zusammensetzung von Lebensmitteln erfasst wird, ist es vielversprechend, verschiedene Datensätze durch eine Data Fusion zu kombinieren.[62,144] In diesem Zusammenhang ist die Anwendung von SMD sinnvoll, um Variablenbeziehungen über mehrere Datensätze hinweg zu

analysieren, was wir im Rahmen der Diplomarbeit von Florian Gaerber durch die Kombination von NMR, LC-MS und ICP-MS Daten von Spargel zeigen konnten.[144] Darüber hinaus wurde durch die Anwendung von SMD auf fusionierte NIR- und LC-MS-Daten gezeigt, dass die Klassifikation zur Unterscheidung von biologisch und konventionell erzeugten Eiern auf Basis der NIR-Spektroskopie weniger aussagekräftig ist, da im Gegensatz zur LC-MS-Analyse keine Informationen über die enthaltenen Carotinoide vorliegen.[145] Durch die oben beschriebene Weiterentwicklung von SMD könnten in Zukunft nicht nur verschiedene analytische Verfahren zur Metabolom- und Elementanalyse, sondern auch Daten aus *genomics*-Experimenten verknüpft und die Beziehungen der Variablen analysiert werden. Dies könnte besonders für Fragestellungen der Lebensmittelauthentifizierung interessant sein, da so die Auswirkungen genetischer Grundlagen auf die Beschaffenheit des Metaboloms hervorträten.

7. Anhang

Für diese Arbeit wurden keine KMR-Chemikalien verwendet.

8. Danksagung

Ich bedanke mich bei Prof. Dr. Stephan Seifert, der mir die fantastische und fürsorgliche Betreuung gewährte und mich in Gesprächen auf die richtigen Ideen hat kommen lassen.

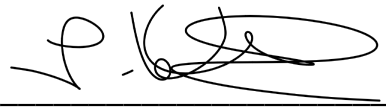
Ich danke meiner Familie, insb. meine Eltern, die mich immer unterstützt, an mich geglaubt und mir auch bei Fehlschlägen den Rücken weiterhin freigehalten haben.

Für meine Freunde, die mich durch gemeinsames Arbeiten über meine Grenzen hinaus haben wachsen lassen. Insbesondere bei Freddi, Nikola, Khang, Felix, Benny und Niklas.

Ich danke Dr. René Bachmann, Thorsten Mix und Dr. Marina Creydt für die hilfreichen Diskussionen und ihre Mitwirkung bei den Publikationen.

9. Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.



25.08.2025
Datum, Unterschrift

10. Literaturverzeichnis

(1) bionity.com. *Die Top 10 der deutschen Zukunftsbranchen*. https://www.bionity.com/de/news/1185818/die-top-10-der-deutschen-zukunftsbranchen.html?utm_source=newsletter&utm_medium=email&utm_campaign=bionityde--2025-03-17--2&WT.mc_id=ca0264 (accessed 2025-04-11).

(2) Niedersächsisches Ministerium für Wissenschaft und Kultur; Volkswagen Stiftung. *Big Data in den Lebenswissenschaften der Zukunft*. https//www.mwk.niedersachsen.de/download/132345/Ausschreibung_Big_Data_in_den_Lebenswissenschaften_der_Zukunft.pdf (accessed 2025-04-11).

(3) Li, R.; Li, L.; Xu, Y.; Yang, J. Machine Learning Meets Omics: Applications and Perspectives. *Briefings in Bioinformatics* **2022**, *23* (1), bbab460. https://doi.org/10.1093/bib/bbab460.

(4) Cebi, N.; Bekiroglu, H.; Erarslan, A. Nondestructive Metabolomic Fingerprinting: FTIR, NIR and Raman Spectroscopy in Food Screening. *Molecules* **2023**, *28* (23), 7933. https://doi.org/10.3390/molecules28237933.

(5) Dou, X.; Zhang, L.; Yang, R.; Wang, X.; Yu, L.; Yue, X.; Ma, F.; Mao, J.; Wang, X.; Zhang, W.; Li, P. Mass Spectrometry in Food Authentication and Origin Traceability. *Mass Spectrometry Reviews* **2023**, *42* (5), 1772–1807. https://doi.org/10.1002/mas.21779.

(6) Novotná, H.; Kmiecik, O.; Gałązka, M.; Krtková, V.; Hurajová, A.; Schulzová, V.; Hallmann, E.; Rembiałkowska, E.; Hajšlová, J. Metabolomic Fingerprinting Employing DART-TOFMS for Authentication of Tomatoes and Peppers from Organic and Conventional Farming. *Food Additives & Contaminants: Part A* **2012**, *29* (9), 1335–1346. https://doi.org/10.1080/19440049.2012.690348.

(7) Wei, F.; Furihata, K.; Koda, M.; Hu, F.; Kato, R.; Miyakawa, T.; Tanokura, M. ^{13}C NMR-Based Metabolomics for the Classification of Green Coffee Beans According to Variety and Origin. *J. Agric. Food Chem.* **2012**, *60* (40), 10118–10125. https://doi.org/10.1021/jf3033057.

(8) Johnstone, I. M.; Titterington, D. M. Statistical Challenges of High-Dimensional Data. *Philos. Trans. Royal Soc.* **2009**, *367* (1906), 4237–4253. https://doi.org/10.1098/rsta.2009.0159.

(9) Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn Comput* **2024**, *16* (1), 45–74. https://doi.org/10.1007/s12559-023-10179-8.

(10) Rajas, F.; Gautier-Stein, A.; Mithieux, G. Glucose-6 Phosphate, a Central Hub for Liver Carbohydrate Metabolism. *Metabolites* **2019**, *9* (12), 282. https://doi.org/10.3390/metabo9120282.

(11) Doan, M. T.; Teitell, M. A. Krebs and an Alternative TCA Cycle! *Cell Res* **2022**, *32* (6), 509–510. https://doi.org/10.1038/s41422-022-00664-4.

(12) Hall, J. E. *Guyton and Hall Textbook of Medical Physiology E-Book: Guyton and Hall Textbook of Medical Physiology E-Book*, 13th ed.; Guyton Physiology Series; Elsevier - Health Sciences Division: Chantilly, 2015.

(13) Europäische Kommission. *The EU Agri-Food Fraud Network, About the Food Fraud Network*. https://food.ec.europa.eu/food-safety/eu-agri-food-fraud-network_en (accessed 2025-04-11).

(14) Xu, L.; Xu, Z.; Liao, X. A Review of Fruit Juice Authenticity Assessments: Targeted and Untargeted Analyses. *Critical Reviews in Food Science and Nutrition* **2022**, *62* (22), 6081–6102. https://doi.org/10.1080/10408398.2021.1895713.

(15) Hansen, J.; Kunert, C.; Münstermann, H.; Raezke, K.-P. R.; Seifert, S. Application of Untargeted Liquid Chromatography-Mass Spectrometry to Routine Analysis of Food Using Three-Dimensional Bucketing and Machine Learning. *Sci Rep* **2024**, *14*, 16594. https://doi.org/10.1038/s41598-024-67459-y.

(16) Hansen, J.; Kunert, C.; Raezke, K.-P.; Seifert, S. Detection of Sugar Syrups in Honey Using Untargeted Liquid Chromatography–Mass Spectrometry and Chemometrics. *Metabolites* **2024**, *14* (11), 633. https://doi.org/10.3390/metabo14110633.

(17) Weesepoel, Y.; Alewijn, M.; Wijtten, M.; Müller-Maatsch, J. Detecting Food Fraud in Extra Virgin Olive Oil Using a Prototype Portable Hyphenated Photonics Sensor. *Journal of AOAC INTERNATIONAL* **2021**, *104* (1), 7–15. https://doi.org/10.1093/jaoacint/qsaa099.

(18) Kamiloglu, S. Authenticity and Traceability in Beverages. *Food Chemistry* **2019**, *277*, 12–24. https://doi.org/10.1016/j.foodchem.2018.10.091.

(19) Selamat, J.; Rozani, N. A. A.; Murugesu, S. Application of the Metabolomics Approach in Food Authentication. *Molecules* **2021**, *26* (24), 7565. https://doi.org/10.3390/molecules26247565.

(20) Wei, Y.; Liu, D. Review of Melamine Scandal: Still a Long Way Ahead. *Toxicol Ind Health* **2012**, *28* (7), 579–582. https://doi.org/10.1177/0748233711416950.

(21) Bonito, G. M.; Gryganskyi, A. P.; Trappe, J. M.; Vilgalys, R. A Global Meta-analysis of *Tuber* ITS rDNA Sequences: Species Diversity, Host Associations and Long-distance Dispersal. *Molecular Ecology* **2010**, *19* (22), 4994–5008. https://doi.org/10.1111/j.1365-294X.2010.04855.x.

(22) Stobbe, U.; Egli, S.; Tegel, W.; Peter, M.; Sproll, L.; Büntgen, U. Potential and Limitations of Burgundy Truffle Cultivation. *Appl Microbiol Biotechnol* **2013**, *97* (12), 5215–5224. https://doi.org/10.1007/s00253-013-4956-0.

(23) Molinier, V.; Murat, C.; Baltensweiler, A.; Büntgen, U.; Martin, F.; Meier, B.; Moser, B.; Sproll, L.; Stobbe, U.; Tegel, W.; Egli, S.; Peter, M. Fine-Scale Genetic Structure of

Natural Tuber Aestivum Sites in Southern Germany. *Mycorrhiza* **2016**, *26* (8), 895–907. https://doi.org/10.1007/s00572-016-0719-y.

(24) Pegler, D. N. Useful Fungi of the World: Morels and Truffles. *Mycologist* **2003**, *17* (4), 174–175. https://doi.org/10.1017/S0269915X04004021.

(25) Hall, I. R.; Yun, W.; Amicucci, A. Cultivation of Edible Ectomycorrhizal Mushrooms. *Trends in Biotechnology* **2003**, *21* (10), 433–438. https://doi.org/10.1016/S0167-7799(03)00204-X.

(26) Zambonelli, A.; Iotti, M.; Puliga, F.; Hall, I. R. Enhancing White Truffle (Tuber Magnatum Picco and T. Borchii Vittad.) Cultivation Through Biotechnology Innovation. In *Advances in Plant Breeding Strategies: Vegetable Crops*; Al-Khayri, J. M., Jain, S. M., Johnson, D. V., Eds.; Springer International Publishing: Cham, 2021; pp 505–532. https://doi.org/10.1007/978-3-030-66969-0_14.

(27) Oliach, D.; Vidale, E.; Brenko, A.; Marois, O.; Andrighetto, N.; Stara, K.; Martínez De Aragón, J.; Colinas, C.; Bonet, J. A. Truffle Market Evolution: An Application of the Delphi Method. *Forests* **2021**, *12* (9), 1174. https://doi.org/10.3390/f12091174.

(28) Reyna, S.; Garcia-Barreda, S. Black Truffle Cultivation: A Global Reality. *For. syst.* **2014**, *23* (2), 317–328. https://doi.org/10.5424/fs/2014232-04771.

(29) Laumont Shop. *Summer Truffle Aestivum*. https://laumontshop.eu/products/summer-truffle-aestivum (accessed 2025-03-11).

(30) Flammer, R.; Flammer, T.; Reil, P. *Trüffeln: Leitfaden zur Analyse der im Handel vorkommenden Arten*, 2. unveränderte Aufl.; IHW-Verlag: Eching, Kr Freising, 2018.

(31) tartufo.com. *Truffle price*. https://www.tartufo.com/en/truffle-prices/ (accessed 2025-04-11).

(32) Zhang, S.; Zhao, Y.; Xia, L.; Shi, Z.; Wang, D.; Wang, J.; Sun, L.; Zhao, M.; Li, J. Chemical Constituents from the Leaves of Malus Pumila Mill. And Their Chemotaxonomic Significance. *Biochemical Systematics and Ecology* **2022**, *105*, 104538. https://doi.org/10.1016/j.bse.2022.104538.

(33) Denver, S.; Jensen, J. D. Consumer Preferences for Organically and Locally Produced Apples. *Food Quality and Preference* **2014**, *31*, 129–134. https://doi.org/10.1016/j.foodqual.2013.08.014.

(34) Carnés, J.; Ferrer, A.; Fernández-Caldas, E. Allergenicity of 10 Different Apple Varieties. *Annals of Allergy, Asthma & Immunology* **2006**, *96* (4), 564–570. https://doi.org/10.1016/S1081-1206(10)63551-X.

(35) Kaeswurm, J. A. H.; Straub, L. V.; Siegele, A.; Brockmeyer, J.; Buchweitz, M. Characterization and Quantification of Mal d 1 Isoallergen Profiles and Contents in Traditional and Commercial Apple Varieties by Mass Spectrometry. *J. Agric. Food Chem.* **2023**, *71* (5), 2554–2565. https://doi.org/10.1021/acs.jafc.2c05801.

(36) Kaeswurm, J. A. H.; Neuwald, D. A.; Straub, L. V.; Buchweitz, M. Impact of Cultivation and Storage Conditions on Total Mal d 1 Content and Isoallergen Profile in Apples. *J. Agric. Food Chem.* **2023**, *71* (35), 12975–12985. https://doi.org/10.1021/acs.jafc.3c02375.

(37) Matthes, A.; Schmitz-Eiberger, M. Apple (Malus Domestica L. Borkh.) Allergen Mal d 1: Effect of Cultivar, Cultivation System, and Storage Conditions. *J. Agric. Food Chem.* **2009**, *57* (22), 10548–10553. https://doi.org/10.1021/jf901938q.

(38) Herzapfelhof. *Alte Apfelsorten Bio-Apfelsaft 5l BIB*. https://www.herzapfelhof.de/Alte-Apfelsorten-Bio-Apfelsaft-5l-BIB?gad_source=1 (accessed 2025-04-11).

(39) Anido, F. L.; Cointry, E. Asparagus. In *Vegetables II*; Prohens, J., Nuez, F., Eds.; Springer New York: New York, NY, 2008; pp 87–119. https://doi.org/10.1007/978-0-387-74110-9_3.

(40) Billau Jungpflanzen. *Backlim*. https://www.billau-jungpflanzen.de/jungpflanzen/backlim (accessed 2025-04-11).

(41) Billau Jungpflanzen. *Cumulus*. https://www.billau-jungpflanzen.de/jungpflanzen/cumulus (accessed 2025-04-11).

(42) Billau Jungpflanzen. *Gijnlim*. https://www.billau-jungpflanzen.de/jungpflanzen/gijnlim (accessed 2025-04-11).

(43) Billau Jungpflanzen. *Grolim*. https://www.billau-jungpflanzen.de/jungpflanzen/grolim (accessed 2025-04-11).

(44) Agrar Heute. *Teurer Genuss zu Ostern. So viel kostet 1 kg deutscher Spargel*. https://www.agrarheute.com/markt/marktfruechte/teurer-genuss-ostern-kostet-kilo-deutscher-spargel-618417.html (accessed 2025-03-25).

(45) Statistisches Bundesamt. *86% des importierten Spargels wurden 2023 während der Saison von März bis Juni eingeführt*. https://www.destatis.de/DE/Presse/Pressemitteilungen/2024/03/PD24_N013_51_41.html (accessed 2025-04-11).

(46) Del Giacco, L.; Cattaneo, C. Introduction to Genomics. In *Molecular Profiling*; Espina, V., Liotta, L. A., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2012; Vol. 823, pp 79–88. https://doi.org/10.1007/978-1-60327-216-2_6.

(47) Bragg, W. L. The Diffraction of Short Electromagnetic Waves by a Crystal. *Scientia* **1929**, *23* (45), 153.

(48) Signor, L.; Boeri Erba, E. Matrix-Assisted Laser Desorption/Ionization Time of Flight (MALDI-TOF) Mass Spectrometric Analysis of Intact Proteins Larger than 100 kDa. *J Vis Exp* **2013**, No. 79, 50635. https://doi.org/10.3791/50635.

(49) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan,

S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Protein Structure Prediction Using Multiple Deep Neural Networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **2019**, *87* (12), 1141–1148. https://doi.org/10.1002/prot.25834.

(50) Google DeepMind; EMBL-EBI. *AlphaFold Protein Structure Database*. https://alphafold.ebi.ac.uk/ (accessed 2025-11-04).

(51) Lenz, C.; Dihazi, H. Introduction to Proteomics Technologies. In *Statistical Analysis in Proteomics*; Jung, K., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2016; Vol. 1362, pp 3–27. https://doi.org/10.1007/978-1-4939-3106-4_1.

(52) Wishart, D. S. Current Progress in Computational Metabolomics. *Brief. Bioinformatics* **2007**, *8* (5), 279–293. https://doi.org/10.1093/bib/bbm030.

(53) Fiehn, O. Metabolomics - the Link between Genotypes and Phenotypes. *Plant Mol. Biol.* **2002**, *48* (1/2), 155–171. https://doi.org/10.1023/A:1013713905833.

(54) Manzoni, C.; Kia, D. A.; Vandrovcova, J.; Hardy, J.; Wood, N. W.; Lewis, P. A.; Ferrari, R. Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences. *Briefings in Bioinformatics* **2018**, *19* (2), 286–302. https://doi.org/10.1093/bib/bbw114.

(55) Aderemi, A. V.; Ayeleso, A. O.; Oyedapo, O. O.; Mukwevho, E. Metabolomics: A Scoping Review of Its Role as a Tool for Disease Biomarker Discovery in Selected Non-Communicable Diseases. *Metabolites* **2021**, *11* (7), 418. https://doi.org/10.3390/metabo11070418.

(56) Arneth, B.; Arneth, R.; Shams, M. Metabolomics of Type 1 and Type 2 Diabetes. *IJMS* **2019**, *20* (10), 2467. https://doi.org/10.3390/ijms20102467.

(57) Nelson, S. D.; Gordon, W. P. Mammalian Drug Metabolism. *J. Nat. Prod.* **1983**, *46* (1), 71–78. https://doi.org/10.1021/np50025a005.

(58) Zhong, P.; Wei, X.; Li, X.; Wei, X.; Wu, S.; Huang, W.; Koidis, A.; Xu, Z.; Lei, H. Untargeted Metabolomics by Liquid Chromatography-mass Spectrometry for Food Authentication: A Review. *Comp Rev Food Sci Food Safe* **2022**, *21* (3), 2455–2488. https://doi.org/10.1111/1541-4337.12938.

(59) Creydt, M.; Hudzik, D.; Rurik, M.; Kohlbacher, O.; Fischer, M. Food Authentication: Small-Molecule Profiling as a Tool for the Geographic Discrimination of German White Asparagus. *J. Agric. Food Chem.* **2018**, *66* (50), 13328–13339. https://doi.org/10.1021/acs.jafc.8b05791.

(60) Lösel, H.; Arndt, M.; Wenck, S.; Hansen, L.; Oberpottkamp, M.; Seifert, S.; Fischer, M. Exploring the Potential of High-Resolution LC-MS in Combination with Ion Mobility Separation and Surrogate Minimal Depth for Enhanced Almond Origin Authentication. *Talanta* **2024**, *271*, 125598. https://doi.org/10.1016/j.talanta.2023.125598.

(61) Sundekilde, U. K.; Eggers, N.; Bertram, H. C. NMR-Based Metabolomics of Food. In *NMR-Based Metabolomics*; Gowda, G. A. N., Raftery, D., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2019; Vol. 2037, pp 335–344. https://doi.org/10.1007/978-1-4939-9690-2_18.

(62) Shakiba, N.; Gerdes, A.; Holz, N.; Wenck, S.; Bachmann, R.; Schneider, T.; Seifert, S.; Fischer, M.; Hackl, T. Determination of the Geographical Origin of Hazelnuts (*Corylus Avellana L.*) by near-Infrared Spectroscopy (NIR) and a Low-Level Fusion with Nuclear Magnetic Resonance (NMR), 2021. https://doi.org/10.33774/chemrxiv-2021-1cr4g.

(63) Mix, T.; Janneschütz, J.; Ludwig, R.; Eichbaum, J.; Fischer, M.; Hackl, T. From Nontargeted to Targeted Analysis: Feature Selection in the Differentiation of Truffle Species (*Tuber* Spp.) Using ^{1}H NMR Spectroscopy and Support Vector Machine. *J. Agric. Food Chem.* **2023**, *71* (46), 18074–18084. https://doi.org/10.1021/acs.jafc.3c05786.

(64) Lösel, H.; Shakiba, N.; Bachmann, R.; Wenck, S.; Le Tan, P.; Creydt, M.; Seifert, S.; Hackl, T.; Fischer, M. Rapid Testing in the Food Industry: The Potential of Fourier Transform near-Infrared (FT-NIR) Spectroscopy and Spatially Offset Raman Spectroscopy (SORS) to Detect Raw Material Defects in Hazelnuts (Corylus Avellana L.). *Food Anal. Methods* **2024**, *17* (3), 486–497. https://doi.org/10.1007/s12161-024-02578-w.

(65) Günther, H. *NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry*, Third, completely revised and updated edition.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2013.

(66) Friebolin, H. *Ein- und zweidimensionale NMR-Spektroskopie: eine Einführung*, 4., vollst. überarb. und aktualisierte Aufl., 1. Nachdr.; Wiley-VCH: Weinheim, 2011.

(67) Ernst, R. R.; Bodenhausen, G.; Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; The International series of monographs on chemistry; Clarendon Press ; Oxford University Press: Oxford [Oxfordshire] : New York, 1987.

(68) Markley, J. L.; Brüschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S. The Future of NMR-Based Metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34–40. https://doi.org/10.1016/j.copbio.2016.08.001.

(69) Bingol, K. Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods. *High-Throughput* **2018**, *7* (2), 9. https://doi.org/10.3390/ht7020009.

(70) Nagana Gowda, G. A.; Raftery, D. Can NMR Solve Some Significant Challenges in Metabolomics? *J. Magn. Reson.* **2015**, *260*, 144–160. https://doi.org/10.1016/j.jmr.2015.07.014.

(71) Fan, T. W.-M.; Lane, A. N. Applications of NMR Spectroscopy to Systems Biochemistry. *Prog. Nucl. Magn. Reson. Spectrosc.* **2016**, *92–93*, 18–53. https://doi.org/10.1016/j.pnmrs.2016.01.005.

(72) Takis, P. G.; Ghini, V.; Tenori, L.; Turano, P.; Luchinat, C. Uniqueness of the NMR Approach to Metabolomics. *TRAC, Trend. Anal. Chem.* **2019**, *120*, 115300. https://doi.org/10.1016/j.trac.2018.10.036.

(73) Kohl, S. M.; Klein, M. S.; Hochrein, J.; Oefner, P. J.; Spang, R.; Gronwald, W. State-of-the Art Data Normalization Methods Improve NMR-Based Metabolomic Analysis. *Metabolomics* **2012**, *8* (Suppl 1), 146–160. https://doi.org/10.1007/s11306-011-0350-z.

(74) Jacob, D.; Deborde, C.; Moing, A. An Efficient Spectra Processing Method for Metabolite Identification from 1H-NMR Metabolomics Data. *Anal Bioanal Chem* **2013**, *405* (15), 5049–5061. https://doi.org/10.1007/s00216-013-6852-y.

(75) Ludwig, C.; Günther, U. L. MetaboLab - Advanced NMR Data Processing and Analysis for Metabolomics. *BMC Bioinformatics* **2011**, *12* (1), 366. https://doi.org/10.1186/1471-2105-12-366.

(76) Huang, K.; Thomas, N.; Gooley, P. R.; Armstrong, C. W. Systematic Review of NMR-Based Metabolomics Practices in Human Disease Research. *Metabolites* **2022**, *12* (10), 963. https://doi.org/10.3390/metabo12100963.

(77) Emwas, A.-H.; Roy, R.; McKay, R. T.; Tenori, L.; Saccenti, E.; Gowda, G. A. N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; Wishart, D. S. NMR Spectroscopy for Metabolomics Research. *Metabolites* **2019**, *9* (7), 123. https://doi.org/10.3390/metabo9070123.

(78) Garcia-Perez, I.; Posma, J. M.; Serrano-Contreras, J. I.; Boulangé, C. L.; Chan, Q.; Frost, G.; Stamler, J.; Elliott, P.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. Identifying Unknown Metabolites Using NMR-Based Metabolic Profiling Techniques. *Nat. Protoc.* **2020**, *15* (8), 2538–2567. https://doi.org/10.1038/s41596-020-0343-3.

(79) Aue, W. P.; Bartholdi, E.; Ernst, R. R. Two-Dimensional Spectroscopy. Application to Nuclear Magnetic Resonance. *The Journal of Chemical Physics* **1976**, *64* (5), 2229–2246. https://doi.org/10.1063/1.432450.

(80) Braunschweiler, L.; Ernst, R. R. Coherence Transfer by Isotropic Mixing: Application to Proton Correlation Spectroscopy. *Journal of Magnetic Resonance (1969)* **1983**, *53* (3), 521–528. https://doi.org/10.1016/0022-2364(83)90226-3.

(81) Reynolds, W. F.; Enríquez, R. G. Choosing the Best Pulse Sequences, Acquisition Parameters, Postacquisition Processing Strategies, and Probes for Natural Product Structure Elucidation by NMR Spectroscopy. *J. Nat. Prod.* **2002**, *65* (2), 221–244. https://doi.org/10.1021/np010444o.

(82) Paul; Magritek. *The 2D TOCSY Experiment*. https://magritek.com/2016/06/14/the-2d-tocsy-

experiment/#:~:text=Whereas%20COSY%20typically%20establishes%20correlations,spin%20systems%20from%20different%20molecules. (accessed 2025-04-11).

(83) Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic ^{1}H NMR Data Sets. *Anal. Chem.* **2005**, *77* (5), 1282–1289. https://doi.org/10.1021/ac048630x.

(84) Arpino, P. J.; Guiochon, G. LC/MS Coupling. *Anal. Chem.* **1979**, *51* (7), 682–701. https://doi.org/10.1021/ac50043a001.

(85) Östman, M. Antimicrobials in Sewage Treatment Plants. **2018**. https://doi.org/10.13140/RG.2.2.30958.87367.

(86) Scott, R. P. W. *Liquid Chromatography Detectors*, 2nd, completely rev. ed ed.; Journal of chromatography library; Elsevier ; Distributors for the U.S. and Canada, Elsevier Science Pub. Co: Amsterdam ; New York : New York, U.S.A, 1986.

(87) Olshina, M. A.; Sharon, M. Mass Spectrometry: A Technique of Many Faces. *Quart. Rev. Biophys.* **2016**, *49*, e18. https://doi.org/10.1017/S0033583516000160.

(88) Chetwynd, A. J.; David, A. A Review of Nanoscale LC-ESI for Metabolomics and Its Potential to Enhance the Metabolome Coverage. *Talanta* **2018**, *182*, 380–390. https://doi.org/10.1016/j.talanta.2018.01.084.

(89) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64–71. https://doi.org/10.1126/science.2675315.

(90) Xiao, J. F.; Zhou, B.; Ressom, H. W. Metabolite Identification and Quantitation in LC-MS/MS-Based Metabolomics. *Trends Analyt Chem* **2012**, *32*, 1–14. https://doi.org/10.1016/j.trac.2011.08.009.

(91) Henglin, M.; Claggett, B. L.; Antonelli, J.; Alotaibi, M.; Magalang, G. A.; Watrous, J. D.; Lagerborg, K. A.; Ovsak, G.; Musso, G.; Demler, O. V.; Vasan, R. S.; Larson, M. G.; Jain, M.; Cheng, S. Quantitative Comparison of Statistical Methods for Analyzing Human Metabolomics Data. *Metabolites* **2022**, *12* (6), 519. https://doi.org/10.3390/metabo12060519.

(92) Reel, P. S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnology Advances* **2021**, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.

(93) Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2021**, *2* (3), 160. https://doi.org/10.1007/s42979-021-00592-x.

(94) Wenck, S.; Creydt, M.; Hansen, J.; Gärber, F.; Fischer, M.; Seifert, S. Opening the Random Forest Black Box of the Metabolome by the Application of Surrogate Minimal Depth. *Metabolites* **2021**, *12* (1), 5. https://doi.org/10.3390/metabo12010005.

(95) Abamba Omwange, K.; Saito, Y.; Firmanda Al Riza, D.; Zichen, H.; Kuramoto, M.; Shiraga, K.; Ogawa, Y.; Kondo, N.; Suzuki, T. Japanese Dace (Tribolodon Hakonensis) Fish Freshness Estimation Using Front-Face Fluorescence Spectroscopy Coupled with Chemometric Analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2022**, *276*, 121209. https://doi.org/10.1016/j.saa.2022.121209.

(96) Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence* **1995**.

(97) Varma, S.; Simon, R. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics* **2006**, *7* (1), 91. https://doi.org/10.1186/1471-2105-7-91.

(98) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009. https://doi.org/10.1007/978-0-387-84858-7.

(99) Sarker, I. H.; Kayes, A. S. M.; Badsha, S.; Alqahtani, H.; Watters, P.; Ng, A. Cybersecurity Data Science: An Overview from Machine Learning Perspective. *J Big Data* **2020**, *7* (1), 41. https://doi.org/10.1186/s40537-020-00318-5.

(100) Francois-Lavet, V.; Henderson, P.; Islam, R.; Bellemare, M. G.; Pineau, J. An Introduction to Deep Reinforcement Learning. **2018**. https://doi.org/10.48550/ARXIV.1811.12560.

(101) Majumder, A. Introduction to Reinforcement Learning. In *Deep Reinforcement Learning in Unity*; Apress: Berkeley, CA, 2021; pp 1–71. https://doi.org/10.1007/978-1-4842-6503-1_1.

(102) Mohammed, M.; Khan, M. B.; Bashier, E. B. M. *Machine Learning*, 0 ed.; CRC Press, 2016. https://doi.org/10.1201/9781315371658.

(103) Limonte, C. P.; Kretzler, M.; Pennathur, S.; Pop-Busui, R.; De Boer, I. H. Present and Future Directions in Diabetic Kidney Disease. *Journal of Diabetes and its Complications* **2022**, *36* (12), 108357. https://doi.org/10.1016/j.jdiacomp.2022.108357.

(104) Jovel, J.; Greiner, R. An Introduction to Machine Learning Approaches for Biomedical Research. *Front. Med.* **2021**, *8*, 771607. https://doi.org/10.3389/fmed.2021.771607.

(105) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2* (1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

(106) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(107) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*, 1st ed.; Routledge, 2017. https://doi.org/10.1201/9781315139470.

(108) Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F. A. A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. *BMC Bioinformatics* **2009**, *10* (1), 213. https://doi.org/10.1186/1471-2105-10-213.

(109) Breiman, L. Bagging Predictors. *Mach Learn* **1996**, *24* (2), 123–140. https://doi.org/10.1007/BF00058655.

(110) Brettschneider, K. C.; Seifert, S. Fusion of Food Profiling Data from Very Different Analytical Techniques. *Current Opinion in Food Science* **2025**, *61*, 101256. https://doi.org/10.1016/j.cofs.2024.101256.

(111) Probst, P.; Wright, M. N.; Boulesteix, A. Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Min & Knowl* **2019**, *9* (3), e1301. https://doi.org/10.1002/widm.1301.

(112) Malley, J. D.; Kruppa, J.; Dasgupta, A.; Malley, K. G.; Ziegler, A. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. *Methods Inf. Med.* **2012**, *51* (01), 74–81. https://doi.org/10.3414/ME00-01-0052.

(113) Nembrini, S.; König, I. R.; Wright, M. N. The Revival of the Gini Importance? *Bioinformatics* **2018**, *34* (21), 3711–3718. https://doi.org/10.1093/bioinformatics/bty373.

(114) Trulson, I.; Klawonn, F.; Holdenrieder, S.; Hoffmann, G. Praktische Herausforderungen beim maschinellen Lernen: Auf die Datenaufbereitung kommt es an. *TD* **2024**, *22* (1), 55–57. https://doi.org/10.47184/td.2024.01.07.

(115) Janitza, S.; Celik, E.; Boulesteix, A.-L. A Computationally Fast Variable Importance Test for Random Forests for High-Dimensional Data. *Adv Data Anal Classif* **2018**, *12* (4), 885–915. https://doi.org/10.1007/s11634-016-0276-4.

(116) Kursa, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36* (11). https://doi.org/10.18637/jss.v036.i11.

(117) Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets. *Brief. Bioinform.* **2019**, *20* (2), 492–503. https://doi.org/10.1093/bib/bbx124.

(118) Seifert, S.; Gundlach, S.; Szymczak, S. Surrogate Minimal Depth as an Importance Measure for Variables in Random Forests. *Bioinform.* **2019**, *35* (19), 3663–3671. https://doi.org/10.1093/bioinformatics/btz149.

(119) Rudnicki, W. R.; Kierczak, M.; Koronacki, J.; Komorowski, J. A Statistical Method for Determining Importance of Variables in an Information System. In *Rough Sets and*

Current Trends in Computing; Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H. S., Słowiński, R., Eds.; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Series Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; Vol. 4259, pp 557–566. https://doi.org/10.1007/11908029_58.

(120) Ishwaran, H.; Kogalur, U. B.; Chen, X.; Minn, A. J. Random Survival Forests for High-Dimensional Data: Random Survival Forests for High-Dimensional Data. *Stat. Anal. Data Min.* **2011**, *4* (1), 115–132. https://doi.org/10.1002/sam.10103.

(121) Wenck, S.; Mix, T.; Fischer, M.; Hackl, T.; Seifert, S. Opening the Random Forest Black Box of 1H NMR Metabolomics Data by the Exploitation of Surrogate Variables. *Metabolites* **2023**, *13* (10), 1075. https://doi.org/10.3390/metabo13101075.

(122) The R Foundation. *What is R?* https://www.r-project.org/about.html (accessed 2025-04-11).

(123) Qi, J.; Li, K.; Shi, Y.; Li, Y.; Dong, L.; Liu, L.; Li, M.; Ren, H.; Liu, X.; Fang, C.; Luo, J. Cross-Species Comparison of Metabolomics to Decipher the Metabolic Diversity in Ten Fruits. *Metabolites* **2021**, *11* (3), 164. https://doi.org/10.3390/metabo11030164.

(124) Xu, L.; Wang, L.; Xu, Z.; Zhang, X.; Zhang, Z.; Qian, Y. Physicochemical Quality and Metabolomics Comparison of the Green Food Apple and Conventional Apple in China. *Food Research International* **2021**, *139*, 109804. https://doi.org/10.1016/j.foodres.2020.109804.

(125) Mie, A.; Laursen, K. H.; Åberg, K. M.; Forshed, J.; Lindahl, A.; Thorup-Kristensen, K.; Olsson, M.; Knuthsen, P.; Larsen, E. H.; Husted, S. Discrimination of Conventional and Organic White Cabbage from a Long-Term Field Trial Study Using Untargeted LC-MS-Based Metabolomics. *Anal Bioanal Chem* **2014**, *406* (12), 2885–2897. https://doi.org/10.1007/s00216-014-7704-0.

(126) Pappoe, J. A.; Mongson, O.; Amuah, C. L. Y.; Opoku-Ansah, J.; Adueming, P. O.-W.; Boateng, R.; Eghan, M. J.; Sackey, S. S.; Anyidoho, E. K.; Huzortey, A. A.; Anderson, B.; Vowotor, M. K.; Teye, E. Classification of Organic and Conventional Cocoa Beans Using Laser-Induced Fluorescence Spectroscopy Combined with Chemometric Techniques. *J Fluoresc* **2023**, *35* (1), 9–19. https://doi.org/10.1007/s10895-023-03499-3.

(127) Bechynska, K.; Sedlak, J.; Uttl, L.; Kosek, V.; Vackova, P.; Kocourek, V.; Hajslova, J. Metabolomics on Apple (Malus Domestica) Cuticle—Search for Authenticity Markers. *Foods* **2024**, *13* (9), 1308. https://doi.org/10.3390/foods13091308.

(128) Creydt, M.; Fischer, M. Food Authentication: Truffle Species Classification by Non-Targeted Lipidomics Analyses Using Mass Spectrometry Assisted by Ion Mobility Separation. *Mol. Omics* **2022**, *18* (7), 616–626. https://doi.org/10.1039/D2MO00088A.

(129) Sibono, L.; Grosso, M.; Tejedor-Calvo, E.; Casula, M.; Marco-Montori, P.; Garcia-Barreda, S.; Manis, C.; Caboni, P. A Critical Analysis of Adaptive Box-Cox Transformation for Skewed Distributed Data Management: Metabolomics of Spanish and Argentinian Truffles as a Case Study. *Analytica Chimica Acta* **2025**, *1345*, 343704. https://doi.org/10.1016/j.aca.2025.343704.

(130) Gao, J.-M.; Zhang, A.-L.; Chen, H.; Liu, J.-K. Molecular Species of Ceramides from the Ascomycete Truffle Tuber Indicum. *Chemistry and Physics of Lipids* **2004**, *131* (2), 205–213. https://doi.org/10.1016/j.chemphyslip.2004.05.004.

(131) Klare, J.; Rurik, M.; Rottmann, E.; Bollen, A.; Kohlbacher, O.; Fischer, M.; Hackl, T. Determination of the Geographical Origin of *Asparagus Officinalis* L. by^{1} H NMR Spectroscopy. *J. Agric. Food Chem.* **2020**, *68* (49), 14353–14363. https://doi.org/10.1021/acs.jafc.0c05642.

(132) Richter, B.; Gurk, S.; Wagner, D.; Bockmayr, M.; Fischer, M. Food Authentication: Multi-Elemental Analysis of White Asparagus for Provenance Discrimination. *Food Chemistry* **2019**, *286*, 475–482. https://doi.org/10.1016/j.foodchem.2019.01.105.

(133) Creydt, M.; Fischer, M. Metabolic Imaging: Analysis of Different Sections of White Asparagus Officinalis Shoots Using High-Resolution Mass Spectrometry. *Journal of Plant Physiology* **2020**, *250*, 153179. https://doi.org/10.1016/j.jplph.2020.153179.

(134) Creydt, M.; Arndt, M.; Hudzik, D.; Fischer, M. Plant Metabolomics: Evaluation of Different Extraction Parameters for Nontargeted UPLC-ESI-QTOF-Mass Spectrometry at the Example of White *Asparagus Officinalis*. *J. Agric. Food Chem.* **2018**, *66* (48), 12876–12887. https://doi.org/10.1021/acs.jafc.8b06037.

(135) Shami, A. A.; Akhtar, M. T.; Mumtaz, M. W.; Mukhtar, H.; Tahir, A.; Shahzad-ul-Hussan, S.; Chaudhary, S. U.; Muneer, B.; Iftikhar, H.; Neophytou, M. NMR-Based Metabolomics: A New Paradigm to Unravel Defense-Related Metabolites in Insect-Resistant Cotton Variety through Different Multivariate Data Analysis Approaches. *Molecules* **2023**, *28* (4), 1763. https://doi.org/10.3390/molecules28041763.

(136) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis – a Marriage of Convenience or a Shotgun Wedding. *Analytica Chimica Acta* **2015**, *879*, 10–23. https://doi.org/10.1016/j.aca.2015.02.012.

(137) Lee, T.-H.; Ullah, A.; Wang, R. Bootstrap Aggregating and Random Forest. In *Macroeconomic Forecasting in the Era of Big Data*; Fuleky, P., Ed.; Advanced Studies in Theoretical and Applied Econometrics; Springer International Publishing: Cham, 2020; Vol. 52, pp 389–429. https://doi.org/10.1007/978-3-030-31150-6_13.

(138) Nyamundanda, G.; Gormley, I. C.; Fan, Y.; Gallagher, W. M.; Brennan, L. MetSizeR: Selecting the Optimal Sample Size for Metabolomic Studies Using an Analysis Based Approach. *BMC Bioinformatics* **2013**, *14* (1), 338. https://doi.org/10.1186/1471-2105-14-338.

(139) Scornet, E.; Biau, G.; Vert, J.-P. Consistency of Random Forests. *Ann. Statist.* **2015**, *43* (4). https://doi.org/10.1214/15-AOS1321.

(140) Wright, M. N.; Ziegler, A. **Ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *J. Stat. Soft.* **2017**, *77* (1). https://doi.org/10.18637/jss.v077.i01.

(141) Živanović, V.; Seifert, S.; Drescher, D.; Schrade, P.; Werner, S.; Guttmann, P.; Szekeres, G. P.; Bachmann, S.; Schneider, G.; Arenz, C.; Kneipp, J. Optical Nanosensing of Lipid Accumulation Due to Enzyme Inhibition in Live Cells. *ACS Nano* **2019**, *13* (8), 9363–9375. https://doi.org/10.1021/acsnano.9b04001.

(142) Lösel, H.; Shakiba, N.; Wenck, S.; Le Tan, P.; Karstens, T.-O.; Creydt, M.; Seifert, S.; Hackl, T.; Fischer, M. Food Monitoring: Limitations of Accelerated Storage to Predict Molecular Changes in Hazelnuts (Corylus Avellana L.) under Realistic Conditions Using UPLC-ESI-IM-QTOF-MS. *Metabolites* **2023**, *13* (10), 1031. https://doi.org/10.3390/metabo13101031.

(143) Voges, L. F.; Jarren, L. C.; Seifert, S. Exploitation of Surrogate Variables in Random Forests for Unbiased Analysis of Mutual Impact and Importance of Features. *Bioinformatics* **2023**, *39* (8), btad471. https://doi.org/10.1093/bioinformatics/btad471.

(144) Gärber, F. *„Klassifizierung Und Charakterisierung von Spargel Mittels Spektroskopischer Und Spektrometrischer Verfahren*; Universität Hamburg: Hamburg, 2022.

(145) Lösel, H.; Brockelt, J.; Gärber, F.; Teipel, J.; Kuballa, T.; Seifert, S.; Fischer, M. Comparative Analysis of LC-ESI-IM-qToF-MS and FT-NIR Spectroscopy Approaches for the Authentication of Organic and Conventional Eggs. *Metabolites* **2023**, *13* (8), 882. https://doi.org/10.3390/metabo13080882.

Gesetzestexte:

Basisverordnung (EG) 178/2002:
VERORDNUNG (EG) Nr. 178/2002 DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 28. Januar 2002 zur Festlegung der allgemeinen Grundsätze und Anforderungen des Lebensmittelrechts, zur Errichtung der Europäischen Behörde für Lebensmittelsicherheit und zur Festlegung von Verfahren zur Lebensmittelsicherheit.

Lebensmittelinformationsverordnung (EG) 1169/2011:
VERORDNUNG (EU) Nr. 1169/2011 DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 25. Oktober 2011 betreffend die Information der Verbraucher über Lebensmittel und zur Änderung der Verordnungen (EG) Nr. 1924/2006 und (EG) Nr. 1925/2006 des Europäischen Parlaments und des Rates und zur Aufhebung der Richtlinie 87/250/EWG der Kommission, der Richtlinie 90/496/EWG des Rates, der Richtlinie 1999/10/EG der Kommission, der Richtlinie 2000/13/EG des Europäischen Parlaments und des Rates, der Richtlinien 2002/67/EG und 2008/5/EG der Kommission und der Verordnung (EG) Nr. 608/2004 der Kommission.