Identifizierung eigenschaftsrelevanter Metabolitencluster

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät II
Chemie, Physik und Mathematik

der Martin-Luther-Universität
Halle-Wittenberg

vorgelegt von

Herrn Dipl. Biol. Mark Haid

geb. am 20. April 1973 in Moers

1. Gutachter: Prof. Dr. Ludger Wessjohann
2. Gutachter: Prof. Dr. Jörg Heilmann

Tag der öffentlichen Verteidigung: 13.11.2020

For all the people I have lost

Inhaltsverzeichnis

Vorwort und Danksagung

Die experimentellen Arbeiten zu dieser Dissertation erfolgten zwischen Mai 2007 und Mai 2012 in der Abteilung Natur- und Wirkstoffchemie am Leibniz-Institut für Pflanzenbiochemie in Halle an der Saale. Die AcorA-Methode wurde 2008 in Zusammenarbeit mit Dr. Katharina Michels und Dr. André Gohr entwickelt. Ich möchte mich bei beiden ganz herzlich für die großartige Zusammenarbeit und Freundschaft bedanken.

Ein ganz großes Dankeschön geht an meinen Doktorvater Prof. Dr. Ludger Wessjohann. Er hat mich trotz meines schon etwas fortgeschrittenen Alters am IPB aufgenommen und mir mit dieser Arbeit an der Schnittstelle zwischen Naturstoffchemie und Chemoinformatik ein spannendes Dissertationsthema überlassen. Mein weiterer Dank gilt Prof. Dr. Jörg Heilmann (Lehrstuhl für Pharmazeutische Biologie, Universität Regensburg), der freundlicherweise seine Expertise für das Zweitgutachten zur Verfügung gestellt hat.

Die erfolgreiche Bearbeitung dieser Dissertation wäre jedoch ohne das „Team Technikum“ nicht möglich gewesen, zumal ich vorher noch keine Berührung mit dem Thema Naturstoffchemie hatte. Ein großer Dank gilt daher Dr. Norbert Arnold für die Einarbeitung in die Naturstoffchemie sowie sämtliche Informationen, Geschichten und Diskussionen rund um das Thema *Sepedonium* (und weitere Pilze). In diesem Zusammenhang seien auch Nicole Hünecke und Monika Kummer genannt, die mich tatkräftig bei der Anzucht der Pilze unterstützt haben.

Ein weiterer großer Dank geht an Dr. Jürgen Schmidt und Dr. Andrea Porzel. Jürgen hat mich in die „Geheimnisse“ des Bruker FT-ICR-MS eingeweiht und stand mit seinem enormen Fachwissen über Massenspektrometrie von Naturstoffen helfend zur Seite. Andrea möchte ich ebenfalls für die vielen Hilfestellungen rund um des Thema Analytik von Naturstoffen danken. Sie hat mit vielen Anregungen und kritischem Hinterfragen wesentlich zum Gelingen dieser Arbeit beigetragen.

Untrennbar mit meiner Arbeit im Technikum ist auch Dr. Katrin Franke verbunden. Ich möchte mich bei Katrin für die vielen Anregungen, Hilfestellungen und naturstoffchemischen Diskussionen bedanken. Bei Annika Denkert möchte ich mich für die Hilfe und Unterstützung bei den Zellkultur Experimenten bedanken.

Ein weiterer großer Dank geht an Gudrun Hahn, Anja Ehrlich, Christine Kuhnt und Martina Lerbs. Alle vier haben mich mit Rat und Tat bei der Aufreinigung und Messung diverser Substanzen unterstützt.

Bei Dr. Tilo Lübken und Dr. Christoph Böttcher möchte ich mich für die Unterstützung bei den UPLC-Qq-TOF-MS Messungen der *Sepedonium* Extrakte und Peptaibole bedan-

ken. Ohne ihre Hilfe wäre die Strukturaufklärung der Peptaibole nicht möglich gewesen.
Unverzichtbar für das gelingen dieser Arbeit war auch das Bioinformatik Team rund um Dr. Steffen Neumann. Vielen Dank für die Hilfe bei diversen R-Skripten, den (bioinformatischen) Diskussionen und die gemeinsame Zeit bei den Mittagessen. Neben Dr. Steffen Neumann seien an dieser Stelle insbesondere Dr. Ralf Tautenhahn und Carsten Kuhl hervorgehoben, die mein Leben auch privat durch viele gemeinsame Unternehmungen und Diskussionen bereichert haben.
Ein weiteres Dankeschön auch an „meine“ beiden Masteranden Kristin Schmatloch und Alexander Maxones. Mit der Unterstützung von Kristin wurde die Extraktionsprozedur von *Sepedonium* entwickelt und optimiert. Alexander danke ich für die Unterstützung bei der Anzucht und Extraktion unzähliger *Sepedonium* Stämme. Danke auch euch beiden für die zahlreichen gemeinsamen Abende bei Konzerten und Tanzveranstaltungen.
Vielen Dank auch an die damaligen Hiwis Dr. Carina Würfel, Julia Mülbradt und Joachim Kutzera für die Unterstützung im Labor und der Bioinformatik.

Ganz besonders möchte ich mich bei Elena und Denis Immel bedanken. Ihr habt mein Leben auf vielfältige Weise bereichert und die wunderschöne Zeit mit euch wird mir ewig in Erinnerung bleiben!

Ein besonderer Dank geht an die gesamte Postamtrunde (Carsten Kuhl, Eva Guttsche, Dr. Jan Grau, Dr. Yvonne Grau, Dr. André Gohr, Dr. Ralf Tautenhahn, Dr. Felix Bollenbeck, Christian Kirchert,...). Mit den vielen Unterhaltungen und Unternehmungen (u. a. Kanutouren auf Saale und Elster, Paragliding, zahlreiche Grillabende und Brunche, Joggingrunden, u. v. m.) sowie der Hilfe bei diversen Umzügen habt ihr meine Zeit in Halle zu einem unervergesslichen Erlebnis gemacht. Danke euch allen!

An dieser Stelle möcht ich mich auch ganz herzlich bei Dr. Caroline Muschet, Jonathan Adam und Julia Schulze Struchtrup für das Korrekturlesen der Dissertation bedanken.

Julia war es auch, die mich in den vergangenen Jahren trotz meiner zahlreichen Fast-Nervenzusammenbrüche bei der Ausarbeitung der Dissertation begleitet hat. Sie hat es verstanden mich immer wieder zu motivieren. Für die gemeinsame Zeit und deine Geduld möchte ich mich ganz herzlich bedanken!

Last but not least, möchte ich mich bei meiner Familie für die moralische und finanzielle Unterstützung in all den Jahren bedanken. Da ich einen Großteil meiner Urlaubszeit in den vergangenen sechs Jahren mit der Fertigstellung dieser Dissertation beschäftigt war, konnte ich leider viel zu selten bei euch sein.

Abkürzungsverzeichnis

ACN	Acetonitril
Aib	α-Aminoisobuttersäure
AcorA	Activity Correlation Analysis
Ala	Alanin
AMP	Adenosinmonophosphat
AmpA	Ampullosporin A
ANN	Artificial Neural Network
AUC	Area Under the Curve
BCD	Bioassay-coupled Chromatography
CART	Classification and Regression Trees
CID	Collision Induced Dissociation
Cov	Kovarianz
CPDB	Comprehensive Peptaibol Database
EryA	Erythromycin A
AEryA	Anhydromerythromycin A
EryAEO	Erythromycin A *N*-oxid
EryB	Erythromycin B
EryC	Erythromycin C
EryE	Erythromycin E
EryF	Erythromycin F
ESI-MS	Elektrospray-Ionisation-Massenspektrometrie
ESS	Explained Sums of Squares
FDA	U. S. Food and Drug Administration
FT-ICR-MS	Fouriertransform-Ion Cyclotron-Massenspektrometrie
Gln	Glutamin
Glu	Glutaminsäure
Gly	Glycin
HIV	Human Immunodeficiency Virus
HTS	High-Throughput Screening
HPLC	High-Performance Liquid Chromatography
HPTLC	High-Performance Thin-Layer Chromatography
IC	Inhibitory Concentration
ITS	Internal Transcriped Spacer

Iva	Isovalin
KSH	Kultursammlung Halle
Lasso	Least Absolute Shrinkage and Selection Operator
LB	Laborbuch
LC	Liquid chromatography
Leu	Leucin
Lxx	Leucin oder Isoleucin
MALDI	Matrix Assisted Laser Disorption/Ionisation
MeOH	Methanol
MIK	Minimale inhibitorische Konzentration
MLR	Multiple lineare Regression
MRSE	Methicillin resistent *Staphylococcus aureus*
MS	Massenspektrometrie
MS/MS	Tandem Massenspektrometrie
MSE	Mean Squarred Error
MSECV	Mean Squarred Error of Cross-Validation
MSEP	Mean Squarred Error of Prediction
NdMeEryA	*N*-demethylerythromycin A
NMR	Nuclear Magnetic Resonance)
NRPS	Non-ribosomal Peptidesynthase
OOB	out-of-bag
OPLS	Orthogonal Partial Least Squares
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCP	Peptidyl Carrier Protein
PLSR	Partial-Least-Squares Regression
Pro	Prolin
QPAR	Quantitative Pattern-Activity Relationship
QSAR	Quantitative Structure Activity Relationship
Rif	Rifampicin
rRNA	Ribosomal Ribonucleic Acid
RMSEP	Root Mean Squarred Error of Prediction
ROC	Receiver Operating Characteristic
RSS	Residual Sums of Squares

sp.	Spezies (Singular)
spp.	Spezies (Plural)
SR	Selectivity Ratio
SV	Säulenvolumen
STOCSY	Statistical Total Correlation Spectroscopy
SVM	Support Vector Machine
TLC	Thin Layer Chromatography
TOCSY	Total Correlation Spectroscopy
TOF	Time Of Flight
Trp	Tryptophan
TSS	Total Sums of Squares
TP	Target Projection
Tyr	Tyrosin
UPLC	Ultraperformance Liquid Chromatrography
Val	Valin
Var	Varianz
VIP	Variable Importance in Projection
Vxx	Valin oder Isovalin
WHO	World Health Organization

Zusammenfassung

Im Laufe der Evolution haben viele Pflanzen, Pilze und Mikroorganismen Sekundärmetaboliten ausgebildet, die u. a. zum Schutz vor Fraßfeinden genutzt werden. Im Gegensatz zu rein synthetischen Verbindungen durchliefen diese Naturstoffe somit einen über Jahrmillionen andauernden Prozess, bei dem sie auf Wechselwirkungen mit Enzymen, Rezeptoren, Biomembranen und anderen Biomolekülen optimiert wurden.
Diese besonderen physiko-chemischen Eigenschaften von Naturstoffen wurden bereits in der Vergangenheit erfolgreich für die Entdeckung und Entwicklung von Medikamenten genutzt und machen Naturstoffe auch heute noch für die pharmazeutische Forschung attraktiv.

Die Identifizierung neuer biologisch aktiver Verbindungen ist jedoch immer noch ein herausfordernder Prozess. Potenzielle Wirkstoffe sind in einer komplexen Matrix aus hunderten oder tausenden von Verbindungen eingebettet. Während viele klassische und Hochdurchsatz Screening Methoden nach einer Art „Brute-Force Algorithmus“ arbeiten, bietet eine **gezielte** Suche nach neuen, interessanten Strukturen in Rohextrakten eine hohe Zeit- und Kostenersparnis.

Ziel dieser Arbeit war die Entwicklung einer einfach zu erlernenden Methode, die eine *in silico* Identifizierung von biologisch aktiven Komponenten in komplexen Mischungen ermöglicht.
Die entwickelte AcorA-Methode nutzt die natürlichen (oder artifiziell induzierten) Konzentrationsunterschiede von Substanzen in einer Reihe von Rohextrakten aus den Testorganismen (oder des Testorganismus). Die Identifizierung der aktivitätsrelevanten Komponenten gelingt über die Erfassung des Bioaktivitäts- und Metabolitenprofils der Extrakte. Signale, deren Intensitäten eine signifikante Korrelation zur Bioaktivität aufweisen, besitzen eine erhöhte Wahrscheinlichkeit, kausal für die Bioaktivität verantwortlich zu sein. Anhand eines Permutationstests wird eine Signifikanzschwelle $cor.thresh$ festgelegt. Signale mit Korrelationskoeffizienten $\rho > cor.thresh$ werden in einer Hitliste annotiert.
Da eine Verbindung vor allem in der Elektrospray-Ionisation-Massenspektrometrie (ESI-MS) meist mehrere Isotopen- und Adduktpeaks verursacht, findet man diese in der Hitliste angereichert. Die hohe Korrelation zwischen diesen Signalen kann z. B. durch Visualisierung in einer Korrelationsmatrix oder eines Korrelationsnetzwerkes zur Dekonvolution der Hitliste genutzt werden.
Um die entwickelte AcorA-Methode zu testen und mit anderen Methoden vergleichen zu können, wurde in einem Proof of Concept Experiment die Suche nach antibiotisch aktiven Verbindungen in Pilzextrakten simuliert. Dazu wurden biologisch inaktive Pilzextrakte (als

Hintergrundmatrix) randomisiert mit verschiedenen Antibiotika (Amoxicillin, Erythromycin, Rifampicin) versetzt und in Hinblick auf ihre wachstumsinhibierende Wirkung von *Bacillus subtilis* charakterisiert. Die hergestellten Extrakte sowie die Antibiotikastammlösungen wurden anschließend mit einem FT-ICR-MS gemessen.
Nach Bestimmung der Signifikanzschwelle über einen Permutationstest bestanden die Hitlisten im Positiv- und Negativ-Ionen-Modus zu 67 % bzw. 41 % aus den Signalen der Antibiotika. Somit wurden die Antibiotikasignale in den Hitlisten weit oben und für den Anwender leicht erkennbar angereichert.

Die Qualität von AcorA wurde anschließend mit einer Reihe von multivariaten Datenanalysemethoden verglichen. Dazu wurden neben der unüberwachten Principal Component Analysis (PCA) vor allem überwachte Methoden getestet. Neben zwei klassischen Methoden (Principal Component Regression (PCR) und Partial-Least-Squares Regression (PLSR)) wurden die Quantitative Pattern-Activity Relationship (QPAR), eine Entscheidungsbaum basierte Methode (Random Forest) sowie verschiedene Penalisierungsmethoden (Ridge Regression, Lasso und Elastic Net) untersucht.

In den FT-ICR-MS Spektren der Antibiotikastammlösungen wurden insgesamt 47 Antibiotikasignale identifiziert und als Grundmenge der für die verschiedenen Datenanalysemethoden zu findenden Peaks festlegt. Die Qualität der Modelle wurde über das Verhältnis aus Richtig-Positiv-Rate (Trefferquote, Sensitivität) und Falsch-Positiv-Rate (1-Spezifität) in ROC-Kurven abgeschätzt. Zur Beurteilung der Variablenselektion wurde der F_1-Wert, der dem harmonischen Mittel aus Genauigkeit (Precision) und Trefferquote (Recall) entspricht, herangezogen.

Für die entwickelte AcorA-Methode wurde ein exzellenter AUC-Wert von 0,99 im Positiv-Ionen-Modus erhalten. Unter den ersten 50 bzw. 100 Signalen der sortierten Liste konnten 38 (Trefferquote = 81 %) bzw. 44 (Trefferquote = 94 %) der 47 Antibiotika Signale annotiert werden. Bezogen auf die Signale in der Hitliste erzielte AcorA einen F_1-Wert von 76 %. AcorA wurde lediglich von der Elastic Net Analyse (F_1 = 77 %) übertroffen. Die Elastic Net Analyse stellt somit ebenfalls eine hoch attraktive Variablenselektionsmethode dar.
Ergebnisse von mittlerer Güte lieferten PLSR-VIP (F_1 = 46 %, AUC = 0,96), Random Forest (F_1 = 45 %, AUC = 0,98) und QPAR (F_1 = 40 %, AUC = 0,99). Insbesondere die QPAR-Methode besitzt mit den höchsten erzielten Trefferquoten unter den Top 50 Peaks (40/47 = 85 %) bzw. Top 100 Peaks (47/47 = 100 %) ein wesentlich höheres Potenzial als der F_1-Wert andeutet. Der zur Ermittlung eines Schwellenwertes verwendete F-Test

ist jedoch deutlich zu konservativ, sodass viele Richtig-Positive Signale nicht in die Hitliste aufgenommen wurden.
Als unüberwachte Methode lieferte die PCA mit einem AUC von 0,90 und Trefferquoten von 60 % (Top 50 Peaks) und 72 % (Top 100 Peaks) erwartungsgemäß die schlechtesten Ergebnisse.

Nachdem die AcorA-Methode in dem Proof of Concept Experiment auf ihre Eigenschaften zur Selektion der aktivitätsrelevanten Signale untersucht wurde, sollte die Eignung der Methode in einem „real case“ Experiment demonstriert werden.
Dazu wurden 21 Pilzstämme der fungicol lebenden Spezies *Sepdonium ampullosporum* extrahiert. Die Extrakte wurden anschließend auf ihre Zytotoxizität gegen humane HT-29 Zellen getestet und die Metabolitenprofile mittels FT-ICR-MS ermittelt. Durch Verwendung von AcorA konnten zwei Hitlisten mit je 115 und 94 Peaks generiert werden, in denen die Massensignale eine hohe Korrelation zur Bioaktivität aufwiesen. Um die Auswertung der Hitlisten zu erleichtern, wurden die Korrelationskoeffizienten und die Peakdichte gegen *m/z* aufgetragen. Bereiche in der Hitliste mit hohen Korrelationskoeffizienten und hoher Peakdichte weisen auf Isotopen- und Adduktpeaks von Verbindungen hin, die statistisch signifikant mit der zytotoxischen Wirkung korreliert sind. Auf diese Weise konnten insgesamt sechs Peakcluster lokalisiert werden. Die Peaks in Peakcluster 1 und 2 sowie in Peakcluster 3-6 wiesen jeweils untereinander einen hohen Grad an Multikollinearität auf. Tatsächlich wurden die Verbindungen innerhalb der Peakcluster 1 und 2 bzw. 3 und 4 jeweils in den selben Pilzstämmen detektiert, was auf einen gemeinsamen biosynthetischen Ursprung schließen lässt. Durch Datenbanksuche und *de novo* Sequenzierung konnte Verbindung **61** als Ampullosporin A identifziert werden. Nach der Isolierung wurde für Ampullosporin A eine moderate zytotoxische Aktivität gegen HT-29 Zellen (IC_{50} = 4,46 µM) ermittelt.
Neben Ampullosporin A wurden für die Verbindungen in den Peakclustern 1-4 durch *de novo* Sequenzierung die Primärstrukturen von 11 weiteren, bislang unbekannten Peptaibolen bestimmt, die signifikant mit der zytotoxischen Aktivität korreliert sind. Da diese Verbindungen nicht isoliert und einzeln getestet wurden, wurde die Zytotoxizität durch Vergleich von Primärstruktur und Polarität (Retentionszeit) mit Ampullosporin A abgeschätzt.
Aufgrund der genannten Kriterien kann für die Verbindungen **62**, **67** und **68** eine zytotoxische Wirkung postuliert werden. Da verschiedene Studien darauf hinweisen, dass der *N*-Terminus essenziell für die Bioaktivität der Ampullosporine ist, kann vermutet werden, dass die Verbindungen **63**, **64**, **69** und **70** keine zytotoxischen Aktivitäten aufweisen und lediglich durch Kokorrelation mit Ampullosporin A in den Hitlisten annotiert wurden. Die Verbindungen **65**, **66**, **71** und **72** besitzen ein für Peptaibole ungewöhnliches *N*-terminales

Fragment, für das massenspektrometrisch lediglich eine Summenformel von $C_7H_{10}NO_3^+$ sowie ein Neutralionenverlust von 44 Da ermittelt werden konnte. In jedem Fall besitzen die Verbindungen eine ähnliche Polarität wie Ampullosporin A, sodass auch für sie eine zytotoxische Wirkung postuliert werden kann.

Die entwickelte AcorA-Methode konnte somit sowohl für die Identifizierung von Antibiotika in einem simulierten Proof of Concept Experiment als auch zur Identifizierung von zytotoxischen Verbindungen in Pilzextrakten erfolgreich verwendet werden. Die Methode ist im Gegensatz zu den multivariaten Methoden vergleichsweise intuitiv und auch für informatisch ungeübte Anwender leicht zu erlernen. Die *in silico* Identifizierung der aktivitätsrelevanten Signale im Rohextrakt erlaubt eine frühzeitige Dereplikation sowie eine zeitsparende *m/z*-geleitete Isolierung.

Summary

In the course of evolution, many plants, fungi, and microorganisms have developed secondary metabolites as defense against enemies. In contrast to synthetic compounds, these natural products were optimised towards interactions with enzymes, receptors, biomembranes, and other biomolecules in a million year long process.
The particular physico-chemical properties of natural products have already been exploited for discovery and development of new drugs, and still make them attractive for pharamceutical research.
However, the identification of new biological entities still is a challenging process. Potential active components are embedded in a matrix of hundreds or thousands of other compounds. While many classic and high-throughput screening methods are based on a "brute-force algorithm", a **guided** search for new interesting compounds would be beneficial in terms of saving time and costs.

The goal of this PhD thesis was to develop a new method that allows an *in silico* identification of biological active components in complex mixtures.
The developed AcorA method uses the natural (or artificially induced) concentration differences of compounds in raw extracts of organisms. The active ingredients can be identified by acquisition and subsequent correlation of the bioactivity- and metabolite profiles of the raw extracts. Signals that exhibit a significant correlation to the bioactivity are suspicious of being causually related to the bioactivity. A significance threshold $cor.thresh$ is determined by a permutation test. Signals with a correlation coefficient $\rho > cor.thresh$ are annotated in a hitlist.

In the electrospray ionisation mass spectrometry (ESI-MS) one molecule usually gives rise to multiple mass signals, consisting of isotopic and adduct peaks. Due to their high multicollinearity, multiple signals of a bioactive component are enriched in the AcorA hitlist and thus support a "real" correlation. Moreover, the hitlist can be deconvoluted by visualisation of the signals in a correlation matrix or a correlation network.

To test the developed AcorA method and to allow comparison with other variable selection algorithms, the discovery of antibiotics in fungal extracts was simulated in a proof of concept experiment. For that purpose, biologically inactive fungi extracts (simulating a background matrix) were randomly spiked with different antibiotics (Amoxicillin, Erythromycin, Rifampicin). The generated extracts were subsequently tested for their growth inhibiting properties against *Bacillus subtilis*. The metabolite profiles of the extracts as

well as the pure antibiotics solutions were determined by FT-ICR-MS. The resulting hitlists of the positive and negative ion mode consisted of 67 % and 41 % antibiotics signals, respectively. Thus, the signals of the antibiotics were highly enriched in the hitlist, which facilitates the identification of the active compounds.

The performance of AcorA was subsequently compared with different multivariate data analysis techniques. Besides the unsupervised principal component analysis (PCA), a special focus was set on several supervised methods. Additionally to classical methods like principal component regression (PCR) and partial-least-squares regression (PLSR), the quantitative pattern-activity relationship (QPAR), a decision tree approach (Random Forest), and several shrinkage methods (Ridge Regression, Lasso, and Elastic Net) were evaluated.

Using FT-ICR-MS for measurement of the antibiotics stock solutions, 47 signals of the antibiotics were identified and determined as basic set to be found by the different algorithmic approaches. The quality of the models as evaluated by the ratio of the true postive rate (aka. sensitivity, hit rate, recall) and the false positive rate (1-specificity) using ROC curves. The F_1-score, defined as the harmonic mean of precision and recall, was used to evaluate the quality of the variable selection procedures.

For the positive ion mode, an excellent AUC score of 0.99 was obtained with the AcorA method. All in all, 38 (hit rate = 81 %) and 44 (hit rate = 94 %) antibiotics signals were annotated beyond the first 50 and 100 signals in the sorted peak list, respectively. In terms of variable selection, AcorA achieved an F_1 score of 76 %. This value was only outperformed by the elastic net analysis (F_1 = 77 %), which turned out to be an excellent variable selection method as well.
Results of medium quality were obtained with PLSR-VIP (F_1 = 46 %, AUC = 0.96), Random Forest (F_1 = 45 %, AUC = 0.98) und QPAR (F_1 = 40 %, AUC = 0.99). However, particularly the QPAR method achieved the highest hits rates beyond the top 50 peaks (40/47 = 85 %) and top 100 peaks (47/47 = 100 %), and thus has a much higher potential for variable selection than the F_1 score pretends. The F-test that is used in QPAR for calculating a threshold for the target projected loadings appears to be too conservative and prevented true positive peaks from being annotated in the hit list.
As expected, the unsupervised PCA performed worst and achieved the lowest AUC (0.90) and hit rates (top 50 = 60 %, top 100 = 72 %) of the methods under investigation.

After AcorA was evaluated in terms of variable selection properties in the proof of concept experiment, its full potential in a “real case“ was demonstrated.
Therefore, 21 fungi strains of the fungicolous species *Sepdonium ampullosporum* were ex-

tracted. The extracts were subsequently tested for cytotoxicity against human HT-29 colon cancer cells. The respective metabolite profiles were determined by FT-ICR-MS. The application of the AcorA method resulted in 2 hitlists, containing 115 and 94 mass signals with a significant correlation to the bioactivity. To facilitate data analysis, the correlation coefficients and the peak density were plotted against *m/z*. Areas in the hitlist with high correlation coefficients and high peak density indicate isotopic and adduct signals of compounds that are statistically significantly associated with the bioactivity.
In this way, six peak clusters could be determined. The peaks within peak cluster 1 and 2 and peak cluster 3-6 exhibited a high degree of multicollinearity. Indeed, the compounds in peak cluster 1 and 2 and peak cluster 3 and 4, respectively, were detected in the same fungi strains, indicating a common biosynthetic origin. A data base search and *de novo* sequencing identified compound **61** as Ampullosporin A. After isolation, a moderate cytotoxic activity against HT-29 cells was determined for Ampullosporin A (IC_{50} = 4.46 µM). The primary structure of 11 other, to date unkown, peptaibols that are significantly correlated with the cytotoxic activty were determined by *de novo* sequencing. Because these compounds were not isolated, their cytotoxcity was appraised by comparing primary structure and polarity (retention time) with Ampullsporin A. Based on these cirteria, a cytotoxic activity can be postulated for **62**, **67**, and **68**. Several studies show that the *N*-terminus of the ampullosporins is essential for their bioactivity. Thus, it can be concluded that compounds **63**, **64**, **69**, and **70** are devoid of cytotoxic activity. Presumably, they were annotated in the hitlist due to high multicollinearity to Ampullosporin A. For the *N*-terminus of the compounds **65**, **66**, **71**, and **72** an unusual fragment with a sum formula of $C_7H_{10}NO_3^+$ an a neutral ion loss of 44 Da were observed. However, these compounds have a similar polarity comparable with Ampullosporin A. Due to this fact, it can be concluded that **65**, **66**, **71**, and **72** possess cytotoxic properties.

In summary, AcorA was successfully applied for the identification of antibiotics in the simulated proof of concept experiment as well as for the identification of cytotoxic peptaibols in fungal extracts. In contrast to most multivariate methods, the concept of AcorA is intuitive and easy to learn even for informatically unexperienced users. Futhermore, the *in silico* identification of the activity relevant signals in raw extracts allows dereplication at an early stage and a time-saving *m/z* guided isolation.

Mathematische Vorbemerkungen

1. Schreibkonventionen und allgemeine Grundlagen der multivariaten Datenanalyse

Dieses einleitende Kapitel soll einige in der Chemometrie häufig verwendete Methoden und Konzepte der multivariaten Datenanalyse erläutern. Im weiteren Verlauf der Arbeit werden die vorgestellten Konzepte als gegeben vorausgesetzt und lediglich Änderungen gegenüber dem Grundalgorithmus dargestellt. Auf diese Weise sollen lange mathematische Ausführungen im Fließtext vermieden werden.

1.1. Allgemeine Schreibkonventionen

Der allgemeinen Konvention folgend werden Vektoren mit Kleinbuchstaben im Fettdruck (z. B. $\mathbf{x}$) und Matrizen mit Großbuchstaben im Fettdruck (z. B. $\mathbf{X}$) dargestellt. Vektoren sind als Spaltenvektoren zu verstehen. Zeilenvektoren werden als transponierte Spaltenvektoren (z. B. $\mathbf{x}^\mathsf{T}$) dargestellt. Matrizen sind so angeordnet, dass die Objekte $x_1, x_2, x_i \ldots, x_n$ (z. B. Proben) in Zeilen und die Variablen $x_{i1}, x_{i2}, x_{ij} \ldots, x_{np}$ (z. B. m/z Werte, Wellenlängen) in Spalten angeordnet sind. Bei einer transponierten Matrix $\mathbf{X}^\mathsf{T}$ sind Zeilen und Spalten miteinander vertauscht:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}; \quad \mathbf{X}^\mathsf{T} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

Die Normierung von Vektoren erfolgt über die euklidische Norm (ℓ_2 Norm):

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2} = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}} \tag{2}$$

In der **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (Lasso) Analyse findet auch die ℓ_1 Norm ihre Anwendung:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |\mathbf{x}_i| \tag{3}$$

Ein Vektor bestehend aus n x 1 wird als $\mathbf{1}_n$ dargestellt:

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{4}$$

Eine n x p Matrix gefüllt mit 1 ergibt sich aus:

$$\mathbf{1}_n\mathbf{1}_p^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{5}$$

Eine Diagonalmatrix $\mathbf{D}$ ist eine quadratische p x p Matrix und enthält auf den Außerdiagonalelementen ausschließlich die Zahl Null:

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix} \tag{6}$$

Ein Spezialfall der Diagonalmatrix ist die Identitätsmatrix $\mathbf{I}_p$, deren Diagonalelemente aus p x 1 besteht:

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{7}$$

Die Inverse einer Matrix ist wie folgt definiert:

$$\mathbf{X}^{-1}\mathbf{X} = \mathbf{I}_\mathrm{p} \tag{8}$$

Weiterhin gilt:

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{9}$$

Für das Produkt aus $(\mathbf{AB})^T$ gilt:

$$(\mathbf{AB})^T = \mathbf{B}^\top\mathbf{A}^\top \tag{10}$$

Eine Matrix is orthogonal, wenn gilt:

$$\mathbf{X}^\top = \mathbf{X}^{-1} \tag{11}$$

Daraus folgt:

$$\mathbf{X}^\top\mathbf{X} = \mathbf{X}\mathbf{X}^\top = \mathbf{I}_\mathrm{p} \tag{12}$$

Das Skalarprodukt (inneres Produkt) zweier Vektoren **a** und **b** ist definiert durch:

$$\mathbf{a}^\top\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta_{a,b} \tag{13}$$

Der Winkel θ zwischen den beiden Vekoren ergibt sich daher über:

$$\cos\theta_{a,b} = \frac{\mathbf{a}^\top\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|} \tag{14}$$

1.2. Zentrierung und Standardisierung von Matrizen

Für viele Anwendungen muss die n x p Datenmatrix zunächst zentriert werden. D. h. von jedem Datenpunkt x_{ij} wird der Spaltenmittelwert $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ abgezogen:

$$\mathbf{c}_\mathrm{j} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j) \tag{15}$$

Hierfür wird die s. g. Zentrierungsmatrix **J** verwendet [1]:

$$\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_\mathrm{n}\mathbf{1}_\mathrm{n}^\top \tag{16}$$

$\mathbf{J}$ ist symmetrisch und idempotent, d. h. $\mathbf{J} = \mathbf{J}^\mathsf{T}$ und $\mathbf{J}^2 = \mathbf{JJ} = \mathbf{J}$. Die Zentrierung eines Spaltenvektors $\mathbf{x}_\mathrm{j}$ erhält man durch die Prämultiplikation mit $\mathbf{J}$:

$$\mathbf{c}_\mathrm{j} = \mathbf{x}_\mathrm{j} - \begin{pmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{pmatrix} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_\mathrm{n}\mathbf{1}_\mathrm{n}^\mathsf{T}\mathbf{x}_\mathrm{j} = \mathbf{J}\mathbf{x}_\mathrm{j} \tag{17}$$

Die zentrierte Datenmatrix $\mathbf{C}$ ergibt sich somit aus:

$$\mathbf{C} = \mathbf{JX} \tag{18}$$

Eine weitere, häufig verwendete Datenpräprozessierungsmethode ist die z-Transformation (auch Autoskalierung oder Standardisierung genannt). Bei der z-Transformation werden die zentrierten Variablen durch die Standardabweichung dividiert (Gl. 19). Auf diese Weise werden die Daten unabhängig von der Größe der Variablen.

$$z_j = \frac{x_{ij} - \bar{x}_j}{\sqrt{Var_j}} \tag{19}$$

Mit der Standardabweichung

$$\sqrt{Var_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} = \sqrt{\frac{1}{n}\left\|\mathbf{J}\mathbf{x}_\mathrm{j}\right\|^2} = \frac{1}{\sqrt{n}}\left\|\mathbf{J}\mathbf{x}_\mathrm{j}\right\| \tag{20}$$

ergibt sich der zentrierte Spaltenvektor $\mathbf{z}_\mathrm{j}$ aus:

$$\mathbf{z}_\mathrm{j} = \frac{1}{\sqrt{Var_j}}\mathbf{J}\mathbf{x}_\mathrm{j} = \frac{1}{\sqrt{Var_j}}\mathbf{c}_\mathrm{j} \tag{21}$$

Mit der Diagonalmatrix $\mathbf{S}$ bestehend aus den Standardabweichungen

$$\mathbf{S} = \begin{pmatrix} \sqrt{Var_j} & & \\ & \ddots & \\ & & \sqrt{Var_p} \end{pmatrix} \tag{22}$$

gelingt die Zentrierung einer Datenmatrix $\mathbf{X}$ über:

$$\mathbf{Z} = \left(\frac{1}{\sqrt{Var_1}}\mathbf{c}_1, \cdots, \frac{1}{\sqrt{Var_p}}\mathbf{c}_p \right) = (\mathbf{c}_1, \cdots, \mathbf{c}_p) \begin{pmatrix} \frac{1}{\sqrt{Var_j}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{Var_p}} \end{pmatrix} \tag{23}$$

$$= \mathbf{CS}^{-1} = \mathbf{JXS}^{-1}$$

1.3. Kovarianz- und Korrelationsmatrix

Ausgangspunkt für viele Algorithmen ist die Kovarianzmatrix $\mathbf{V}$ (Gl. 25). Die Kovarianz entspricht dem Skalarprodukt (inneres Produkt) der zentrierten Variablenvektoren $\mathbf{c}_i = \mathbf{Jx}_j$ und $\mathbf{c}_k = \mathbf{Jx}_k$ [1]:

$$Cov = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k}) = \frac{1}{n}\left(\mathbf{Jx}_j\right)^T \mathbf{Jx}_k = \frac{1}{n}\mathbf{x}_j^\top \mathbf{Jx}_k = \frac{1}{n}\mathbf{c}_j^\top \mathbf{c}_k \tag{24}$$

In der Matrixrepräsentation ergibt sich:

$$\mathbf{V} = \frac{1}{n}\begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_p^\top \end{pmatrix} \mathbf{J}^\top\mathbf{J} \left(\mathbf{x}_1 \cdots \mathbf{x}_p \right) = \frac{1}{n}\mathbf{X}^\top\mathbf{J}^\top\mathbf{JX} = \frac{1}{n}\mathbf{X}^\top\mathbf{JX} \tag{25}$$

Auf den Diagonalelementen, d. h. im Falle von $j = k$, befindet sich die Varianz einer Variable $\mathbf{x}_j$. Die Kovarianzmatrix wird daher auch als Varianz-Kovarianzmatrix bezeichent. Wird die Kovarianz zusätzlich auf die Standardabweichung normiert, erhält man die Korrelationskoeffizienten:

$$\rho = \frac{Cov}{\sqrt{Var_j}\sqrt{Var_k}} = \frac{\mathbf{x}_j^\top \mathbf{Jx}_k}{\sqrt{\mathbf{x}_j^\top \mathbf{Jx}_j}\sqrt{\mathbf{x}_k^\top \mathbf{Jx}_k}} = \frac{(\mathbf{Jx}_j)^T \mathbf{Jx}_k}{\sqrt{(\mathbf{Jx}_j)^T \mathbf{Jx}_j}\sqrt{(\mathbf{Jx}_k)^T \mathbf{Jx}_k}} = \frac{\mathbf{c}_j^\top \mathbf{c}_k}{\|\mathbf{c}_j\| \, \|\mathbf{c}_k\|} = \frac{1}{n}\mathbf{C}^\top\mathbf{C} \tag{26}$$

Der Korrelationskoeffizient entspricht somit dem Winkel $\cos\theta$ zwischen zwei Vektoren $\mathbf{c}_j$ und $\mathbf{c}_k$. Sind die beiden Vektoren orthogonal, d. h. θ beträgt 90°, wird $\cos\theta = 0$ und die beiden Vektoren sind somit unkorreliert. Eine positive Korrelation erhält man bei Winkeln $0 < \theta < 90°$, eine negative Korrelation bei Winkeln $90° < \theta < 180°$. Die Korrelationsmatrix

ergibt sich analog zur Kovarianzmatrix aus:

$$\mathbf{K} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} = \frac{1}{n}\mathbf{S}^{-1}\mathbf{X}^\top\mathbf{J}\mathbf{X}\mathbf{S}^{-1} = \mathbf{S}^{-1}\mathbf{V}\mathbf{S}^{-1} \tag{27}$$

Wird eine Datenmatrix $\mathbf{X}$ bereits vor der Berechnung der Kovarianz/Korrelationsmatrix zentriert/z-transformiert, so reduziert sich die Berechnung der Kovarianz- respektive Korrelationsmatrix auf $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$. Dies kann mit der Statistiksoftware R beispielsweise über den Befehl scale(X,center = TRUE, scale = FALSE) oder scale(X,center = TRUE, scale = TRUE) erreicht werden.

1.4. Grundlagen der (multiplen) linearen Regression

Grundlage für viele der multivariaten Regressionsmethoden ist die multiple lineare Regression (MLR). In der MLR soll ein funktioneller Zusammenhang $y = f(x)$ zwischen einigen unabhängigen Variablen x (Regressoren, Kovariablen, Prädiktoren) und einer abhängigen Variable y (Regressand) erfasst werden. Die Vektoren $\mathbf{x}^\top$ sind in der Regel Spektren, bestehend aus einzelnen Variablen $x_1, x_2, \ldots, x_p$. Der Regressand y ist häufig die Konzentration des Analyten oder wie in dem hier diskutierten Fall die Bioaktivität. Die Regressionskoeffizienten β_i geben den Anteil an, mit dem die entsprechende Variable x_i an der Vorhersage des Regressanden y beteiligt ist. Da jede Messung mit einem Fehler behaftet ist, wird ein Fehlerterm ϵ zur Regressionsgleichung hinzugefügt (Gl. 28).

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_p x_p + \epsilon \tag{28}$$

In Matrixschreibweise ist der Ansatz für die MLR eines Datensatzes wie folgt darstellbar:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{29}$$

Da die tatsächlichen Regressionskoeffizienten β_i unbekannt sind, werden diese durch einen Vektor $\mathbf{b}$ geschätzt. Auf diese Weise wird ein neuer Regressandenvektor

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{30}$$

berechnet, der $\mathbf{y}$ umso besser schätzt, je geringer die Residuen $\mathbf{e}$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{31}$$

sind. Zur Abschätzung des Fehlers wird die Summe der quadrierten Residuen (Fehlerquadratsumme, FQS) verwendet:

$$\|\mathbf{e}\|^2 = \mathbf{e}^\top\mathbf{e} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 + x_i b)^2 \tag{32}$$

Eine optimale Schätzung für den Koeffizientenvektor $\mathbf{b}$ wird erhalten, wenn die Fehlerquadrate minimiert werden (Methode der kleinsten Fehlerquadrate, engl.: ordinary least squares, OLS):

$$\min \| \mathbf{y} - \hat{\mathbf{y}} \|^2 = \min_{\mathbf{b}} \| \mathbf{y} - \mathbf{Xb} \|^2 \tag{33}$$

Die Lösung des Minimierungsproblems ist in Gleichung 34 angegeben und wird als Kleinste-Quadrate-Schätzer (KQ-Schätzer) bezeichnet:

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \tag{34}$$

Der Erwartungswert E für y ist also:

$$E(\mathbf{y}) = \hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \tag{35}$$

Die Minimierung der Restvarianz $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ entspricht einer Orthogonalprojektion von $\mathbf{y}$ auf einen Unterraum (Spaltenraum) von $\mathbf{X}$. Dies ist in Abbildung 1 am Beispiel von zwei Variablen $\mathbf{x}_1$ und $\mathbf{x}_2$ veranschaulicht. Für ein lineares Modell ist der Least-Squares Schätzer

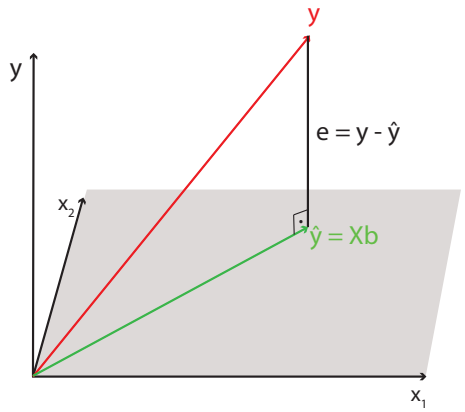


Abbildung 1.: Orthogonalprojektion von y auf die Hyperebene aus x_1 und x_2.

der beste erwartungstreue (unverzerrte) Schätzer mit minimaler Varianz (Gauß-Markow-Theorem) [2]. Hierfür müssen folgende Bedingungen erfüllt sein:

1. die Residuen müssen linear unabhängig sein (Orthogonalität, $Cov(e_i, e_j) = 0$)
2. die Residuen besitzen im Mittel einen Wert von Null ($E(e_i) = 0$)
3. die Residuen müssen gleichverteilt sein (Homoskedastizität)

Wenn wie im Falle von Massen- oder auch IR-Spektren die Anzahl von x-Variablen die Anzahl der Proben übersteigt, existiert die Inverse von $\mathbf{X}^\top\mathbf{X}$ nicht. In diesem Falle wird die s. g. generalisierte (oder auch Pseudoinverse) Matrix verwendet:

$$\mathbf{X}^+ = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \tag{36}$$

Durch Einsetzen von Gleichung 36 in Gleichung 34 kann nun der Koeffizientenvektor berechnet werden:

$$\mathbf{b} = \mathbf{X}^+\mathbf{y} \tag{37}$$

Mit steigender Anzahl von x-Variablen liefert die MLR selbst unter Verwendung der Pseudoinversen keine zuverlässigen Ergebnisse mehr, da zunehmend Multikollinearität auftritt. (Multi)Kollinearität bezeichnet den Umstand, dass unter mehreren x-Variablen (zufällige) Korrelationen auftreten, die Variablen also untereinander linear abhängig sind. Durch die kollinearen x-Variablen werden die x-Variablen, die die Zielgröße tatsächlich beschreiben, verschleiert. Beispielsweise würden die Konzentrationen inaktiver Zersetzungsprodukte einer bioaktiven Substanz automatisch zu einer falsch positiven Korrelation zur Bioaktivität führen. Weiterhin führt Multikollinearität dazu, dass die Regressionskoeffizienten nur sehr ungenau geschätzt werden können, d. h. die Varianz der Least-Squares-Schätzer nimmt mit steigender Korrelation zwischen der x-Variablen zu [2]:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2 \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)} \tag{38}$$

Der Anteil der Multikollinearität in einem Datensatz lässt sich über den Varianzinflationsfaktor (VIF) abschätzen:

$$VIF_j = \frac{1}{1 - R_j} \tag{39}$$

1.5. Modellvalidierung

Um die Güte eines Modells quantitativ abschätzen zu können, existieren im Wesentlichen zwei verschiedene Verfahren:

1. Das Bestimmtheitsmaß R^2

2. Der mittlere quadratische Fehler (Mean Squarred Error (MSE))

Das Bestimmtheitsmaß R^2 Die Gesamtvarianz (Total Sums of Squares (TSS)) eines Regressionsmodells setzt sich aus der erklärten Varianz (Explained Sums of Squares (ESS)) und der Restvarianz (Residual Sums of Squares (RSS)) zusammen:

$$TSS = ESS + RSS \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{40}$$

Das Bestimmheitsmaß gibt den Anteil der erklärten Varianz im Regressionsmodell an.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{41}$$

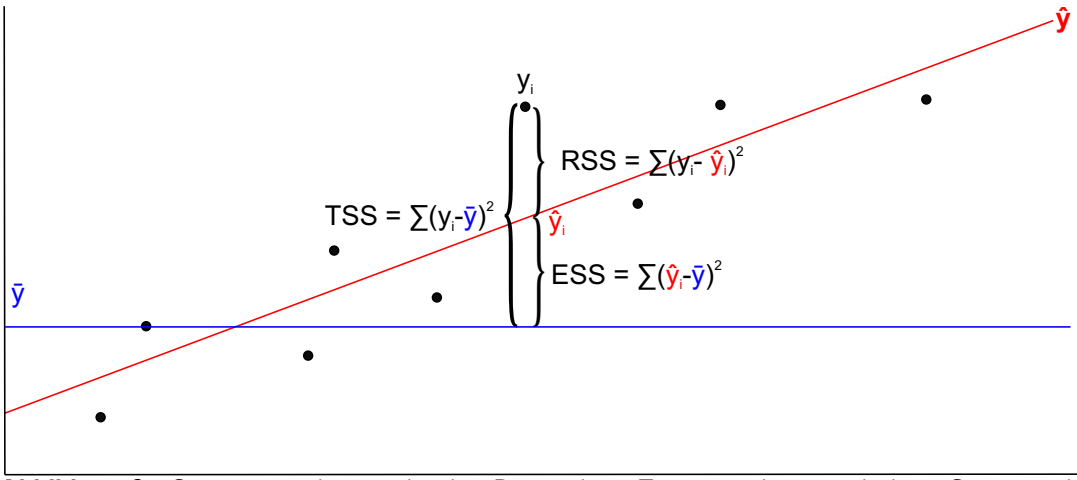


Abbildung 2.: Streuungszerlegung in der Regression. Zusammenhang zwischen Gesamtvarianz (TSS), erklärter Varianz (ESS) und Restvarianz (RSS). TSS = ESS + RSS

Speziell beschreibt es den Anteil der erklärten Varianz in den x-Variablen an der Varianz der abhängigen Variablen **y**:

$$R^2 = \frac{\|\mathbf{J}\hat{\mathbf{y}}\|^2}{\|\mathbf{J}\mathbf{y}\|^2} = 1 - \frac{\|\mathbf{e}\|^2}{\|\mathbf{J}\mathbf{y}\|^2} = 1 - \frac{\|\mathbf{J}\mathbf{X}\mathbf{b}\|^2}{\|\mathbf{J}\mathbf{y}\|^2} \tag{42}$$

Weiterhin kann das Bestimmtheitsmaß R^2 auch als quadrierter Korrelationskoeffizient zwischen **y** und $\hat{\mathbf{y}}$ aufgefasst werden. R^2 gilt streng genommen nur für den Fall, dass der untersuchte Zusammenhang tatsächlich linear ist. Bei Abweichungen wird R^2 zwar kleiner, es ermöglicht jedoch keine Aussage darüber, ob das Modell tatsächlich richtig spezifiziert ist. Weiterhin erlaubt das Bestimmheitsmaß keine Aussage über die Qualität der Vorhersage bei einem unabhängigen Testdatensatz, der nicht an der Bildung des Modells beteiligt war.

Mittlerer Vorhersagefehler (Mean Squarred Error of Prediction (MSEP)) Als Maß für die Vorhersagegüte eines Modells wird in der Regel der mittlere Vorhersagefehler (Mean-Squared Error Of Prediction, MSEP) berechnet. Dieser gibt die mittlere Abweichung der Messwerte **y** von den vorhergesagten Werten $\hat{\mathbf{y}}$ an:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2 = Var(\hat{y}) + (Bias(\hat{y}))^2 + Var(\epsilon) \tag{43}$$

Anschaulicher als der MSEP ist dessen Quadratwurzel (Root Mean Squarred Error of Prediction (RMSEP)). Auf diese Weise erhält der Vorhersagefehler die gleiche Dimension wie die Zielgröße y:

$$RMSEP = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{44}$$

Zur Ermittlung des MSEP wird der Datensatz zunächst in einen Trainings- und einen Testdatensatz aufgeteilt (ca. 2/3 Training, 1/3 Test). Der Trainingsdatensatz wird zur Erstellung des Modells verwendet. Anschließend wird die Qualität der Vorhersage anhand des Testdatensatzes mit Hilfe des RMSEP berechnet. Bei hinreichend großen Datensätzen liefert die beschriebene Vorgehensweise gute Ergebnisse. Häufig ist jedoch die Anzahl der Proben ein limitierender Faktor. Stehen nur wenig Proben zur Erstellung des Modells zur Verfügung, kann das Modell nur sehr ungenau geschätzt werden. Um dieses Problem zu umgehen, kann eine s. g. Kreuzvaldierung durchgeführt werden.

Kreuzvalidierung In der Kreuzvalidierung wird der vorhandene Datensatz zufällig in k Abschnitte aufgeteilt (mit k normalerweise $5 < k \leq 10$). Anschließend wird ein Teildatensatz ausgelassen und das Modell auf dem verbliebenen Datensatz trainiert. Nach Erstellung des Modells kann dessen Güte auf dem ausgelassenen Abschnitt getestet werden. Dieses Verfahren wird so lange wiederholt, bis alle k-Abschnitte jeweils einmal ausgelassen wurden. Der Vorhersagefehler des Modells wird über den Mean Squarred Error of Cross-Validation (MSECV) berechnet:

$$MSECV = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2 \tag{45}$$

Da jeder der k-Abschnitte sowohl als Trainings- als auch als Validierungsdatensatz diente, wurde das Modell nie auf einem vollständig unabhängigen Testdatensatz getestet. Die Vorhersagegüte des Modells wird somit überschätzt (MSECV < MSEP). Ein bessere Einschätzung des Modells erhält man über eine doppelte Kreuzvalidierung, die gegebenenfalls auch mehrfach ausgeführt werden kann [3, 4].

Teil I.

Einleitung

1. Einleitung

1.1. Naturstoffe als Basis für Arzneimittel

Die Natur stellt ein nahezu unerschöpfliches Arsenal von chemischen Verbindungen zur Verfügung, die als Leitsubstanzen für pharmazeutische Produkte dienen können. Der Erfolg von Naturstoffen als Basis für Therapeutika beruht u. a. auf ihrer hohen chemischen Diversität und biochemischen Spezifität als Ergebnis einer evolutionären Entwicklung [5, 6]. Es ist deshalb nicht überraschend, dass etwa 60 % aller 1211 niedermolekularer Substanzen („small molecules"), die zwischen 1981 und 2014 auf dem pharmazeutischen Markt eingeführt wurden, auf Naturstoffen basieren oder von ihnen inspiriert wurden [7]. Auf dem Gebiet der Krebsmedikamente lag der Anteil der von der Natur inspirierten Verbindungen, die zwischen 1940 und 2010 für die Krebstherapie zugelassen wurden, sogar bei 75 % [8]. Weiterhin befanden sich 2013 etwa 100 Naturstoffe oder davon abgeleitete Verbindungen in klinischen Studien [9].

Historisch reicht die gezielte Verwendung von Naturstoffen in Form von Pflanzenextrakten bis in die frühen Hochkulturen in Mesopotamien und Ägypten [10]. Über die Jahrhunderte wurde das Wissen über die Wirkung von Heilpflanzen (Pharmakognosie) mündlich überliefert und von Zeit zu Zeit in verschiedenen Werken schriftlich fixiert [11]. Als herausragende Persönlichkeiten arabisch-europäischer Heilkunst seien an dieser Stelle Galen (Galenos von Pergamon, ca. 129-200), Avicenna (Abū Alī al-Husain ibn Abd Allāh ibn Sīnā, 980-1037) und Paracelsus (Theophrastus Bombast von Hohenheim, 1493-1541) genannt [12].

In der Regel wurden die Arzneidrogen, d. h. Bestandteile von Pflanzen oder Pilzen, getrocknet und entweder direkt oder als Tinktur angewendet. Diese Vorgehensweise ist auch heute noch in der Naturheilmedizin in Gebrauch. Man denke beispielsweise an Extrakte des Johanniskrauts (*Hypericum perforatum*), die zur Behandlung bei depressiven Verstimmungen eingesetzt werden [13]. Eine schlaffördernde Wirkung wird den Baldrian (*Valeriana officinalis*)-, Hopfen (*Humulus lupulus*)-, Passionsblumen (*Passiflora incarnata*)- und Melissenextrakten (*Melissa officinalis*) nachgesagt [14]. Auch die Kamille (*Matricaria chamomilla*) dürfte in Form von Tee in fast jedem deutschen Haushalt zu finden sein [15].

Mit dem Beginn der modernen Chemie im 18. Jahrhundert wurde zunehmend die Bedeutung einzelner Wirkkomponenten in Extrakten erkannt. Bereits 1804 isolierte Sertürner das analgetisch wirkende Morphin (**1**) aus *Papaver somniferum* (Schlafmohn). Bis zur vollkommenen Aufklärung der Struktur und kommerziellen Vermarktung vergingen jedoch noch weitere 70 Jahre. In der Folgezeit wurden eine Reihe weiterer Morphinderivate (Opioide) wie z. B. Fentanyl (**2**) [16, 17], Codein (**3**) und Tramadol (**4**) [18] entwickelt, die bis heute unerlässlich in der Schmerztherapie sind. Ein weiteres Blockbuster Medikament, das seinen Ursprung in einem pflanzlich basierten Naturstoff hat, ist das Biguanid Metformin (**5**). Aufgrund seiner blutglukosesenkenden Wirkung wird Metformin heute als Primärtherapie bei Typ 2 Diabetes verwendet [19]. Doch bereits Ende des 18. Jahrhunderts wurden Extrakte aus *Galega officinalis* (Geißraute) zur Behandlung von Diabetessymptomen eingesetzt [20]. Als Wirkstoff fungiert das in den Blättern enthaltenen Guanidin-Derivat Galegin (**6**) [21]. 1918 erkannte Watanabe bei Versuchen mit Hasen, dass die Injektion von Guanidinhydrochlorid zu Hypoglykämie führt [22]. Das Wissen um die blutglukosesenkende Wirkung der Guanidine geriet in den 1930er und 1940er Jahren in Vergessenheit, bis Jean Sterne 1957 die erste Studie zur Verwendung von Metformin publizierte [23].
Nach der Markteinführung in Großbritannien/Europäische Wirtschaftsgemeinschaft (1958) und in den USA (1995) wurden im Jahre 2014 alleine in den USA über 86 Millionen Einheiten Metformin Hydrochlorid verschrieben [24]. Seit 2011 befindet sich Metformin in der WHO Liste der „essential medicines“ [25] und in jüngster Zeit werden sogar antikanzerogene Effekte von Metformin diskutiert [26].
Ein weiteres Beispiel für einen Wirkstoff, der bereits vor seiner Entdeckung als Heilmittel in Gesamtextrakten Verwendung fand und später Leitsubstanz für eine ganze Klasse von Arzneimitteln wurde, ist die Acetylsalicylsäure (**7**) als Derivat des Salicins (**8**). Das

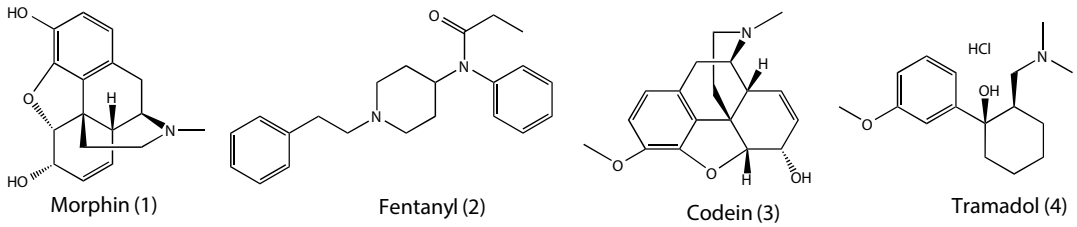
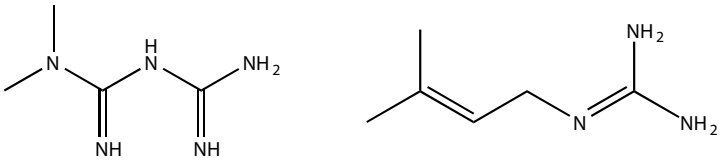


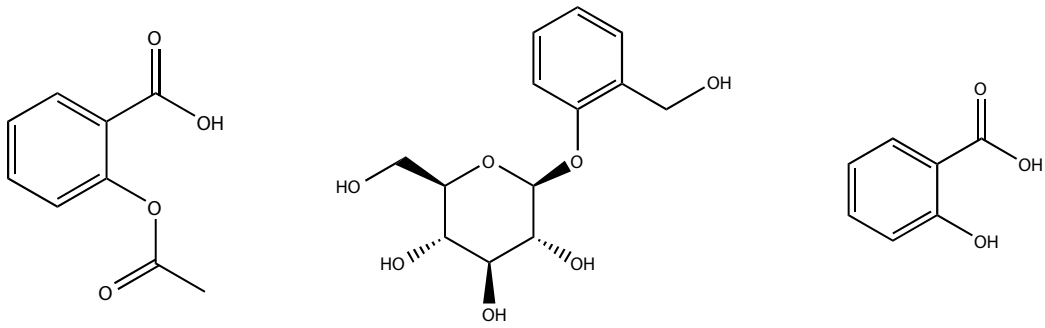
Abbildung 1.1.: Morphin und weitere Opioide



Metformin (5) Galegin (6)

Abbildung 1.2.: Guanidinderivate Metformin (**5**) und Galegin (**6**).

Salicin, das im Körper zum eigentlichen Wirkstoff Salicylsäure (**9**) umgesetzt wird, ist ein Bestandteil der Weidenrinde (*Salix*) und wurde nachweislich bereits 1934 v.Chr. als Analgetikum zur Linderung von Schmerzen verwendet [27]. Nachdem Buchner das Salicin aus Weidenrinde isoliert hatte (1828), gelangen in der Folgezeit die Synthese der Salicylsäure (Piria, 1838) und schließlich der besser verträglichen Acetylsalicylsäure (Gerhardt, 1853). Nach der Markteinführung im Jahre 1897 startete die Acetylsalicylsäure unter dem Handelsnamen Aspirin® einen weltweiten Siegeszug als Antipyretikum und Analgetikum. Für die Aufklärung des Wirkmechanismus der Acetylsalicylsäure (Inhibition der Cyclooxygenase) erhielt John Vane 1982 als einer von drei Laureaten den Nobelpreis für Medizin und Physiologie [28]. Bereits zuvor wurde der Nobelpreis für Medizin und Physiologie an Forscher



Acetylsalicylsäure (7) Salicin (8) Salicylsäure (9)

Abbildung 1.3.: Der Naturstoff Salicin (**8**) aus der Weidenrinde ist die Basis für das Medikament Aspirin® mit dem Inhaltsstoff Acetylsalcylsäure (**7**), die in den eigentlichen Wirkstoff Salicylsäure (**9**) metabolisiert wird.

vergeben, die durch die Entdeckung von potenten Naturstoffen die Grundlage für bahnbrechende neue Therapien legten. Im Jahre 1945 wurden Sir Alexander Fleming, Ernst Boris Chain und Sir Howard Walter Florey „für die Entdeckung von Penicillin (**10**) und seinen heilsamen Effekt bei verschiedenen infektiösen Krankheiten“ geehrt. Für die Entdeckung des Streptomycins (**11**) erhielt sieben Jahre später Selman Abraham Waksman ebenfalls den Nobelpreis für Medizin und Physiologie.

Viele der bekannten Antibiotika sind inzwischen durch leichtfertigen und zu häufigen Einsatz unwirksam geworden. Insbesondere die s. g. Methicillin resistenten *Staphylococcus aureus* Stämme (MRSA) sind zu einer Bedrohung für Menschen geworden [29]. Soll bei dem „Wettstreit“ zwischen Mensch und Mikrobe der Mensch den Kopf vorne haben, ist die Entdeckung von neuen Wirkstoffen und Wirkmechanismen essentiell.

Doch nicht nur zur Bekämpfung bakterieller Infektionen werden neue Medikamente benötigt. Durch die demografische Entwicklung in den Industrienationen steigt die Inzidenz von Alterserkrankungen [30, 31] wie Demenz, Parkinson, rheumatoide Arthritis und Krebs. Beispielsweise schätzt die International Diabetes Federation (IDF), dass die Anzahl der Menschen mit Diabetes von 415 Millionen im Jahre 2015 auf etwa 642 Millionen im Jahre 2040 ansteigen wird [32].

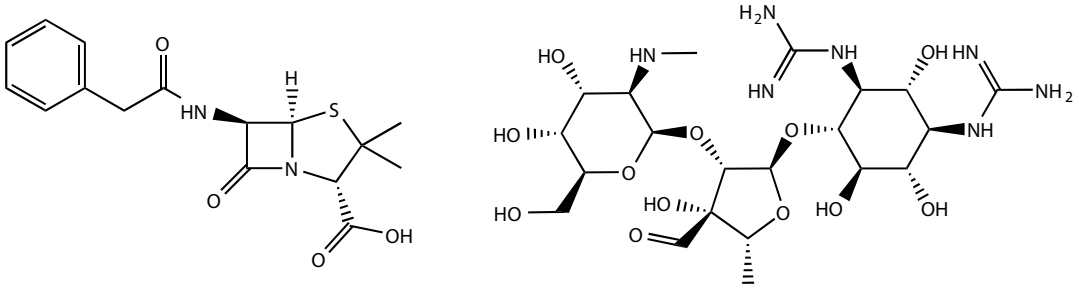


Abbildung 1.4.: Beispiele für Antibiotika: Beta-Lactam Antibiotikum Penicillin (**10**) und das Aminoglykosid Streptomycin (**11**).

1.2. Klassische Methoden zur Identifizierung biologisch aktiver Substanzen

Die Entdeckung und Strukturaufklärung neuer Naturstoffe, die als Leitstrukturen für die Entwicklung neuer Therapeutika dienen können, ist immer noch ein herausfordernder Prozess. Durch einen ethnopharmakologischen Ansatz, d. h. den Einbezug von tradiertem Wissen über Heilmittel (Pflanzen, Pflanzenteile, Pilze etc.), etwa der traditionellen chinesischen Medizin, kann bereits zu Beginn einer Studie eine Präselektion von interessanten Ausgangsmaterialien getroffen werden [33, 11].
Doch Rohextrakte biologischer Matrizes sind komplexe Mischungen aus hunderten oder tausenden Verbindungen mit unterschiedlichen chemo-physikalischen Eigenschaften. Die Suche nach der/den aktive(n) Komponente(n) entspricht der sprichwörtlichen Suche nach der Nadel im Heuhaufen. Grundsätzlich können die Testverfahren, die für das Screening nach neuen Verbindungen eingesetzt werden, in Phänotyp- und Target-basierte Verfahren kategorisiert werden [34].

Phänotyp-basierte Methoden sind dadurch charakterisiert, dass keine Annahmen über molekulare Wirkungsmechanismen vorausgesetzt werden. Die Testung von Extrakten, Fraktionen oder Einzelsubstanzen erfolgt an „lebenden“ Systemen, die einen hohen Bezug zur therapeutischen Relevanz besitzen [35]. Beispielsweise werden Extrakte, Einzelverbindungen oder Substanzbibliotheken bei der Suche nach neuen Antibiotika direkt an entsprechenden Mikroorganismen getestet, ohne dass der molekulare Wirkmechanismus bekannt sein muss. Durch die Testung an Zellkulturen, Gewebekulturen oder Mikroorganismen werden Faktoren, die die Wirkung einer bioaktiven Komponente beeinflussen können, in das Screening mit einbezogen [36]. So verfügen beispielsweise Bakterien mit ihrer Zellwand über einen wirkungsvollen Schutz vor Xenobiotika [37]. Noch effektiver sind Effluxtransporter, die Xenobiotika aus den Zellen hinaus transportieren (ABC-Transporter [38]). Diese spielen eine wichtige Rolle bei der Entwicklung von Resistenzen gegen Antibiotika [39, 40] und Chemotherapeutika [41].
Da Phänotyp-basierte Methoden hypothesenfrei arbeiten, sind sie ideal dazu geeignet, Medikamente mit ganz neuen Wirkmechanismen zu entdecken (s. g. First-in-Class Medikamente). Tatsächlich wurden 12 von 14 Naturstoff-basierten First-in-Class Medikamente, die zwischen 1999 und 2013 von der amerikanischen Behörde für Lebens- und Arzneimittel (FDA) zugelassen wurden, durch Phänotyp-basierte Assays entdeckt. Demgegenüber wurden lediglich zwei Naturstoff-basierte First-in-Class Medikamente durch Target-basierte Me-

thoden entdeckt [42]. Ein Nachteil der Phänotyp-basierten Methoden ist der meist deutlich geringere Probendurchsatz im Vergleich zu den Target-basierten Methoden [34].

Target-basierte Methoden entsprechen einer Hypothesen-geleiteten Suche nach neuen Wirkstoffen. Dabei wird das Wissen über molekulare Wirkmechanismen genutzt, um gezielt nach Verbindungen zu suchen, die definierte Targets beeinflussen. Targets sind in vielen Fällen Rezeptoren und Enzyme, die bei Signaltransduktionsprozessen eine wichtige Rolle spielen [43, 44, 34]. Beispielsweise sind G-Protein gekoppelte Rezeptoren und Kinasen in die Entstehung von Krebs und vielen weiteren Erkrankungen involviert [45, 46, 47]. In ihrer Übersichtsarbeit beschreiben Eder *et al.* G-Protein-gekoppelte Rezeptoren, Kinasen, Proteasen und Ionenkanäle als häufigste molekulare Targets von First-In-Class Medikamenten [42].
Ein wesentlicher Vorteil der Target-basierten Methoden ist die enorme Menge an Substanzen, die innerhalb kurzer Zeit getestet werden können. Der Einsatz von Hochdurchsatz Plattformen erlaubt das Screening von bis zu 100.000 Substanzen pro Tag [48, 44, 49]. Durch den reduktionistischen Ansatz werden jedoch wichtige Aspekte wie Bioverfügbarkeit und Pharmakokinetik ausgeblendet. So besteht eines der größten Probleme beim Target-basierten Antibiotika-Screening darin, dass viele der *in vitro* positiv getesteten Substanzen die bakterielle Zellwand nicht durchdringen können und somit für die weitere Entwicklung nicht in Betracht kommen [50].

High-throughput Screening (HTS) In der pharmazeutischen Industrie werden primär HTS-Methoden zur Entdeckung neuer bioaktiver Komponenten verwendet. Durch voranschreitende Miniaturisierung (bis zu 3456 Kavitäten pro Mikrotiterplatte) und automatisierte Verfahren können etwa 100.000 Substanzen pro Tag in zumeist Target-basierten Assays getestet werden [49]. Aus Sicht der Informatik entspricht das HTS also einer Brute-Force Methode.
Zu Beginn der 1990er Jahre bestanden die verwendeten Substanzbibliotheken aus ca. 10.000 bis 100.000 vor Ort verfügbaren Chemikalien; getestet wurden u. a. auch Farbstoffe und Feinchemikalien [51]. Die physiko-chemischen Eigenschaften viele dieser Verbindungen (z. B. Löslichkeit, Toxizität etc.) waren jedoch wenig wirkstoffähnlich, sodass der Erfolg des HTS für die Entwicklung neuer Medikamente zunächst ausblieb. Initial brachte auch der Einsatz der Kombinatorischen Chemie, u. a. wegen mangelnder Diversität und Wirkstoffähnlichkeit, nicht den gewünschten Erfolg. [52, 53, 51].
Inzwischen sind Substanzbibliotheken mit einer Größe zwischen 500,000 und 2 Millionen Substanzen im Einsatz [51]. Im Gegensatz zu den 1990er Jahren sind die Bibliotheken auf Wirkstoffähnlichkeit (drug-likeness) und Leitsubstanzähnlichkeit (lead-likeness) optimiert,

sodass Hitraten von über 60 % erzielt werden [51]. Auch dem Problem von Screening-Artefakten (Falsch-Positive Hits) wird durch sekundäre Gegen- bzw. tertiäre Validierungsscreenings wirksam begegnet [44, 54]. Basierend auf [55] wurden zwischen 1996 und 2014 von der FDA insgesamt 29 Medikamente zugelassen, die durch HTS-Methoden entdeckt wurden.
In den letzten Jahren hat dabei der Anteil der Phänotyp-basierten Assays im HTS deutlich zugenommen. Bei Novartis werden inzwischen nahezu 50 % der Screenings mit Phänotyp-basierten Assays durchgeführt [56]. Diese Entwicklung ist u. a. auch durch neuartige Zellkultur Assays getrieben, die durch CRISPR-Cas Editing maßgeschneiderte Krankheitsmodelle simulieren [36].
Problematisch bleibt jedoch die Testung von Rohextrakten. Zum einen geben einige häufig auftretende Verbindungen wie z. B. Tannine Falsch-Positive Signale in Screenings (s. g. „Dark Chemical Matter“) [57, 58, 59]. Zum anderen ist die Identifizierung und Isolierung einer bioaktiven Substanz aus einem positiv getesteten Extrakt sehr zeit- und arbeitsintensiv [57].

Klassische Naturstoffisolierung In der akademischen Forschung wird vielfach noch die klassische Naturstoffisolierung praktiziert. D. h. aus einem biologischen Rohmaterial (Pflanzenteile, Pilz etc.) wird zunächst ein Rohextrakt hergestellt. Aus diesem Rohextrakt werden über eine Vielzahl weiterer Aufreinigungsschritte einzelne Substanzen isoliert [60, 61, 62, 63]. Anschließend erfolgen Strukturaufklärung, Identifizierung und Testung der isolierten Verbindungen in einem Bioassay.
Der grundlegende Nachteil dieser Methode besteht darin, dass sehr viel Zeit (und Geld) in die Isolierung einzelner Substanzen investiert wird, die unter Umständen keine relevante biologische Aktivität besitzen oder aber bereits als biologisch aktive Komponenten bekannt sind (Dereplikation).

Bioaktivitäts-geleitete Fraktionierung Um möglichst nur die aktiven Komponenten zu isolieren, hat sich die s. g. Aktivitäts-geleitete Fraktionierung (engl. Bioactivity Guided Isolation) etabliert [64]. Bei diesem Verfahren werden Aliquots der Fraktionen, die bei jedem Aufreinigungsschritt gesammelt werden, auf ihre biologische Aktivität untersucht. Positiv getestete Fraktionen werden vereinigt und einem weiteren Reinigungsschritt unterzogen. Auf diese Weise wird sichergestellt, dass tatsächlich nur die aktiven Komponenten isoliert und identifiziert werden.
Doch auch dieses Verfahren ist mit Problemen behaftet. Da bei jedem Aufreinigungschritt

Substanzverluste auftreten, kann die messbare Bioaktivität gerade bei niedrig abundanten Metaboliten unter die Nachweisgrenze sinken, sodass die bioaktive Komponente nicht identifiziert werden kann. Zudem kann die biologische Aktivität niedrig konzentrierter Metaboliten auch irrtümlich einer höher konzentrierten Verbindung in der gleichen Fraktion zugeschrieben werden. So berichten Beutler *et al.*, dass sie die anti-HIV Aktivität in Fraktionen aus *Maprounea africana* zunächst einigen Triterpenen zuschrieben. Nach der vollständigen Isolierung zeigten diese jedoch keine Bioaktivität. Erst später bemerkten Beutler *et al.*, dass die zuvor gemessen Bioaktivität von einigen sehr niedrig konzentrierten Phorbolestern ausging, die jedoch unter dem Detektionslimit des verwendeten NMR-Spektrometers lagen [65].
Nicht selten geht die Gesamtaktivität eines Rohextraktes von mehreren additiv oder synergistisch wirkenden Verbindungen mit eher geringen Einzelaktivitäten aus [66, 67]. Auch in diesem Falle würde die Aktivität in den Fraktionen sinken, da die Substanzen im Laufe der Aufreinigung vereinzelt werden [68, 11].
Ein weiterer Nachteil der Aktivitäts-geleiteten Fraktionierung ist die, je nach verwendetem Bioassay, zeitaufwendige Testung der Extraktfraktionen, die in jedem Aufarbeitungsschritt entstehen. Beispielsweise kann für Zellkultur- und Mikroorganismen-basierte Assays inklusive Anzucht und Testung ca. eine Woche Arbeitszeit veranschlagt werden, während der keine weitere Aufreinigung möglich ist.
Wie auch bei der klassischen Naturstoffisolierung ist die vollständige Strukturaufklärung und damit auch Identifizierung erst nach der Isolierung der Substanz möglich. Mithilfe von hochauflösenden LC-MS und NMR-spektroskopischen Messungen kann aber schon zu einem früheren Zeitpunkt in dem Aufreinigungsprozess mit der Dereplikation begonnen werden [69] (s. Abschnitt Dereplikation). Die Dereplikation selbst ist jedoch ebenfalls ein zeitaufwendiger Prozess, der zudem auch nicht den direkten Bezug zur Bioaktivität herstellt, sondern lediglich einzelne Substanzen identifiziert.

Bioautographie und Bioassay-coupled Chromatography (BCD) Bioautographie und BCD stellen Phänotyp- und Target-basierte Screening Methoden dar, die auch für den akademischen Einsatz geeignet sind. In beiden Methoden werden Rohextrakte zunächst chromatographisch aufgetrennt. Die separierten Substanzen werden anschließend über entsprechende Reportersysteme (z. B. Enzym-Assay, Rezeptor-Assay) auf ihre Bioaktivität getestet und im Idealfall mit einem Massenspektrometer identifiziert [70].

Bei Thin Layer Chromatography (TLC) oder High-Performance Thin-Layer Chromatography (HPTLC) werden Enzyme oder andere ausgewählte molekulare Targets auf einer Dünn-

schichtplatte immobilisiert [71]. Findet nach chromatographischer Trennung eine Reaktion mit dem Target statt, kann die Substanz z. B. mit einem kolorimetrischen Verfahren detektiert und nach Abrasion der entsprechenden Stelle mithilfe eines Massenspektrometers identifiziert werden. Für diese Methode sind in [71] eine Reihe von Reportersystemen beschrieben. U. a. erlaubt ein von Gu *et al.* entwickelter TLC-basierter Assay die Suche nach Dipeptidyl Peptidase 4 Inhibitoren (DPP-4) [72]. DPP-4 Inhibitoren sind wertvolle Therapeutika bei Typ 2 Diabetes, da sie die Regulation des Blutglukosespiegels über die Deaktivierung von Inkretinhormonen ermöglichen [73].
In einem Phänotyp-basierten Ansatz (direkte Bioautographie) können antifungale oder antibakterielle Wirkstoffe durch Eintauchen oder Einsprühen einer entwickelten TLC-Platte mit einer Pilz- bzw. Bakterienlösung durch die Ausbildung von Hemmhöfen detektiert werden [74]. Die Anwendung der direkten Bioautographie ist jedoch auf eine Reihe von Organismen beschränkt, die auf TLC-Platten wachsen können [75].

LC-BCD oder LC-BCD-MS erlauben eine On-line Kopplung von Chromatographie, Assay- und Detektionssytem. Dabei werden Rohextrakte zunächst über einer HPLC-Säule aufgetrennt. Das Eluat wird über einen Splitter parallel zu einem Massenspektrometer und zu einer Reaktionszelle geleitet, die das Reportersystem enthält. Auf diese Weise kann eine zeitliche Korrelation zwischen positiver Reaktion (z. B. Bindung an einen Rezeptor oder Inhibition eines Enzyms) und einem Massensignal hergestellt werden [76]. Mit einem solchen System gelang van Elswijk *et al.* ein automatisiertes Screening nach neuen Liganden des β-Estrogenrezeptors [77]. In einer anderen Variante befindet sich das Reportersystem seriell zwischen HPLC und Massenspektrometer [78]. De Boer *et al.* nutzten eine Serienschaltung zur Untersuchung von Naturstoffextrakten in Hinblick auf Inhibitoren von Cathepsin B, einem Enzym, das bei der Genese verschiedener Tumorvarianten involviert ist [79]
Ein limitierender Faktor für den Einsatz von Bioautographie und BCD-Methoden ist die Inkompatibilität von Enzymen und Rezeptoren bei der Verwendung in der Chromatographie üblicher organischer Lösungsmittel [80, 81].

Dereplikation Unabhängig von der Art der Screening Methode ist die s. g. Dereplikation, d. h. die frühzeitige Identifizierung bereits bekannter Substanzen, eine der Hauptengpässe bei der Suche nach neuen bioaktiven Verbindungen [82, 83]. Corley und Durley von Monsanto schätzten 1994, dass für jede frühzeitig dereplizierte Substanz ca. 3 Monate Arbeit und 50000 US-Dollar eingespart werden könnten [84]. In einer systematischen Übersichtsarbeit über Dereplikationsstrategien beschreiben Hubert *et al.* fünf verschiedene Konzepte [83] (Tabelle 1.1). Mit der Ausnahme der Strategie DEREP5, die DNA-Sequenzen zur taxononi-

schem Klassifizierung von Antibiotika produzierenden Mikroorganismen verwendet, besteht der gemeinsame Nenner der übrigen vier Verfahren darin, die unbekannten Substanzen in Rohextrakten oder Extraktfraktionen mithilfe von Substanz- oder Spektrendatenbanken zu identifizieren.
Moderne Dereplikationsverfahren basieren hauptsächlich auf LC-MS, LC-MS/MS, GC-MS und LC-NMR Spektroskopie [85, 86]. Die häufig eingesetzte massenspektrometrische Identifizierung von Substanzen ist nicht trivial und fehleranfällig. Prinzipiell ist die Bestimmung einer Summenformel über die Messung der akkuraten Masse möglich [87]. Allerdings steigt die Anzahl der isomeren Verbindungen exponentiell mit dem Masse zu Ladungsverhältnis, sodass auch mit hochauflösenden Massenspektrometern wie FT-ICR-, Orbitrap- oder Q-TOF-Geräten häufig keine eindeutige Identifizierung erreicht werden kann [88]. Die Akquisition von Fragmentspektren und Bestimmung von Retentionszeiten kann eine eindeutige Identifizierung einzelner Extraktkomponenten erleichtern [89]. Der Vergleich mit öffentlichen Spektrenbibliotheken [90] wie METLIN [91] oder ReSpect [92] kann jedoch beeinträchtigt sein, da die Fragmentspektren unterschiedlicher Gerätetypen - in Abhängigkeit des verwendeten Massenanalysators - deutliche Unterschiede aufweisen [93, 94]. Erschwerend kommt hinzu, dass die Intensitäten der Massensignale zusätzlich durch Messparameter wie Fragmentierungsenergie, Declustering Potential, Quellenspannung etc. beinflusst werden können. Durch geschickte Erstellung von Konsenusspektren und elaborierte Bewertungsalgorithmen kann die Trefferquote erhöht werden [95, 96, 97, 98]. Die Hauptfehlerquelle für die Nicht-Identifizierung einer Substanz besteht jedoch immer noch darin, dass sie nicht in den Spektraldatenbanken verzeichnet ist [89]. In diesem Fall muss die Struktur der unbekannten Substanz aufwendig bestimmt werden. In der CASMI Challenge 2016, bei der anhand von MS/MS-Spektren die zugrunde liegenden Molekülstrukturen ermittelt werden mussten, benötigte die drittplatzierte Arbeitsgruppe ca. 5 Stunden Arbeit pro Substanz [99].
Die Dereplikation ist somit ein zeitraubender Prozess, der zudem keinen direkten Bezug zur Bioaktivität herstellt. Interessant wäre daher eine Methode, die bereits in den Rohextrakten die aktivitätsrelevanten Signale identifzieren und die Dereplikation somit auf ein Minimum reduzieren könnte.

Tabelle 1.1.: Dereplikationsstrategien. Tabelle übernommen aus [83].

Ziel	DEREP1 Identifizierung der Hauptkomponenten in Einzelextrakten	DEREP2 Identifizierung bei Aktivitätsgeleiteter Fraktionierung	DEREP3 Metaboliten Profiling von Rohextrakten	DEREP4 Targeted Profiling	DEREP5 Taxonomische Identifizierung
Targeted	Nein	Nein	Nein	Ja	Nein
Biologische Assays	Unabhängig	Systematisch	Unabhängig	Unabhängig	Unabhängig
Proben	Einzelextrakt	Einzelextrakt	Extraktkollektion	Einzelextrakt oder Kollektion	Extraktkollektion
Fraktionierung	Meistens	Ja	Nein	Meistens Nein	Nein
Analyseverfahren	LC-MS, GC-MS, NMR	LC-MS, GC-MS, NMR	LC-MS, GC-MS, NMR	Meistens LC-MS	Gensequenzierung
Chemoinformatische, statistische Methoden	Manchmal	Manchmal	Systematisch	Manchmal	Systematisch
Identifizierung	Metaboliten Datenbank	Metaboliten Datenbank	Metaboliten Datenbank	Metaboliten Datenbank	Gendatenbank

1.3. In silico Methoden zur Vorhersage von Bioaktivitäten

Moderne massenspektrometrische und spektroskopische Methoden erlauben die simultane Analyse einer Vielzahl von Metaboliten. In Analogie zur Genomik, bei der die Analyse der Gesamtheit aller Gene als „*Genomics*“ bezeichnet wird, hat sich der Begriff „*Metabolomics*“ für die Analyse der Gesamtheit aller Metaboliten in einer Probe etabliert [100].
Bei der metabolomischen Messung von Naturstoffextrakten entstehen hochdimensionale Datensätze, für deren Analyse komplexe mathematische Analysemethoden benötigt werden.
Je nach Standpunkt des Betrachters lassen sich diese Methoden der Chemometrie (Chemiker), dem Machine Learning (Datenanalysten) oder ganz allgemein der multivariaten Datenanalyse (Statistiker, Metabolomics Community) zuordnen. Grundsätzlich lassen sich diese Methoden in überwachte und unüberwachte Methoden kategorisieren. Zu letzterer Kategorie gehören beispielsweise die PCA und verschiedene Clusterverfahren (hierarchische Clusteranalyse, K-Means Clustering, Self-Organizing Maps). Die unüberwachten Methoden können dazu genutzt werden, die inhärenten Informationen eines hochdimensionalen Datensatzes innerhalb von nur zwei oder drei Dimensionen abzubilden. Die PCA wird beispielsweise häufig als erster Schritt bei der Analyse von metabolomischen Datensätzen

verwendet. Mit ihrer Hilfe kann untersucht werden, ob ein hochdimensionaler Datensatz einer homogenen Verteilung unterliegt oder sich in Gruppen aufspaltet. Sie eignet sich daher auch zur Analyse von Ausreißerproben.
Im Unterschied zu den unüberwachten Methoden wird bei den überwachten Methoden die Zielvariable y in den Modellierungsprozess miteinbezogen. Generell wird versucht eine abhängige Variable y aus einer Anzahl von unabhängigen Variablen $\mathbf{x}^T$ (Regressoren, Prädiktoren) vorherzusagen. Der Regressand y kann dabei sowohl eine kategoriale (z. B. Mutante vs. Wildtyp, Krank vs. Gesund, mehrkategorielle Variablen) als auch eine kontinuierliche Variable sein. In ersterem Fall spricht man von Diskriminanzanalyse; ist y kontinuierlich, handelt es sich um eine Regressionsanalyse.
Zu den klassischen überwachten Methoden zählen sowohl die multiple lineare Regression (MLR) als auch mathematisch komplexere Algorithmen wie beispielsweise die Partial-Least-Squares Regression (PLSR) [101, 102, 103, 104], die Partial-Least-Squares-Diskriminanzanalyse (PLS-DA) und die lineare Diskriminanzanalyse (LDA). In den vergangenen Jahrzehnten kamen, u. a. ermöglicht durch immer leistungsfähigere Computer, weitere Verfahren hinzu. Als Beispiele seien hier Support Vector Machines (SVM) [105, 106], Entscheidungsbaum-basierte Verfahren wie CART und Random Forest [107], und durch die Biologie inspirierte Methoden (künstliche neuronale Netze/Deep Learning, Genetische Algorithmen [108, 109]) genannt.
Im Gegensatz zu den univariaten Methoden, die jeweils nur eine einzelne Variable betrachten, werden bei den multivariaten Methoden während der Datenmodellierung die Abhängigkeiten aller x-Variablen untereinander berücksichtigt [110, 111]. Sie bieten daher ideale Voraussetzungen für die Analyse metabolomischer Datensätze und werden für die unterschiedlichsten Fragestellungen u. a. in der (Pflanzen)biochemie [112], Naturstoffchemie [113], klinischen Chemie und den Ernährungswissenschaften [114, 115] verwendet [110, 116, 117, 118, 119, 120].

Bereits seit den 1980er Jahren nutzen Chemiker insbesondere die klassischen Methoden zum Zwecke der Kalibration und zur Modellierung von Struktur-Wirkungs-Beziehungen (Quantitative Structure Activity Relationship, QSAR). Die QSAR hat zum Ziel, die biologische Aktivität eines Moleküls aus einer Reihe von molekularen Deskriptoren wie beispielsweise physikochemische Faktoren (Lipophilie, Polarisierbarkeit...), topologische Eigenschaften (Diederwinkel) und quantenchemische Deskriptoren (HOMO/LUMO Energien) vorherzusagen [121, 122]. Neben der PLSR [123] haben sich in jüngerer Zeit auch hier moderne Methoden wie Random Forest [124, 125, 126], SVM [127] und penalisierte Regressionsmethoden wie Lasso [128, 129] etabliert. Einige Vergleichsstudien zu multivariaten Methoden

in der QSAR findet man in [122, 130, 131, 132, 133, 126].

In Anlehnung an QSAR-Analysen entwickelten Wang und Cheng 2006 ein QCAR (Quantitative Composition-Activity Relationship) genanntes Verfahren [134], mit dessen Hilfe das ideale Mischungsverhätltnis verschiedener traditioneller chinesischer Heilpflanzen unter Maximierung des biologischen Effekts vorhergesagt werden kann. Um auch mögliche synergistische Effekte erfassen zu können, verwendeten sie nicht-parametrische Methoden wie künstliche neuronale Netze und SVM Regression und erzielten damit tatsächlich bessere Ergebnisse als mit klassischer MLR [135].
Da für diese Verfahren fortgeschrittene Kenntnisse in Programmierung und Datenmodellierung notwendig sind, haben sich in den 2000er Jahren jedoch zunächst die in Software Paketen wie SIMCA (Umetrics Inc, Umeå) und SPSS zugänglichen multivariaten Methoden wie PCA, PCR und PLSR zur Vorhersage der Bioaktivität in Naturstoffextrakten durchgesetzt [136, 137, 138, 139]. Beispielsweise nutzten Cardoso-Tarketa *et al.* PCA und PLS-DA zur Diskriminierung von aktiven und inaktiven Extrakten aus *Galphimia glauca* und konnten auf diese Weise anxiolytisch und sedativ wirkende Metaboliten (Galphimine) identifizieren [137].
Eine umfassende Untersuchung verschiedener multivariater Regressionsmethoden zur Vorhersage der antioxidativen Eigenschaften von grünem Tee aus chromatographischen Fingerprints machten Dumrey et al. [138]. Ähnlich wie Rajalahti [140] kommen sie zum Schluss, dass MLR und PCA bzw. PCR für diese Erfordernisse relativ ungeeignet sind. Erstere, da diese nur gut funktioniert, wenn die Anzahl der vorhandenen Peaks in einem Spektrum die Anzahl der Proben nur unwesentlich übersteigt. Dies ist bei Spektraldaten aus Naturstoffextrakten praktisch nie der Fall. Die PCA bzw. PCR leiden an dem Problem, dass der zugrunde liegende Algorithmus nach den Variablen mit der größten Varianz sucht. Dadurch werden automatisch größere Peaks im Spektrum bevorzugt. Vor allem aber steht die Größe der Varianz nicht unmittelbar in Bezug zur gesuchten Bioaktivität.
Bessere Ergebnisse liefern die PLS und hier insbesondere die orthogonalen PLS (OPLS) Methoden [141, 138, 142]. Beide Algorithmen maximieren die Hauptkomponenten nicht nur in Bezug auf die Varianz in $\mathbf{X}$, sondern auch in Hinblick auf die Kovarianz zwischen den Spektraldaten und der Zielvariablen (hier also der biologischen Aktivität). Die OPLS hat hier den entscheidenden Vorteil, dass die zur Bioaktivität unkorrelierte (orthogonale) Varianz in $\mathbf{X}$ entfernt wird. Dies führt zu Modellen mit geringerer Komplexität und besserer Interpretierbarkeit.
Durch die Verfügbarkeit der Partial-Least-Squares Methoden in SIMCA und anderen automatisierten Softwarelösungen [143, 144] sind PLSR, PLS-DA, OPLSR und OPLS-DA

inzwischen zu Standardmethoden in Metabolomics Analysen geworden [145]. Das SIMCA-Paket wurde jedoch auch erfolgreich zur Identifizierung von bioaktivitäts-relevanten Metaboliten in Naturstoffextrakten eingesetzt. Dazu nutzten verschiedene Arbeitsgruppen zunächst die OPLS-DA, um den Bezug zwischen der Bioaktivität und NMR- [146, 147], MS- [148, 149, 150] oder HPLC-Spektren [142] von Rohextrakten bzw. Extraktfraktionen herzustellen. In einem zweiten Schritt wurden die aktivätsrelevanten Metaboliten entweder mithilfe der Variable Importance in Projection (VIP) Methode [151] oder mittels des s. g. S-Plots [152] selektiert. Beide Methoden ermitteln den Einfluss der einzelnen Variablen auf das OPLS-DA-Modell.

Der OPLS recht ähnlich ist die von Kvalheim eingeführte „Target Projection“ (TP) [153, 154, 155, 156, 157]. Analog zur OPLS extrahiert auch die TP die für die Bioaktivität relevante Variation in $\mathbf{X}$ auf eine einzelne latente Variable [156]. Als Ausgangpunkt zur Berechnung der Target Projected Scores und Loadings dient hier der maximal mit der vorhergesagten Bioaktivität assoziierte Regressionskoeffizientenvektor einer PLS Regression. Doch sowohl bei der OPLS als auch der TP tritt das Problem auf, dass Substanzen in hoher Konzentration aber geringer Kovarianz zur biologischen Aktivität gegenüber Substanzen mit geringer Konzentration aber hohen Kovarianzwerten bevorzugt werden. Um dieses Problem zu umgehen, führte Rajalahti den s. g. „Selectivity Ratio Plot“ ein (SR-Plot), der als Maß für die Signifikanz einer gefundenen Kovarianz angesehen werden kann [158, 159]. Unter dem Namen „Quantitative Pattern Activity Relationship“ (QPAR) führten Chau und Kvalheim die Kombination aus TP und SR-Plot zur Identifizierung von bioaktiven Substanzen in Naturstoffextrakten ein [160]. In einem Proof-Of-Principle Experiment konnten mit dieser Methode verschiedene Substanzen sogar in der Reihenfolge ihrer antioxidativen Kapazität in einer Mischung richtig prognostiziert werden [161].
Mit einer als Biochemometrics bezeichneten Methode, die eine Variation zu QPAR darstellt, identifizierten Kellog *et al.* die antibiotisch wirkenden Verbindungen in Rohextrakten der endophytischen Pilzgattungen *Pyrenochaeta* und *Alternaria* [162]. Dazu regressierten sie mithilfe einer PLSR die Bioaktivitätsdaten (Wachstumsinhibtion von *Staphylococcus aureus*) auf die entsprechenden UHPLC-HRMS-Spektren der Pilzextrakte. In einem zweiten Schritt nutzten sie den SR-Plot zur Selektion der aktivitätsrelevanten Metaboliten (Alternariol monomethyl ether, Altersetin, Macrosphelide A), die anschließend isoliert und positiv auf ihre Wachstumsinhibition von *Staphylococcus aureus* getestet wurden.
In einer 2017 publizierten Studie verwendeten Britton *et al.* die Biochemometrics Methode (PLSR + SR-Plot) um synergistisch wirkende Metaboliten in Rohextrakten von *Hydrastis canadensis* (Kanadische Orangenwurzel) zu identifizieren und während der anschließenden

aktivitäts-geleiteten Fraktionierung „verfolgen“ zu können [163].

In den zuvor beschriebenen Studien hat die Verwendung der klassischen multivariaten Methoden die Identifizierung der aktivitäts-relevanten Metaboliten ermöglicht (Tabelle 1.2). Doch die Modellierung hochdimensionaler Datensätze birgt auch einige Gefahren. Der hohe Grad an Multikollinearität kann zu instabilen Schätzungen der Regressionskoeffizienten führen [165, 2]. Die mit der Zielvariable am stärksten assoziierten x-Variablen können somit durch Störvariablen („Confounder“) verschleiert werden [166]. In den vergangenen Jahren wurden einige penalisierte Regressionsverfahren wie z. B. Lasso (Least Absolute Shrinkage and Selection Operator) [167] und Elastic Net entwickelt, die auch ein höheres Maß an Multikollinearität tolerieren können [168, 169]. Sowohl Lasso als auch Elastic Net verfügen zudem über eine intrinsische Variablenselektion und wurden zu diesem Zweck u. a. bereits zur Analyse von hochdimensionalen Microarray [170, 171, 172, 173], Targeted-Metabolomics- [174], Imaging-MS [175] und QSAR-Datensätzen [132, 131, 129] verwendet. Die Verwendung dieser s. g. Regularisierungsmethoden zur Identifizierung bioaktiver Komponenten in Extrakten ist allerdings bislang noch nicht durch Publikationen dokumentiert.

Die hohe Multikollinearität in massenspektrometrischen und spektroskopischen Datensätzen birgt noch eine zweite Gefahr. So sind insbesondere die Partial-Least-Squares Methoden anfällig für Überanpassung [176, 177, 178], d. h. das Modell wird zu stark an den Trainings-

Tabelle 1.2.: Studien zur Prädiktion der Bioaktivität und Selektion von aktivitätsrelevanten Metaboliten in Naturstoffextrakten.

Methode	Multivariates Verfahren	Variablenselektion	Analytik	Ziel	Literatur
QCAR	ANN, SVM	-	-	Prädiktion	[135]
Metabolic Profiling	PCA, PLS-DA	Loadings	NMR	Prädiktion, Variablenselektion	[137, 136]
Metabolic Profiling	MLR, PCR, PLSR, OPLSR	Regressions-koeffizienten	HPLC	Prädiktion	[138, 139]
Metabolic Profiling	PCA, OPLSR	Regressions-koeffizienten	HPLC, LC-MS	Variablenselektion	[142, 164]
Metabolic Profiling	OPLS-DA	VIP	NMR	Variablenselektion	[146, 147]
Metabolic Profiling	OPLS-DA	S-Plot	LC-MS, GC-MS,NMR	Variablenselektion	[148, 149, 150]
QPAR	Target Projection	SR-Plot	HPLC, MALDI-TOF-MS	Variablenselektion	[160, 161, 158]
Biochemometrics	PLSR	SR-Plot	LC-MS	Variablenselektion	[162, 163]

datensatz angepasst. Durch die hohe Anzahl an x-Variablen lässt sich im Extremfall auch eine zufällige Variablenkombination finden, die y in vermeintlich geeigneter Weise vorhersagt [179, 180]. Tatsächlich sind überangepasste Modelle nur schlecht generalisierbar und können somit zu Falsch-Positiven Ergebnissen führen.
Für die Datenverarbeitung und die Durchführung der chemoinformatischen Analysen benötigt der Experimentator somit sehr gute Kenntnisse auf dem Gebiet der multivariaten Statistik. Um keine fehlerhaften Ergebnisse zu produzieren, muss er mit den Tücken, die bei der Modellierung auftreten können, vertraut sein [181]. In der Realität fehlen den meisten Naturstoffchemikern, die täglich mit der Isolierung und Strukturaufklärung neuer Verbindungen beschäftigt sind, diese speziellen Kenntnisse.
Ein Ziel dieser Arbeit war daher die Entwicklung einer effektiven, aber einfach zu erlernenden Methode, mit deren Hilfe die aktivitätsrelevanten Metaboliten in komplexen Naturstoffextrakten identifiziert werden können.

1.4. Peptaibiotika und Peptaibole

Der Bezeichnung „Peptaibiotika“ ist ein Oberbegriff für eine Klasse von Peptidantibiotika, die im Wesentlichen aus den Peptaibolen *in sensu stricto* und den Lipo- bzw. Lipoaminopepaibolen bestehen [182, 183]. Peptaibiotika sind nicht-ribosomal synthetisierte **Pept**ide, die als charakteristisches Merkmal einen hohen Anteil an α, α-dialkylierten Aminosäuren enthalten. Neben der namensgebenden **a**lpha-Amino**i**so**b**uttersäure (**Aib**) ist dies vor allem Isovalin (**Iva**). Die C-terminale Carboxylgruppe liegt zum Alkoh**ol** reduziert vor. Der *N*-Terminus ist bei Peptaibolen in aller Regel acetyliert. Lipopepaibole sind am *N*-Terminus mit Octadecan-, Decan- oder (Z)-Dec-4-ensäure verestert [184, 185]. Lipoaminopeptaibole tragen an einem *N*-terminalen Prolin oder Hydroxyprolin C_4-C_{15} verzweigte oder unverzweigte Fettsäuren. An Position 2 befindet sich bei allen bekannten Lipoaminopeptaibolen 2-amino-6-hydroxy-4-methyl-8-oxo-decansäure (AHMOD) [183].
In Peptaibiotika treten noch ein Reihe weiterer nicht-proteinogener Aminosäuren auf, die in zwei Arbeiten von Degenkolb übersichtlich dargestellt sind [186, 183].
Alle bislang bekannten Peptaibole bestehen aus 5 bis 22 Aminosäuren und besitzen eine Masse zwischen 500 und 2200 Da. Durch Sequenzalignments konnten Chugh und Wallace die Peptaibole in 9 Subfamilientypen (SF-Typen) einteilen [187]. Seit der Entdeckung des ersten Peptaibols (Alamethicin (**12**)) durch Meyer und Reusser im Jahre 1967 [188] sind mehr als 1000 Peptaibole charakterisiert und in der Comprehensive Peptaibol Database (CPDB) dokumentiert worden [189]. Durch moderne, LC-MS basierte Screening-Methoden (Peptaibiomics) kommen jedes Jahr weitere Peptaibole hinzu [190, 185, 183, 191].
Ein Vergleich zwischen den in der CPDB dokumentierten Peptaibolen und der Pilzdatenbank Mycobank [192, 193] offenbart, dass Peptaibole bislang fast ausschließlich bei Pilzen der Ordnung *Hypocreales* (Krustenkugelpilzartigen) gefunden wurden. In der CPDB sind 1331 Peptaibole aus 34 Gattungen dokumentiert (Stand: August 2017). Mit 989 Peptaibolen (70 %) stammt ein Großteil der aufgeführten Peptaibole aus der Gattung *Trichoderma* (Abb.1.6).

Struktur und Bioaktivität Zahlreiche Strukturuntersuchungen zeigen, dass Peptaibiotika, u. a. durch einen hohen Anteil an Aib, helikale Konformationen annehmen [194, 195, 187, 196, 197, 198]. Weit verbreitet sind die klassische α-Helix, 3_{10}-Helices oder gemischte $\alpha/3_{10}$-Helices [199, 200]. Durch die Kombination aus hydrophoben und hydrophilen Aminosäuren besitzen Peptaibiotika eine amphipathische Natur, die ihnen in Kombination mit einer helikalen Konformation membranaktive Eigenschaften verleiht.

Für die biologische Aktivität der Peptaibiotika (und kationischen antimikrobiellen Peptiden) werden im Wesentlichen drei verschiedene Mechanismen diskutiert, die in ihrer Konsequenz zu einer Desintegration der Zellmembran und somit zur Zerstörung der Zielzelle führen [201].

Alle Modelle beschreiben gemeinsam, dass in niedrigen Konzentrationen die Peptaibolmonomere sich zunächst parallel zur Oberfläche an eine Membran an- bwz. einlagern (Abb. 1.7 a, b). Gemäß dem s. g. „Barrel-Stave“ Modell kommt es nach Überschreiten einer kritischen Konzentration zur Oligomerisierung der Monomere, sodass im Zentrum ein hydrophiler Kanal entsteht, während die Außenseiten mit der Membran interagieren [202]. Auf diese Weise bilden sich fassartige Membrankanäle (Abb. 1.7 c). Beispielsweise zeigen Röntgenstrukturanalyse [203], Rastertunnel- [204] und Rasterkraftmikroskopie [205], dass Alamethicin in Phospholipidmembranen hexamere Kanäle ausbildet [195]. Neben Alamethicin ist der „Barrel-Stave“ Mechanismus für eine Reihe weiterer Peptaibole wie z. B. Trichogin GA IV [206, 207] und Antiamoebin I [208] sehr gut belegt.

Das „Toroidal-Poren“ Modell (auch „wormhole model“) geht davon aus, dass die Phospholipide der Zellmembranen durch die Peptaibole zurückgedrängt werden, sodass die polaren

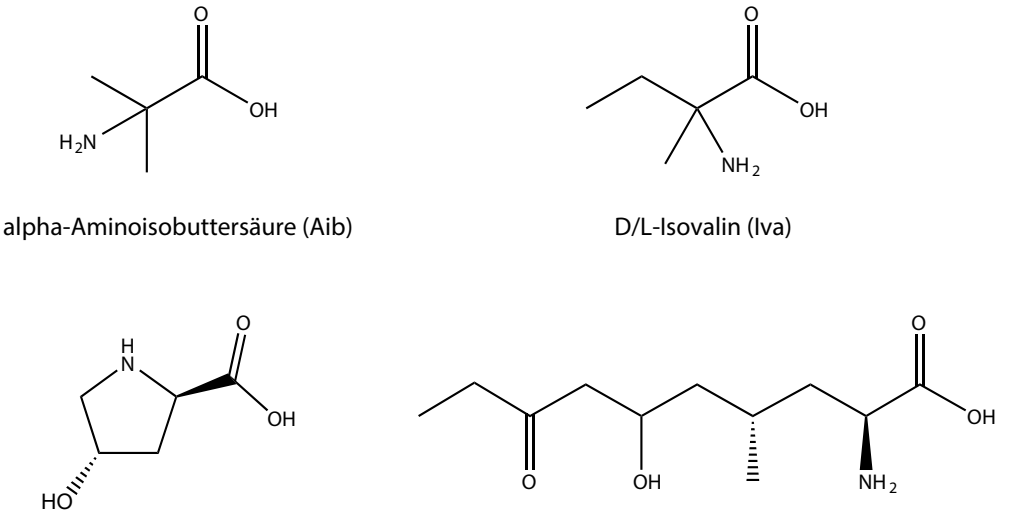


Abbildung 1.5.: Einige nicht-proteinogene Aminosäuren in Peptaibolen.

Kopfgruppen der äußeren und inneren Membran an den Grenzflächen zu den Peptaibolen miteinander in Kontakt treten und die Membrandoppelschicht sich somit in sich selbst zurückfaltet (Abb. 1.7 d) [209]. Typische Vertreter für diesen Mechanismus sind die kationischen Peptide Melittin (Hauptbestandteil des Bienengifts) [210] und Magainin (Hautsekret des afrikanischen Krallenfrosches) [211, 212]. Für Peptaibole ist bisher lediglich für das aus *Trichoderma* sp. isolierte SPF-5506-A_4 eine nach dem „Toroidal-Poren" Modell beschriebene Wirkungsweise nachgewiesen [213].

Als dritter *Modus Operandi* wird ein detergenzartiger Mechanismus diskutiert („Carpet-Modell") [212, 201]. Überschreiten die Peptaibole eine kritische Konzentration, dringen sie in die Zellmembran ein und solubilisieren sie unter Ausbildung mizellarer Strukturen (Abb. 1.7 e, f). Sato *et al.* vermuten sogar, dass alle antimikrobiellen Peptide aufgrund ihres amphiphatischen Charakters in hohen Konzentrationen über einen detergenzartigen Mechanismus agieren [201]. Verschiedene Untersuchungen an Ampullosporin A haben ge-

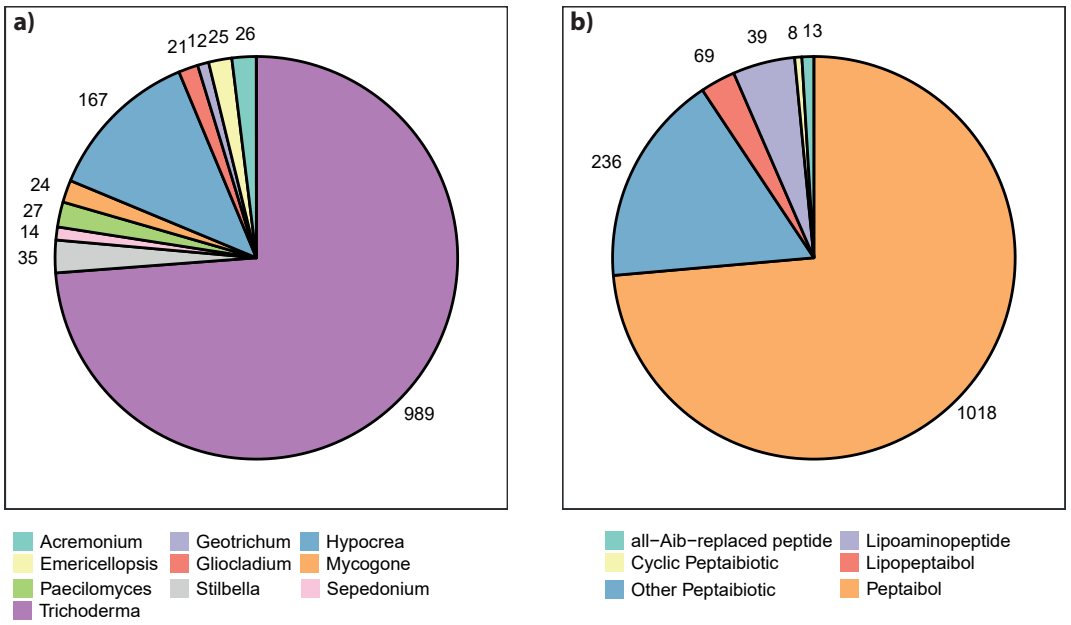


Abbildung 1.6.: Auszug aus der CPDB (Stand August 2017). a) Anzahl der Peptaibole pro Genus. Genera mit weniger als 10 Peptaibolen sind aus Gründen der Übersichtlichkeit nicht berücksichtigt (n = 75 Peptaibole). b) Anzahl der Peptaibole nach Klassen.

zeigt, dass es seine Wirkung sowohl über einen detergenzartigen Effekt als auch über einen Membranporen Mechanismus entfaltet [198, 214, 215].

Angesichts der beschriebenen Eigenschaften ist es nicht verwunderlich, dass Peptaibole eine breites Spektrum an biologischen Reaktivitäten besitzen. Sehr häufig werden antibiotische Aktivitäten, insbesondere gegen Gram-positive Bakterien, beschrieben [216, 217, 218, 219, 220, 221]. Wachstumsinhibitorische Wirkungen gegen (phyto)pathogene Pilze sind für Atroviridine und Neoatroviridine aus *Trichoderma atroviride* sowie für Tylopeptine und Chilenopeptine aus *Sepedonium* spp. dokumentiert [219, 222, 223]. Stadler [224] und Kim *et al.* [225] wiesen für die Microspermine und Chrysospermine antivirale Eigenschaften ge-

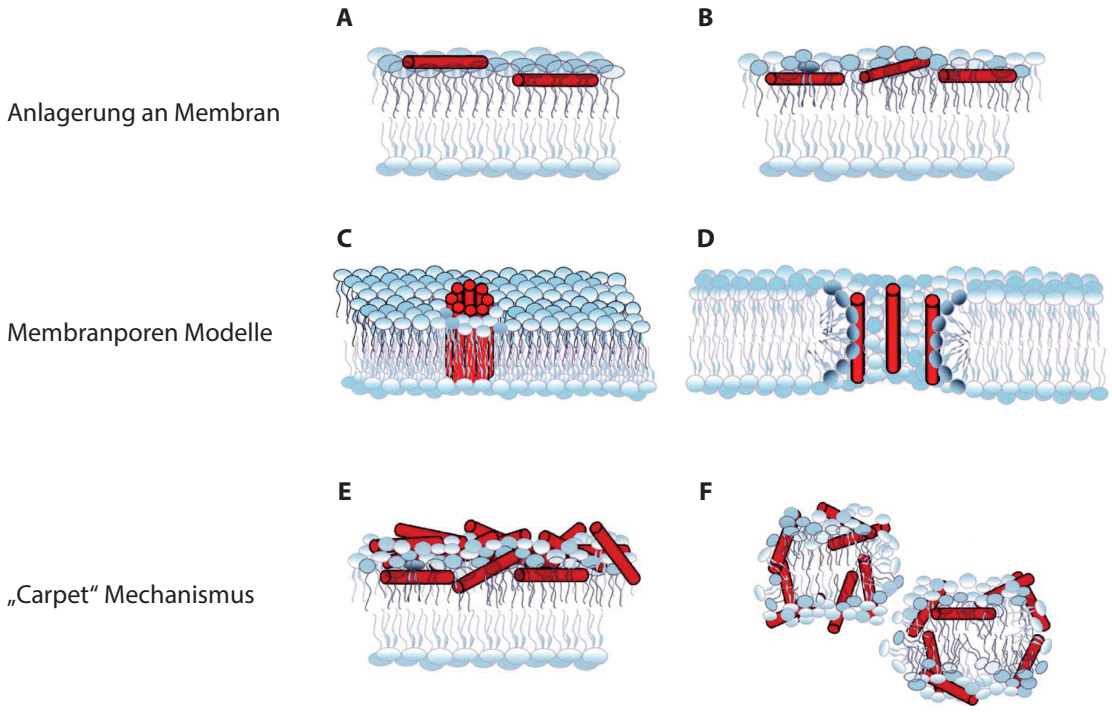


Abbildung 1.7.: Mechanistischer Ablauf der Membrandisruption durch Peptaibole.
Anlagerung der Peptaibole parallel zur Membran (a, b). **Membranporen Modelle**: Wird eine spezifische Peptaibolkonzentration überschritten, bilden sich fassartige („Barrel-Stave“ Mechanismus) (c) oder toroidale („wormhole“ Modell) (d) Membrankanäle. **„Carpet“ Mechanismus**: Sukzessive Erhöhung der Peptaibolkonzentration führt zu einer teppichartigen Schicht (e) und schließlich zur Desintegration der Membran unter Bildung von mizellaren Strukturen (f). Modifiziert nach [201]

gen das HI-Virus bzw. den Tabakmosaikvirus nach. Weiterhin sind für Chyrospermine und Ampullosporine neuroleptikaartige Wirkungen bekannt, die mit denen des Chlorpromazins vergleichbar sind [226, 227, 228, 218].
Diverse Peptaibole aus *Trichoderma* spp. besitzen zytotoxische, antibiotische und antihelmintische Aktivitäten [216]. Neben antihelmintischen Eigenschaften wirken Cephaibole wachstumshemmend auf Mycoplasmen und einige Ektoparasiten [229].
Ebenfalls weit verbreitet sind zytotoxische Aktivitäten gegen eine große Varietät von humanen Zellen [230, 219, 221, 231, 216]. Die wachstumsinhibierende Wirkung ist jedoch nicht alleine auf die Membranporen bildende Aktivität beschränkt. So konnten Du *et al.* zeigen, dass Gichigamin A die äußere Zellmembran passieren kann und intrazellulär die Funktion der Mitochondrien beeinträchtigt [232]. Weithin induziert Trichokonin VI den programmierten Zelltod von hepatozellulären Karzinomazellen über Apoptose und Autophagie Mechnismen [233]. Die Wirkung der Efrapeptine gegen Brustkrebszellen basiert vermutlich auf der Hemmung des Chaperons Hsp90 und der ATP-Synthase in den Mitochondrien [234, 235].

Evolutionsbiologisch werden die Peptaibole vermutlich zu zwei Zwecken genutzt. Zum einen erlauben die membrandisruptiven Eigenschaften eine wirksame Verteidigung gegen Fraßfeinde und Nahrungskonkurrenten. Im Falle von den auf *Boleticus* parasitierenden *Sepedonium* Spezies könnten die Peptaibole auch dazu genutzt werden, die Zellmembran ihres Wirtes zu lysieren um damit die benötigten Nährstoffe freizusetzen [236].

Biosynthese Die Synthese der Peptaibole erfolgt mit Hilfe der s. g. Non-ribosomal Peptidesynthase (NRPS) [237]. Die NRPS sind Multienzymkomplexe ähnlich den Polyketidsynthasen, die jedoch anstatt aktivierter Acyl-Gruppen auf aktivierte Aminosäuren zurückgreifen [238, 239]. Die NRPS sind modular aufgebaut, wobei jedes Modul aus drei Domänen besteht. Die Adenylierungsdomäne (A) erkennt spezifisch eine Aminosäure und aktiviert diese über die Hydrolyse von ATP zum Aminoacyl-Adenylat. Die Thiolierungsdomäne (T) - auch Peptidyl Carrier Protein (PCP) genannt - enthält einen konservierten Serinrest, auf den im Apoenzym zunächst die 4'-Phosphopantethein-Einheit des Koenzym A übertragen wird. Das aktivierte Holoenzym kann anschließend die Aminosäure aus dem Aminoacyl-Adenylat aufnehmen. Die Elongation der Peptaibolkette erfolgt mit Hilfe der Kondensationsdomäne (C), wobei die zuletzt aktivierte Aminosäure mit der naszierenden Kette verknüpft wird (Abbildung 1.8).
Das terminale Modul einer NRPS enthält eine zusätzliche Dehydrogenasedomäne, die für die Reduktion der C-terminalen Carboxylgruppe verantwortlich ist, was schließlich zur Abspaltung des fertigen Oligopeptids führt. In den NRPS sind die einzelnen Module sequen-

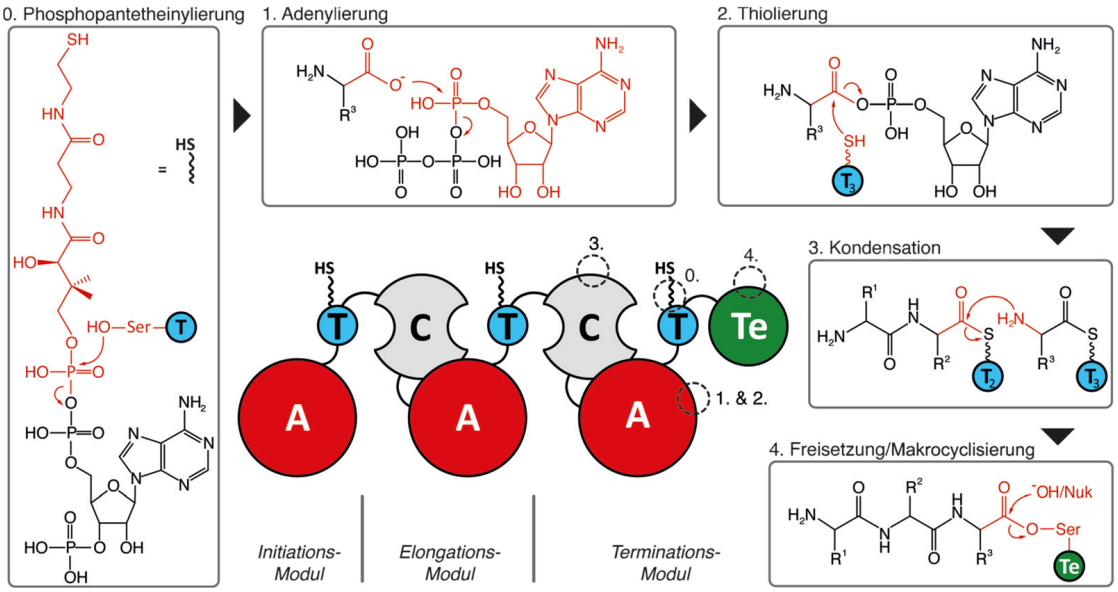


Abbildung 1.8.: Schematischer Ablauf der Biosynthese von Peptaibolen an der NRPS. (0.) Zunächst erfolgt die Aktivierung der Thiolierungs-Domänen (T) über eine 4'-Phosphopantethein Einheit. (1.) Erkennung und Aktivierung einer Aminosäure unter Freisetzung von Pyrophosphat in der Adenylierungs Domäne (A). (2.) Übertragung der aktivierten Aminosäure auf die T-Domäne. (3.) Kondensation von zwei Aminosäuren zwischen T_n und T_{n+1} unter Vermittlung der Kondensationsdomäne (C). (4.) Reduktion des C-Terminus und Freisetzung des Oligopeptids in der Terminationsdomäne. Abbildung übernommen aus [240].

ziell angeordnet, sodass die Abfolge der Module (aufgrund der Substratspezifität der A-Domänen) die Peptaibolsequenz determiniert. Die Sequenzen der Substratbindungsstellen innerhalb der A-Domänen sind phylogenetisch stark konserviert, sodass die Peptaibolsequenz aus einer bekannten Nukleotidsequenz mit einiger Sicherheit vorhergesagt werden kann [241, 242, 243, 244, 245]. In *Trichoderma* spiegelt die Phylogenie der A-Domänen jedoch nicht die Evolution der Spezies wider [244]. In der NRPS ist die Erkennung der Aminosäuren in der A-Domäne an einigen Sequenzpositionen promiskuitiv, d. h. in diesen Modulen werden verschiedene, strukturell ähnliche Aminosäuren in das Peptid eingebaut. Hieraus resultierten mikroheterogene Peptaibolmischungen, die für eine Spezies charakteristisch sind. Häufig findet man an den variablen Positionen Permutationen aus Ala, Aib und Iva, Gln und Glu oder Trp und Phe (Tabelle 1.4). Die variablen Sequenzen werden u. a. von der Zusammensetzung des Nährmediums bestimmt [243]. Neben den variablen

Sequenzpositionen sind andere Positionen sehr konservativ und somit vermutlich essentiell für die biologische Aktivität.
Die Mikroheterogentität der Peptaibole hat evolutionsbiologisch den Vorteil, dass über ein Enzym eine Vielfalt von leicht unterschiedlichen Peptaibolen gebildet werden kann. Dadurch erhöht sich die Wahrscheinlichkeit ein für den jeweiligen Feind wirksames Peptaibol zu erhalten.
In der Regel kann eine NRPS nur eine Klasse von Peptaibolen mit definierter Länge herstellen. Vor allem bei *Trichoderma* sind jedoch auch Fälle bekannt, bei denen Peptaibole mit unterschiedlicher Länge von einer einzigen NRPS synthetisiert werden [245, 243]. So konnte Mukherjee zeigen, dass die Synthese von 11- und 14-Aminosäure Peptaibolen in *Trichoderma virens* von einer einzigen NRPS ausgeht, die aus 14 Modulen besteht. Die Primärsequenzen der daraus gebildeten 11- und 14-Residuen Peptaibole sind mit der Ausnahme der Positionen 4-6 identisch. Da das Tex2 Gen an den relevanten Stellen keine Introns enthält, wurde alternatives Spleißen als möglicher Mechanismus ausgeschlossen. Stattdessen wird angenommen, dass die Biosynthese der 11-Residuen Peptaibole durch s. g. Modul-Skipping verläuft, d. h. bei der Bildung der Peptaibolkette werden einzelne Module übersprungen [245, 243]. Dies wird vermutlich durch die helikale Anordnung der einzelnen Module in der NRPS ermöglicht [246].
Es gibt bislang keine genomischen Analysen zu *Sepedonium* spp., sodass die für die Biosynthese der Peptaibole verwendeten Gene nicht bekannt sind. Reiber hat jedoch aus *S. ampullosporum* zwei Proteine isolieren können, die mutmaßlich für die Synthese von Ampullosporin A (AmpA) verantwortlich sind [237]. Ihrem Modell zufolge, synthetisiert das kleinere Protein HMWP2 ($\approx$ 350 kDa) die ersten beiden *N*-terminalen Aminosäuren inklusive der Acetylierung des Trp^{1}, während das größere Protein HMWP1 ($\approx$ 1,5 MDa) die Oligopeptidkette bis zum terminalen $LeuOH^{15}$ verlängert.

1.5. Die Gattung Sepedonium

Als Testorganismus diente in dieser Arbeit der flaschenporige Goldschimmel *Sepedonium ampullosporum* Damon. *S. ampullospermum* ist eine Art der Gattung *Sepedonium*, einem fungicol lebenden Ascomyceten, der nahezu ausschließlich Pilze aus der Ordnung Boletales (Röhrlinge) befällt. Die Gattung *Sepedonium* wurde 1809 von H.F. Link etabliert [247] und ist als asexuelle Nebenfruchtform (Anamorph) von verschiedenen Arten von *Hypomyces* (Fr.) Tul. & C. Tul. anzusehen. Der Pilz ist kosmopolitisch verbreitet [248], wobei die sexuelle Hauptfruchtform (Teleomorph) in Europa nur selten beobachtet wurde [249, 250]. Die asexuelle Verbreitung von *Sepedonium* erfolgt über zwei oder mehr synanamorphe Formen. Die weißen Phialokonidien dienen der schnellen Verbreitung der Spezies, während die gelblichen Aleuriokonidien bei Resourcenknappheit und ungünstigen klimatischen Bedingungen ausgebildet werden und somit eine Überdauerungsform von *Sepdonium* darstellen. *S. ampullosporum* besitzt als Besonderheit eine dritte Synanamorphe, die durch die namensgebenden flaschenförmigen Konidien charakterisiert ist [251, 249].

In der Internetdatenbank Mycobank sind insgesamt 58 Einträge für *Sepedonium* aufgeführt [192, 193]. Viele dieser Funde stammen jedoch aus dem 19. und frühen 20. Jahrhundert und beruhen allein auf der Beschreibung des Habitus. Zu dieser Zeit war noch nicht bekannt, dass Pilze mit z.T. vollständig unterschiedlicher Morphologie eine Holomorphe, bestehend aus sexueller Haupt- und asexueller Nebenfruchtform, darstellen können. Eine auf Phänotypen beruhende Klassifizierung führte in der Vergangenheit daher immer wieder zu Fehlbestimmungen und Diskussionen unter den Mykologen [252, 253, 251]. Die Schwierigkeiten ergeben sich daraus, dass in der Natur häufig lediglich die Nebenfruchtform gefunden wird. Um verwandtschaftliche Beziehungen zwischen verschiedenen Anamorphen auf morphologischer Ebene zu etablieren, ist es jedoch notwendig die vollständige Holomorphe in die taxonomischen Überlegungen einzubeziehen, da Pilze mit einer morphologisch ähnlichen Nebenfruchtform durchaus gänzlich unterschiedlichen Klassen und Familien entstammen können [253]. Hinzu kommt eine gewisse Variationsbreite in Größe und Farbgebung der Konidien. So argumentiert Damon, die ursprünglich bezeichnete Form *Sepedonium macrosporum* Sacc & Cav sei lediglich als Varietät von *Sepedonium chrysospermum* mit leicht vergrößerten Sporen anzusehen [251]. Auf der anderen Seite etablierten Besl et al. eine ursprünglich als *Sepedonium cf. chrysospermus* bezeichnete Form als *Sepedonium microspermum*, nachdem diese zusammen mit der Hauptfruchtform *Hypomyces microspermum* gefunden wurde [250]. Eine systematische, auf Morphologie basierende Klassifikation von *Hypomyces* Tul. und dessen Anamorphe *Sepedonium* Link wurde 1989 von Rogerson und

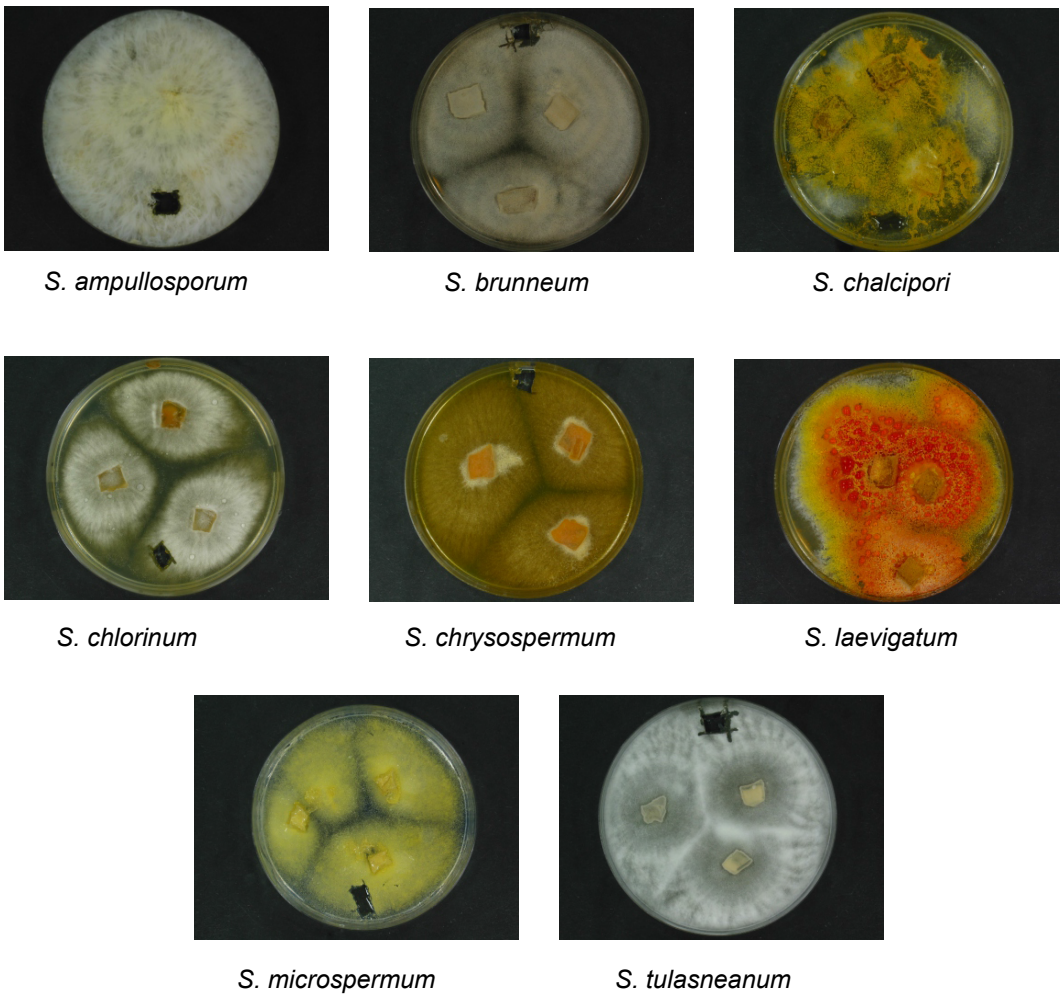


Abbildung 1.9.: Agar-Agar Kulturen der bekannten *Sepedonium* Spezies.

Samuels durchgeführt [248]. In ihrer Arbeit beschreiben sie zehn auf Röhrlingen parasitierende *Hypomyces* Arten. Fünf der *Hypomyces* Arten wurden zusammen mit einer bereits bekannten Nebenfruchtform gefunden. Fünf weitere Fundstücke wurden lediglich als *Sepedonium sp.* bezeichnet. Die Teleomorphe zu *S. ampullosporum* ist bislang noch unbekannt [253, 254].

Eine erste molekularbiologische Klassifizierung auf Basis von 5,8S Ribosomal Ribonucleic Acid (rRNA) inklusive der internen Spaceerregionen ITS1 und ITS2 erfolgte 1989 von Sahr

Tabelle 1.3.: Überblick über die boleticolen Ascomyceten *Hypomyces* und *Sepedonium*.

Teleomorph	**Anamorph**	**Wirt**	Pigmente des Anamorphs
H. completus	S. brunneum	Suillus pictus	**14**, **13**
H. melanocarpus	S. sp.	Tylopilus spp.	
H. transformans	S. sp.	Suillus spp.	
H. badius	S. sp.	Boletaceae	
H. boletiphagus	S. sp.	Boletaceae	
H. chlorinigenus	S. chlorinum	Boletaceae, Agricaceae	
H. melanochlorus	S. sp.	Boletus sp.	
H. microspermus	S. microspermum	Xerocomus porosporus	**17**, **18**
		Xerocomus chrysenteron	**17**, **18**, Anthrachinone
		X. rubellus, C. communis, cf. Chrysenteron	ohne Pigmente
H. chrysospermus	S. chrysospermum	Boletaceae	**17**, **18**
H. tulasneanus	S. tulasneanum	Boletus sp.	unbekannt
?	S. laevigatum		**17**, **18**
?	S. chalcipori	Chalciporus piperatus	**17**, **18**
?	S. ampullosporum	Boletus sp., Chaliporus piperatus?	**17**, **18**, **14**, **13**, **16**

et al. [254]. Anhand dieser Arbeit konnten die fünf bereits von Rogerson und Samuels [253] namentlich aufgeführten Spezies (*S. chlorinum, S. tulasneanum, S. ampullosporum, S. chrysospermum, S. brunneum*) auf molekularbiologischer Ebene bestätigt werden. Die phylogenetischen Daten unterstützen weiterhin die von Rogerson und Samuels als *S. cf. chrysospermum* und von Besl etablierte Spezies *S. microspermum* sowie die von Helfer [255] beschriebene *Sepedonium* Art *S. chalcipori*. Weiterhin lieferten die molekularbiologischen Daten Hinweise auf eine neue Art, die als *S. laevigatum* bezeichnet wurde. In einer jüngeren Arbeit von Otto *et al.* wurde die molekularbiologische Klassifizierung von Sahr bestätigt [222] (Abbildung 1.10).

Als Humanpathogen spielt *Sepedonium* nur eine untergeordnete Rolle. Bislang sind lediglich einige Einzelfälle dokumentiert, bei denen es sich um Sekundärinfektionen immungeschwächter Personen, etwa im Rahmen einer Pneumonie nach Stammzelltransplantation [256] oder AIDS [257], handelt.

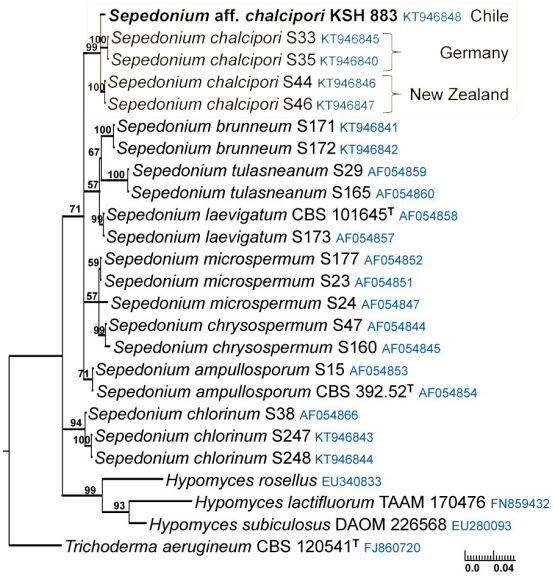


Abbildung 1.10.: Der Stammbaum von *Sepedonium* spp., basierend auf den ITS1 und ITS2 Sequenzen. Modifiziert nach [222].

1.5.1. Sekundärmetaboliten aus Sepedonium

Die ersten Arbeiten zu Sekundärmetaboliten aus *Sepedonium* stammen aus dem Jahr 1957 von Shibata *et al.* [258]. Durch Extraktion des Pilzmyzels von *S. ampullosporum* konnten die Anthrachinonderivate Rugulosin (**13**), Skyrin (**14**) und Chrysophanol (**15**) gewonnen werden. **13** und **14** sind Kondensationsprodukte des Emodins (**16**) und besitzen eine antibiotische Wirkung gegen verschiedene Gram-positive Bakterien, u. a. *Staphylococcus aureus* [259].

1965 isolierten Divekar *et al.* die beiden gelbfarbigen Tropolone Sepedonin (**17**) und Anhydrosepedonin (**18**) aus *Sepedonium chrysospermum* [260]. Während der Anzucht der Pilze entsteht **18** vermutlich artifiziell aus **17** durch pH-Änderungen im Kulturmedium [260]. In der Tat war in einem MRM-basierten Screening Verbindung **17** nach einem mehrwöchigen Wachstum von *Sepedonium* spp. in Malz-Pepton Medium nur noch in Spuren nachweisbar [261]. Stattdessen wurden größere Mengen des Dehydratationsprodukts **18** sowie ein weiteres Sepedonin Artefakt (1-O-methylsepedonin (**19**)) detektiert [261]. Dieselbe Studie zeigt weiterhin, dass diese Verbindungen ausschließlich in den mit gelben Aleuriokonidien ausgestatteten Stämmen *S. ampullosporum*, *S. microspermum*, *S. chalcipori* und *S. chrysospermum* auftritt. In den farblosen Stämmen *S. tulasneanum* und *S. brunneum* waren

weder **18** noch **19** nachweisbar. Dies kann als Hinweis dafür gewertet werden, dass **17** für die gelbliche Farbe der Aleuriokonidien verantwortlich ist. Neben den Tropolonderivaten konnten Quang *et al.* eine neue Verbindung namens Ampullosin (**20**) charakterisieren.

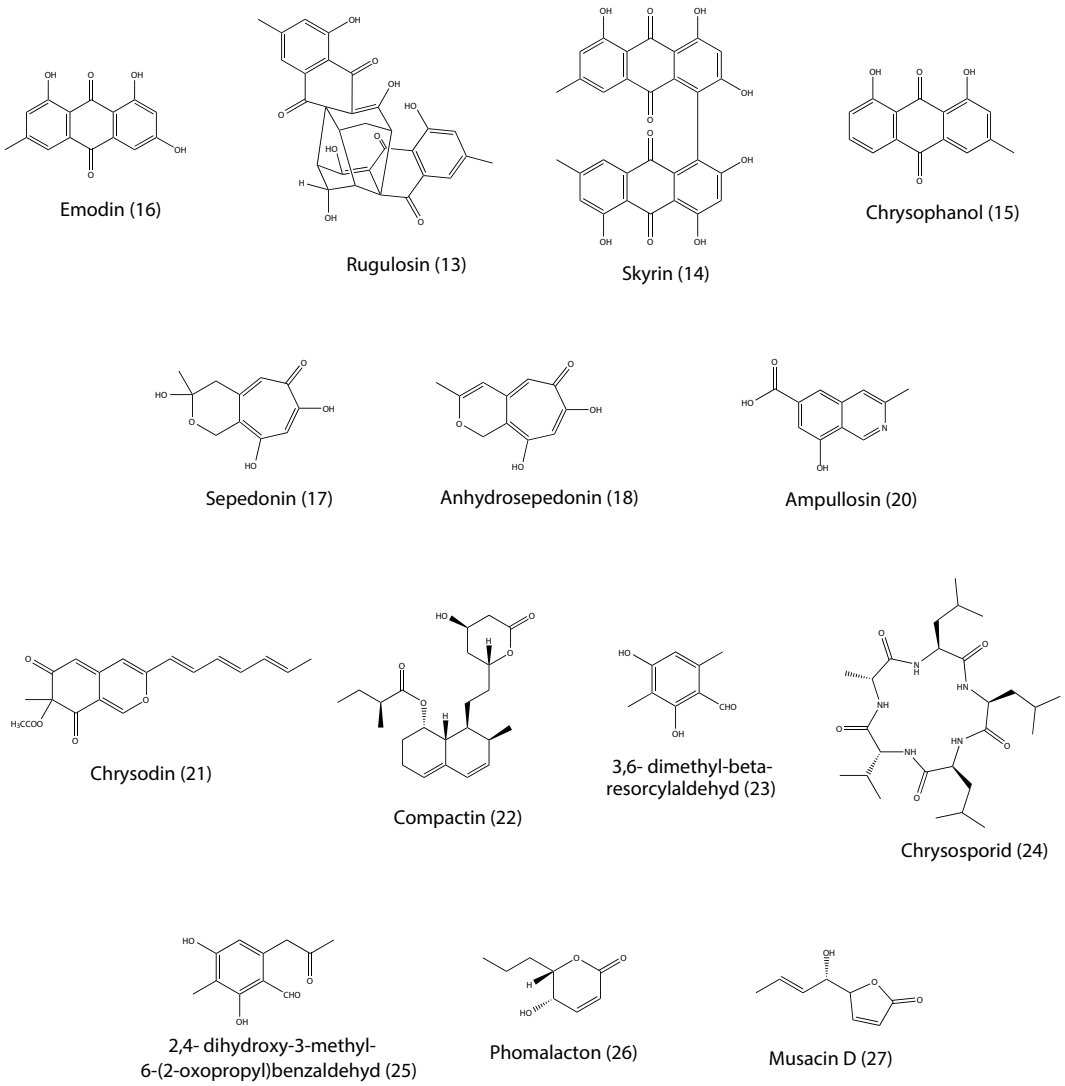


Abbildung 1.11.: Sekundärmetaboliten aus *Sepedonium* spp..

Das Isochinolin Alkaloid **20** wurde mit der Ausnahme von *S. brunneum* und *S. tulasneanum* in allen untersuchten *Sepedonium* Stämmen detektiert.
In den 1970er und 1980er Jahren gelangen die Charakterisierung des antifungal wirkenden Azaphilons Chrysodin (**21**) (1973, Closse und Hauser [262]) und der Nachweis des bereits zuvor aus *Penicillium brevicompactum* isolierten Polyketids Compactin (**22**) (1976, Brown [263, 264]) in *S. chrysospermum*. Compactin ist auch unter dem Namen Mevastatin bekannt und ist ein Inhibitor der 3-hydroxy-3-methylglutaryl-CoA Reduktase (HMG-CoA), einem Schlüsselenzym der Cholesterolbiosynthese. Mevastatin wurde zur Leitsubstanz der pharmakologischen Substanzklasse der Statine, die als Lipidsenker im Handel sind.
Mitova *et al.* untersuchten 2006 das Sekundärmetabolitenprofil von einem in Neuseeland isolierten *S. chrysospermum* [265]. Im Kulturfiltrat entdeckten sie das bereits bekannte 3,6-dimethyl-β-resorcylaldehyd (**23**) und das zyklische Pentapeptid Chrysosporid (**24**), das über schwache zytotoxische Eigenschaften verfügt. Aus dem Extrakt des festen Mediums gelangen die Nachweise von 2,4-dihydroxy-3-methyl-6-(2-oxopropyl)benzaldehyde (**25**), einem Intermediat in der Synthese der Azaphilone [266], sowie einer Mischung aus Phomalacton (**26**) und Musacin D (**27**).

1.5.2. Peptaibole aus Sepedonium spp.

Die Gattung *Sepedonium* ist vor allem für die Biosynthese von Peptaibolen bekannt. Einen Überblick über die derzeit bekannten Peptaibole aus *Sepedonium* gibt Tabelle 1.4.
Die erste Studie zu Peptaibolen in *Sepedonium* stammt von Dornberger *et al.* [218]. 1995 isolierten sie die SF-1 Typ Peptaibole Chrysospermin A-D (**28**-**31**) aus *Apiocrea chrysospermus* (= *Hypomyces chrysospermus*). Typisch für die Chrysospermine sind ein Acetylphenylalanin am *N*-Terminus, ein Prolin an Position 14 sowie ein C-terminales Tryptophanol. Die Chrysospermine inhibieren das Wachstum verschiedener Gram-positiver Bakterien [218], induzieren die Pigmentbildung von *Phoma destructiva* [218] und besitzen eine Neuroleptika-ähnliche Aktivität, die mit der von Chlorpromazin vergleichbar ist [227]. Weiterhin wurden auch zytotoxische Effekte sowie inhibitorische Eigenschaften gegen das Tabakmosaikvirus nachgewiesen [225].
Aus einem australischen *S. chrysospermum* isolierten Mitova *et al.* das Chrysaibol (**32**) [221]. Im Gegensatz zu den anderen aus *S. chrysospermum* isolierten Peptaibolen besitzt **32** ein *N*-terminales Acetyltryptophan und ein für *Sepedonium* ungewöhnliches C-terminales Alaninol. Chrysaibol zeigt eine zytotoxische Wirkung gegen die murine Krebszellinie P388 (IC_{50} = 6,61 µM)) sowie eine antibiotische Wirkung gegen das Gram-positive Bakterium *Bacillus subtilis* (IC_{50} = 1,54 µM).

In *S. ampullosporum* wurden die Ampullosporine A-E4 (**61**-**40**) identifiziert, die der Klasse der SF-6 Typ Peptaibole zugeordnet werden [227, 228]. Typisch für diese Klasse sind Peptaibole mit 15 Aminosäuren, die *N*-terminal ein Acetyltryptophan sowie C-terminal ein Leucinol tragen. Ferner fehlt das in Peptaibolen abundant auftretende Prolin [187].
Die Ampullosporine sind in Hinblick auf Struktur und Reaktivität sehr gut untersucht. 2003 publizierten Kronen *et al.* die Kristallstruktur von Ampullosporin A (AmpA, **61**) [197]. Das Molekül besitzt eine α-helikale Struktur zwischen Ac-Trp^1 und Aib^{13} sowie einen C-Terminus mit β-schleifenartiger Struktur. Innerhalb von Membranen nimmt AmpA vermutlich durch sterische Restriktionen eine gemischte $\alpha/3_{10}$-helikale Struktur an [198]. Durch die Kombination aus hydrophoben und hydrophilen Abschnitten besitzt AmpA einen amphiphatischen Charakter. Während die hydrophilen Stukturen über die β-Carboxylamidgruppen der Glutamine an den Positionen 7, 11 und 14, sowie über die Carbonylsauerstoffe von Aib^{10} und Gln^{11} ausgebildet werden, ermöglichen die hydrophoben Abschnitte, bestehend aus Trp^1, Leu^5, Leu^{12} und $LeuOH^{15}$, die Integration in die Zellmembran [198]. Für die Ampullosporine sind verschiedenste biologische Aktivitäten beschrieben, die auf ihre membranaktiven Eigenschaften zurückgeführt werden können [227, 226, 228, 267, 268]. So ist

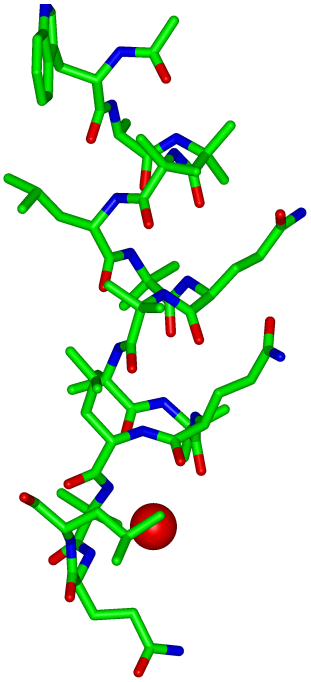


Abbildung 1.12.: Kristallstruktur von AmpA (**61**) ermittelt durch Röntgenstrukturanalyse von [197].

für alle Ampullosporine eine neuroleptikaartige Induktion von Hypothermie bei Mäusen sowie die Induktion der Pigmentbildung von *Phoma destructiva* beschrieben [228, 267]. Durch Untersuchungen an Modellmembransystemen vermuten Eid [214], Bortolus [215] und Salnikov [198], dass die membranaktiven Eigenschaften von AmpA sowohl auf einer spannungsabhängigen Bildung von Membranporen als auch auf einem detergenzartigen Effekt (Carpet Mechanismus) beruhen. Interessanterweise scheint der *N*-Terminus für die Aktivität von besonderer Bedeutung zu sein. Im Gegensatz zu den Ampullosporinen verursachte Desactyltryptophanyl AmpA weder einen Anstieg der Membranleitfähigkeit an Modellmembranen noch eine Induktion der Pigmentbildung bei *Phoma destructiva* [269, 270, 267].
Neben den Ampullosporinen wurde das aus nur fünf Aminosäuren bestehende Peptaibolin (**41**) in *S. ampullosporum* nachgewiesen [271]. Peptaibolin ist das kürzeste der derzeit bekannten Peptaibole; es verfügt dennoch über schwache antibiotische und antifungale Aktivitäten gegen *Bacillus subtilis* ATCC 6633 (MIK: 100 µg/mL) respektive *Candida albicans* (MIK: 100 µg/mL).

Während der Naturstofftage in Irsee stellte Stadler 2001 die Microspermine A-H (**42**-**49**)

in einer Posterpräsentation vor [224]. Die Microspermine wurden aus einem als *S. microspermum* beschriebenen Pilz isoliert und besitzen eine hohe Sequenzhomologie zu den Chrysosperminen. Sie beginnen *N*-terminal allerdings mit Acetyltryprophan (anstatt Acetylphenylalanin) und enden auf Leucinol (anstatt Tryptophanol). Für die Microspermine wurde von Stadler eine antivirale Wirkung gegen HI-Viren nachgewiesen [224].

Mit der Isolierung und Charakterisierung der Tulasporine A-D (**50**-**51**) aus *S. tulasneanum* gelangen Otto *et al.* erstmals der Nachweis, dass auch *Sepedonium* spp. mit ovalförmigen Aleuriokonidien Peptaibole produzieren [223]. Die Tulasporine gehören zu den SF1-Typ Peptaibolen und besitzen eine hohe Sequenzhomologie zu den Chrysosperminen. Der Unterschied besteht im Wesentlichen aus einem konservierten Glutamin anstatt eines Alanins an Sequenzposition 12 (Tabelle 1.4). Die Tulasporine zeigen moderate Aktivitäten gegen phytopathogene Pilze wie *Botrytis cinerea* und *Phytophthora infestans* ($IC_{50} \approx 50$ µM) [223].

Aus *S. chalcipori* wurden die Chalciporine A (**54**) und B (**55**) [272], Tylopeptin A (**56**) und B (**57**) [220, 272, 222] sowie die Chilenopeptine A (**58**) und B (**59**) [222] isoliert. Während die ersten beiden Gruppen den SF6-Typ Peptaibolen zuzuordnen sind, besitzen die Tylopeptine einen für *Sepedonium* spp. sehr ungewöhnlichen *N*-Terminus aus Ac-Aib. Dieser ist zwar typisch für Peptaibole aus *Trichoderma* spp. (z. B. Stilboflavine [273], Hypophelline [191], Trichobrachine [274]), jedoch stellt die Arbeit von Otto *el al.* der erste Nachweis für Peptaibole mit Ac-Aib in *Sepedonium* spp. dar.
Chalciporin A besitzt ähnlich wie die Tylopeptine eine wachstumsinhibierende Aktivität gegen Gram-positive Bakterien [275, 220]. Moderate antifungale Aktivitäten gegen phytopathogene Pilze wie *Botrytis cinerea* und *Phytophthora infestans* wurden von Otto *et al.* für die Tylopeptine und Chilenopeptine nachgewiesen [222].

Lee *et al.* isolierten Boletusin (**60**) von einem allgemein als *Boletus* sp. bezeichneten Pilz [276]. Es wird jedoch vermutet, dass der Wirtspilz mit *Sepedonium* sp. infiziert war, da das Boletusin zusammen mit den Chrysosperminen gefunden wurde [186]. Eigene (unpublizierte) Arbeiten bestätigen einen Fund von Boletusin in *S. laevigatum* und *S. microspermum*, nicht jedoch in *S. chrysospermum*, sodass der exakte Ursprung von Boletusin nicht bekannt ist. Boletusin zeigt eine moderate antibiotische Aktivität gegen Gram-positive Bakterien wie *Bacillus subtillis*, *Staphylococcus aureus* und *Corynebacterium lilium* [276].

Tabelle 1.4.: Publizierte Peptaibolsequenzen aus *Sepedonium* spp..

Peptaibol	Aminosäuresequenz	Literatur
S. ampullosporum		
Ampullosporin A (**61**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Aib**-**Aib**-**Aib**-Gln-Leu-**Aib**-Gln-LeuOH	[227]
Ampullosporin B (**34**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Ala**-**Aib**-**Aib**-Gln-Leu-**Aib**-Gln-LeuOH	[228]
Ampullosporin C (**35**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Aib**-**Ala**-**Aib**-Gln-Leu-**Aib**-Gln-LeuOH	[228]
Ampullosporin D (**36**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Aib**-**Aib**-**Ala**-Gln-Leu-**Aib**-Gln-LeuOH	[228]
Ampullosporin E1 (**37**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Ala**-**Aib**-**Aib**-Gln-Leu-**Ala**-Gln-LeuOH	[228]
Ampullosporin E2 (**38**	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Aib**-**Ala**-**Ala**-Gln-Leu-**Aib**-Gln-LeuOH	[228]
Ampullosporin E3 (**39**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Aib**-**Aib**-**Ala**-Gln-Leu-**Ala**-Gln-LeuOH	[228]
Ampullosporin E4 (**40**)	AcTrp-Ala-Aib-Aib-Leu-Aib-Gln-**Ala**-**Ala**-**Aib**-Gln-Leu-**Aib**-Gln-LeuOH	[228]
Peptaibolin (**41**)	AcLeu-Aib-Leu-Aib-PheOH	[271]
S. chalcipori		
Chalciporin A (**54**)	AcTrp-Val-Aib-**Val**-Ala-Gln-Ala-Aib-**Ser**-**Leu**-Ala-Leu-Aib-Gln-LeuOH	[272]
Chalciporin B (**55**)	AcTrp-Val-Aib-**Val**-Ala-Gln-Ala-Aib-**Gln**-**Aib**-Ala-Leu-Aib-Gln-LeuOH	[272]
Tylopeptin A (**56**)	AcTrp-Val-Aib-**Iva**-Ala-Gln-Ala-Aib-Ser-Aib-Ala-Leu-Aib-Gln-LeuOH	[272]
Tylopeptin B (**57**)	AcTrp-Val-Aib-**Aib**-Ala-Gln-Ala-Aib-Ser-Aib-Ala-Leu-Aib-Gln-LeuOH	[272]
Chilenopeptin A (**58**)	AcAib-Ser-**Trp**-Aib-Pro-Leu-Aib-Aib-Gln-Aib-Aib-Gln-Aib-Leu-PheOH	[222]
Chilenopeptin B (**59**)	AcAib-Ser-**Phe**-Aib-Pro-Leu-Aib-Aib-Gln-Aib-Aib-Gln-Aib-Leu-PheOH	[222]
S. chrysospermum		
Boletusin (**60**) [1]	AcPhe-Aib-**Ala**-Aib-**Iva**-Leu-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-**Aib**-Aib-Aib-Gln-TrpOH	[276]
Chrysospermin A (**28**)	AcPhe-Aib-**Ser**-Aib-**Aib**-Leu-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-**Aib**-Aib-Aib-Gln-TrpOH	[218]
Chrysospermin B (**29**)	AcPhe-Aib-**Ser**-Aib-**Aib**-Leu-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-**Iva**-Aib-Aib-Gln-TrpOH	[218]
Chrysospermin C (**30**)	AcPhe-Aib-**Ser**-Aib-**Iva**-Leu-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-**Aib**-Aib-Aib-Gln-TrpOH	[218]
Chrysospermin D (**31**)	AcPhe-Aib-**Ser**-Aib-**Iva**-Leu-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-**Iva**-Aib-Aib-Gln-TrpOH	[218]
Chrysaibol (**32**)	AcTrp-Aib-Aib-Leu-Val-Gln-Aib-Aib-Aib-Gln-Leu-Aib-Pro-Gln-AlaOH	[221]
S. microspermum		
Microspermin A (**42**)	AcTrp-**Aib**-Ser-Aib-**Iva**-**Trp**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Aib**-Gln-LeuOH	[224]
Microspermin B (**43**)	AcTrp-**Aib**-Ser-Aib-**Aib**-**Trp**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Iva**-Gln-LeuOH	[224]
Microspermin C (**44**)	AcTrp-**Iva**-Ser-Aib-**Iva**-**Trp**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Aib**-Gln-LeuOH	[224]
Microspermin D (**45**)	AcTrp-**Aib**-Ser-Aib-**Iva**-**Trp**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Iva**-Gln-LeuOH	[224]
Microspermin E (**46**)	AcTrp-**Iva**-Ser-Aib-**Iva**-**Trp**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Iva**-Gln-LeuOH	[224]
Microspermin F (**47**)	AcTrp-**Iva**-Ser-Aib-**Iva**-**Leu**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Aib**-Gln-LeuOH	[224]
Microspermin G (**48**)	AcTrp-**Aib**-Ser-Aib-**Iva**-**Leu**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Iva**-Gln-LeuOH	[224]
Microspermin H (**49**)	AcTrp-**Iva**-Ser-Aib-**Iva**-**Leu**-Gln-Gly-Aib-Aib-Ala-Ala-Aib-Pro-Aib-Aib-**Iva**-Gln-LeuOH	[224]
S. tulasneanum		
Tulasporin A (**50**)	AcPhe-Aib-Ser-Aib-**Aib**-Leu-Gln-**Gly**-Aib-Aib-Gln-Ala-Aib-Pro-Aib-Aib-Gln-TrpOH	[223]
Tulasporin B (**51**)	AcPhe-Aib-Ser-Aib-**Iva**-Leu-Gln-**Gly**-Aib-Aib-Gln-Ala-Aib-Pro-Aib-Aib-Gln-TrpOH	[223]
Tulasporin C (**52**)	AcPhe-Aib-Ser-Aib-**Aib**-Leu-Gln-**Ala**-Aib-Aib-Gln-Ala-Aib-Pro-Aib-Aib-Gln-TrpOH	[223]
Tulasporin D (**53**)	AcPhe-Aib-Ser-Aib-**Iva**-Leu-Gln-**Ala**-Aib-Aib-Gln-Ala-Aib-Pro-Aib-Aib-Gln-TrpOH	[223]

1.6. Ziele der Arbeit

Die Ziele der Dissertation lassen sich wie folgt zusammenfassen:

1. Das primäre Ziel dieser Arbeit war die Entwicklung einer Massenspektrometrie-basierten Methode, die die *in silico* Identifizierung von aktivitätsrelevanten Metaboliten in Naturstoffextrakten ermöglicht. Die Methode sollte über die folgenden Charakteristika verfügen:
   - Identifizierung der aktivitätsrelevanten Signale möglichst schon im Rohextrakt
   - Geringer Zeitaufwand und einfache Handhabung („Quick & Easy“)
   - Datenanalyse auch für chemoinformatisch ungeübte Anwender geeignet
2. In einem zweiten Schritt sollte die Qualität der entwickelten Methode in Hinblick auf die Identifizierung von aktivitätsrelevanten Metaboliten mit einer Reihe von klassischen und modernen multivariaten Methoden untersucht werden.
3. Die Anwendbarkeit der entwickelten Methode sollte unter Laborbedingungen demonstriert werden. Dazu wurden zunächst die bioaktiven Komponenten in Extrakten von verschiedenen *S. ampullosporum* Stämmen durch die *in silico* Methode identifiziert. Signifikant mit der Bioaktivität korrelierte Verbindungen wurden anschließend isoliert und z.T. kausal auf ihre biologische Aktivität getestet.

Teil II.

Material und Methoden

2. Material und Methoden

2.1. Geräte

In dieser Arbeit wurden die folgenden Laborgeräte verwendet:

Funktion	**Gerät**	**Hersteller**
CO_2-Inkubator	CB 150	Binder
Dampfdrucksterilisation	Varioklav® 75 S	H+P
Mikroskop (Zellkultur)	CKX41	Olympus
Sterilarbeiten Pilze & Bakterien	Laminarbox KS 12	Kendro
Sterilarbeiten Zellkultur	Herasafe KSP	Thermo Scientific
Ultraschallbad	Super RK510-H	Bandelin Sonorex
Ultraschallbad	RK100	Bandelin Sonorex
Vakuumdestillation	IR-Dancer 360	Hettlab
Wasserbad	WNB	Memmert
Zentrifuge (Zellkultur)	Centrifuge 5810 R	Eppendorf

2.2. Pilzmaterial

2.2.1. S. ampullosporum

Es wurden die in Tabelle 2.1 aufgeführten Stämme verwendet.

Tabelle 2.1.: Verwendete *S. ampullosporum* Stämme

Stamm	KSH	Regensburg Nummer (S)	Wirt	leg./det.	Extraktmenge [mg]
S. ampullosporum	490	212	*Boletus radicans*		5,4
S. ampullosporum	498	223	*Boletus miniatoolivaceus*	Aleksandrovic	1,2
S. ampullosporum	499	227	*Boletus miniatuspallescens*	Aleksandrovic	1,5
S. ampullosporum	500	228	*Boletus miniatuspallescens*	Aleksandrovic	2,1
S. ampullosporum	502	230	*Boletus chromapes*	Aleksandrovic	1,4
S. ampullosporum	504	234	*Boletus retipes*	Aleksandrovic	1,5
S. ampullosporum	522	269	*Boletus spec.*		1,2
S. ampullosporum	523	270			1,6
S. ampullosporum	531	285	*Boletus fechtneri*	Besl	1,3
S. ampullosporum	532	286	*Boletus rhodopurpureus*	Besl	1,3
S. ampullosporum	533	287	*Boletus fechtneri*	Besl	1,6
S. ampullosporum	534	288	*Boletus calopus*	Besl	2,6
S. ampullosporum	537	290	*Xerocomus roseoalbidus*	Besl	1,9
S. ampullosporum	538	294	*Pisolithus arhizus*	Besl	1,5
S. ampullosporum	544	304	*Chalciporus piperatus*		3,0
S. ampullosporum	549	310	*Boletus retipes*		2,3
S. ampullosporum	559	330	*Boletus luteocupreus*	Bode	3,7
S. ampullosporum	560	331	*Boletus luridus* (?)	Bode	2,3
S. ampullosporum	561	333	*Xerocomus riparielllus*	Bode	1,3
S. ampullosporum	568	340	*Boletus rhodoxanthus*	Bode	1,6
S. ampullosporum	635	379	*Pisolithus arhizus*	Besl	1,7
S. ampullosporum	641	158	*Boletus radicans*	Sahr	1,7

2.2.2. Hygrophorus spp.

Für die Proof of Concept Studie wurden die in Tabelle 2.2 aufgeführten *Hygrophorus* Spezies verwendet.

Tabelle 2.2.: Verwendete *Hygrophorus* Spezies

Spezies	Fundort	Datum	leg./det.	Kollektion
H. agathosmus (Fr.) Fr. 1838	Prügel Burkunstadt	17.10.05	Arnold	38/05
H. chrysodon (Batsch.) Fr. (1838)	Uffenheim Naturwald Reservat Wolfsee	26.09.94	Arnold	130/94
H. lucorum Kalchbr. (1874)	Freyburg	30.10.07	Arnold	63/07
H. olivaceoalbus (Fr.) Fr. (1838)	Smaland (Schweden)	17.-24.09.07	Arnold	-
H. pustulatus (Pers.) Fr. (1838)	Neudorf Harz	25.11.03	Arnold	46/03

2.3. Anzucht von Sepedonium spp.

Kulturmedien Für die Anzucht von *Sepedonium* spp. wurden die in Tabelle 2.3 angegebenen Kulturmedien verwendet. Die Dampfdrucksterilisation erfolgte für 15 min bei 120 °C und 1,1 bar im Autoklaven (Fa. H+P).

Anzucht von Sepedonium Agarkulturen Die Anzucht von *Sepedonium* spp erfolgte in Petri-Schalen mit ca. 20 mL sterilisiertem MPA Medium (Tabelle 2.3). Für die Anzucht eines Stammes wurde zunächst eine Cryo-Kultur (gelagert in flüssigem N_2) aufgetaut. Durch mehrfaches Einstechen des Cryo-Röhrchens in das Agarmedium wurden die im Cryo-Röhrchen enthaltenen Sporen freigegeben. Nach etwa einwöchigem Wachstum bei

Tabelle 2.3.: Malz-Pepton Medium (MP). Zur Herstellung von Agarkulturen wird zusätzlich Agar hinzugegeben [MPA].

Malz	10 g
Pepton	2,5 g
Aqua dest.	ad 1 L
Agar-Agar (Fluka)	15 g

Raumtemperatur wurden 3 ca. 1 cm^2 große Stücke vom Rand der Kolonie (hohe Dichte an Phialokonidien) auf eine MPA-Petrischale überführt und 21 Tage lang bei 24 °C (Schüttler) inkubiert. Um zu verhindern, dass Kondenswasser auf die Pilzkulturen tropft, wurden die Petrischalen auf den Kopf, d.h. mit dem Agar noch oben, in den Schüttler gestellt. Um eine größere Menge an Pilzmaterial zu erhalten, wurden je drei Agarkulturen pro Stamm auf diese Weise angezogen.

Anzucht von Sepedonium Emerskulturen Für die Anzucht der Pilze wurden 1 L Erlenmeyerkolben (mit luftdurchlässigem Wattestopfen) autoklaviert und unter sterilen Bedingungen mit je 150 mL autoklaviertem MP-Medium befüllt. Als Inokulum diente ein ca. 1 cm^2 großes Agarstück, das aus dem Randbereich (weiße Phialokonidien) der *Sepedonium* Kolonie entnommen wurde.

2.4. Extraktionsmethoden

2.4.1. Extraktion von Sepedonium spp. Agarkulturen

Nach mehrwöchigem Wachstum von *Sepedonium* spp. zeigten die Agarkulturen eine deutliche Färbung, die von dem jeweils verwendeten Stamm abhängig ist. Um auch die sekretierten Metaboliten erfassen zu können, wurden die Agarkulturen als Ganzes extrahiert. Dazu wurden pro Stamm je drei Agarkulturen in einem Mörser mit flüssigem N_2 vereint und zu einem Pulver zerstoßen. Das gefrorene Pulver kann auf diese Weise über längere Zeit bei -80 °C gelagert werden.
Das gefrorene Pulver wurde in einem 500 mL Erlenmeyerkolben mit 100 mL MeOH 30 Mintuen lang im Ultraschallbad (Sonorex, 320 W, 35 kHz) extrahiert. Anschließend wurde das Pulver mit einem Faltenfilter (Grade 6, 110 mm) vom Extrakt separiert. Der Filterkuchen wurde noch 2x auf diese Weise extrahiert, so dass schließlich 300 mL Methanolextrakt pro Pilzkultur erhalten wurden. Der Extrakt wurde anschließend *in vacuo* bis zur Trockne eingeengt und ausgewogen.
Um den hohen Anteil an Zuckern zu reduzieren, wurde die Probe anschließend mit einer selbstgefertigten Diaion HP-20 (Supelco) Kartusche gereinigt. Die Kartuschen (Chromabond®, PP, PE Filter, 15 mL) wurden so hergestellt, dass das Massenverhältnis von Diaion zu Extrakt 3 : 1 beträgt.
Die nachfolgenden Reinigungsschritte wurden mit einer Chromabond® Vakuumkammer

(Macherey-Nagel) durchgeführt, da hiermit bis zu 12 Proben parallel bearbeitet werden können. Die Säule wurde zunächst 3x mit je 1 Säulenvolumen (20 mL) MeOH aktiviert und dann mit 3 Säulenvolumen dest. H_2O äquilibriert. Die Probe wurde in einer Massenkonzentration von β= 2 mg/mL in dest. H_2O gelöst und auf die Säule aufgetragen. Zucker, sowie weitere stark polare Verbindungen wurden anschließend mit 100 mL H_2O von der Säule eluiert. Die an der Säule gebundenen Verbindungen wurden anschließend mit 20 mL MeOH eluiert, *in vacuo* bis zur Trockne eingeengt (IR-Dancer 360, Hettlab AG) und ausgewogen.
In einem weiteren Reinigungsschritt wurden stark apolare Verbindungen und eventuelle Schwebstoffe mit einer Chromabond® LV-C_{18}ec Kartusche (Macherey-Nagel, 500 mg, 15 mL) abgetrennt. Dazu wurde die Säule zunächst 3x mit 15 mL MeOH äquilibriert. Die Probe wurde in MeOH (1 mg/mL) gelöst, auf die Säule aufgetragen und mit 5 mL MeOH eluiert. Anschließend wurde die Probe *in vacuo* bis zur Trockne eingeengt und mit einer Analysewaage (Sartorius) ausgewogen.

2.4.2. Extraktion von Hygrophorus spp.

Für das Proof of Concept Experiment wurden von Michels fünf Extrakte verschiedener *Hygrophorus* Spezies zur Verfügung gestellt. Die Herstellung der Extrakte ist in [277] beschrieben. Nach Extraktion des zerkleinerten Pilzmyzels mit den in Tabelle 2.4 angegebenen Lösungsmitteln, wurden alle Extrakte zunächst mit dem Rotationsverdampfer zur Trockne eingeengt und nach Aufnahme in MeOH mit einer Chromabond RP18 Kartusche (500 mg/ 3 mL, Macherey-Nagel) filtriert.

Tabelle 2.4.: Überblick über die im Proof of Concept Experiment verwendeten *Hygrophorus* spp. Extrakte (10 µg/mL) und deren Wachstumsinhibition gegenüber *Bacillus subtilis*.

Spezies	Extrakt	Lösungsmittel	Wachstumsinhibition (%)
H. lucorum	HE1	80 % MeOH	25
H. chrysodon	HE2	80 % MeOH	8
H. agathosmus	HE3	100 % EtOAc	19
H. pustulatus	HE4	100 % MeOH	17
H. olivaceoalbus	HE5	100 % MeOH	12

2.5. Chromatographische Methoden

2.5.1. Größenausschlusschromatographie (SEC)

Als stationäre Phase für die Größenauschlußchromatographie wurde Sephadex LH20 (Pharmacia) verwendet.

Tabelle 2.5.: Verwendete SEC-Systeme

SEC-System	Säulendimension	Fließmittel	Flussrate	Fraktionsvolumen
SEC1	74 x 3,5 cm	MeOH	1,0 mL/min	10 mL
SEC2	76 x 3,0 cm	MeOH	1,2 mL/min	10 mL

2.5.2. Adsorbtionschromatographie

Chromatographie mit Diaion HP-20 Um den hohen Zuckeranteil in den Rohextrakten zu reduzieren, wurde Diaion HP-20 als Säulenmaterial verwendet. Diaion HP-20 ist ein kugelförmiges Polymerharz aus Divinyl- (Styrol) und Vinylbenzol. Die Polymer-Matrix bindet mittelpolare bis hydrophobe Verbindungen und eignet sich somit zur Entfernung von polaren Substanzen, wie z.B. Zuckern und Salzen. Voruntersuchungen an einem *Sepedonium* Rohextrakt haben gezeigt, dass die Säulenkapazität bei einem Verhältnis Diaion/Extrakt $\geq$ 3 nicht überschritten wird. Vor der chromatographischen Trennung wurde das Diaion mit MeOH aktiviert und anschließend mit 3 Säulenvolumen H_2O äquilibriert. Die Probe wurde in H_2O gelöst und auf die Säule aufgetragen. Die polaren Verbindungen wurden mit 1-2 L H_2O eluiert. Nach diesem Waschschritt wurden die an der stationären Phase gebundenen Substanzen mit 250 mL MeOH eluiert.

Chromatographie mit Kieselgel Stark apolare Verbindungen und eventuelle Schwebstoffe wurden mit einer Chromabond® C_{18}ec Kartusche (Macherey-Nagel, 500 mg, 3 mL) abgetrennt. Dazu wurde die Säule zunächst 3x mit 1 Säulenvolumen MeOH äquilibriert. Die Probe wurde in MeOH (1 mg/mL) gelöst, auf die Säule aufgetragen und mit 5 mL MeOH eluiert.

2.5.3. HPLC

Die Aufreinigung und Isolierung von Verbindung **61** erfolgte mit den in Tabelle 2.6 angegebenen HPLC-Systemen. Als stationäre Phase wurde jeweils mit Octadecylsilan derivatisiertes Silicagel (RP18) verwendet.

- Säule 1: YMC-Pack ODS-A, 150x20 mm, 5 µm Partikelgröße
- Säule 2: YMC-Pack ODS-A, 150x4,6 mm, 5 µm Partikelgröße

Tabelle 2.6.: Verwendete HPLC-Systeme

HPLC-System	Pumpensystem	Detektor	Datenanalyse
HPLC1	Merk L-7150, Degasser	Hitachi L-7400 UV Detector	Data-Jet-Integrator
HPLC2	Knauer Wellchrom 2x K-1001, Degasser	UV Detektor K-280	Knauer Software Eurochrom 2000

Zur Aufreinigung wurden die folgenden Gradientensysteme verwendet:

Tabelle 2.7.: Gradientensystem HPLC1. Laufmittel: A = H_2O, B = ACN

Zeit	Flussrate [mL/min]	A (%)	B (%)
0,0	10,0	60	40
30,0	10,0	0	100

Tabelle 2.8.: Gradientensystem HPLC2 (isokratisch). Laufmittel: A = H_2O, B = ACN

Zeit	Flussrate [mL/min]	A (%)	B (%)
0,0	10,0	40	60
30,0	10,0	40	60

2.6. Biotests

2.6.1. Bestimmung der antibiotischen Aktivität

Die antibiotische Aktivität der Rohextrakte wurde an grampositiven Bakterien der Art *Bacillus subtilis* getestet. Dazu wurde der von Michels etablierte Wachtumsinhibitionsassay verwendet [278]. Der verwendete Stamm *Bacillus subtilis* 168 (PAbrB-iyfp) exprimiert eine im Hinblick auf die Translationseffizienz verbesserte Variante des gelb-fluoreszierenden Proteins (IYFP). Das IYFP wurde unter die Kontrolle des Promotors des abrB-Gens gestellt [279]. AbrB ist ein Protein, dass u.a. die Sporulationsinitiation reprimiert und ausschließlich während der Wachstumsphase exprimiert wird. Die Zunahme der Fluoreszenz ist proportional zur Konzentration der lebenden Zellen [277].

2.6.1.1. Kulturmedien

Die Anzucht von *Bacillus subtilis* erfolgte mit TY-Medium (Tab. 2.9). Die Sterilistion erfolgte durch Dampfdrucksterilisation im Autoklaven (15 min, 120 °C, 1,1 bar, H+P). Zur Selektion der Bakterien wurde die Chloramphemicol-Lösung durch einen Sterilfilter (0,2 µm, Nalgene) filtriert und dem autoklavierten Nährmedium hinzugegeben. Die Stammhaltung erfolgte auf TY-Agarplatten.

Tabelle 2.9.: TY-Medium (*Bacillus subtilis)*. Zur Herstellung von TY-Agarschalen wird die angegebene Menge Agar hinzugegeben.

Chloramphenicol (1g/L)	5 mL (5 µg/mL)
Hefeextrakt (Serva)	5 g (0,5 %)
NaCl (Roth)	10 g (1 %)
Trypton (BD)	10 g (1 %)
Aqua Dest.	Ad 1 L
Agar Agar (Fluka)	15 g (1,5 %)

2.6.1.2. Testvorschrift

Für den Biotest wurde zunächst eine Vorkultur hergestellt. Dazu wurden 50 mL autoklaviertes TY-Medium in einem sterilen 100 mL Erlenmeyerkolben mit *Bacillus subtilis* 168 inokuliert. Die Vorkultur wurde nun etwa 24 h bei 30 °C inkubiert. Anschließend wurde

die Zellzahl mit Hilfe einer Neubauer-Zählkammer bestimmt (Gleichung 2.1) und auf die erforderliche Zellzahl von 1x10^4 Zellen/mL mit TY-Medium verdünnt.

$$\text{Zellzahl pro Großquadrat x 250 x } 10^3 = \text{Zellzahl/mL} \tag{2.1}$$

Für die Durchführung des Tests wurden schwarze 96-Well Platten mit flachem Boden (BD Falcon™) verwendet. Alle zu testenden Extrakte und Substanzen wurden als Triplikate in drei verschiedenen Konzentrationen (1 mg/mL, 0,1 mg/mL, 0,01 mg/mL) getestet. Bei jedem Test wurde eine Kontrolle der Eigenfluoreszenz (K_1) der Testsubstanz respektive des zu testenden Extrakts als Einfachbestimmung durchgeführt. Weiterhin wurden auf jeder Mikrotiterplatte eine Positivkontrolle des Bakterienwachstums (K_2) sowie eine Bezugskontrolle (K_3, entspricht 100 % Bakterienwachstum) als Hexaplikate mitgeführt. Als Referenzsubstanz diente Erythromycin, das als Triplikat in drei Konzentrationen (1 mg/mL, 0,1 mg/mL, 0,01 mg/mL) getestet wurde. Nach Zugabe der Bakteriensuspension im letzten Pi-

Tabelle 2.10.: Pipettierschema *Bacillus subtilis* Assay

TY	300 µL TY-Medium
K_1	270 µL TY-Medium + 30 µL Probe
K_2	270 µL TY-Medium + 30 µL *Bacillus subtilis* (10^4 Zellen)
K_3	240 µL TY-Medium + 30 µL MeOH + 30 µL Bakteriensuspension (10^4 Zellen)
Probe	240 µL TY-Medium + 30 µL Probe + 30 µL Bakteriensuspension (10^4 Zellen)

pettierschritt, wurde sofort, d. h. zum Zeitpunkt t_0, die Intensität der Fluoreszenzemission im Mikrotiterplatten-Reader (Genios Pro, Tecan) gemessen. Nach einer Inkubationsdauer von 15 h bei 30 °C wurde die Fluoreszenzemission ein weiteres Mal gemessen (t_{15}). Die Geräteparameter sind in Tabelle 2.11 zusammengefasst. Die Wachstumsinhibition wurde mit folgender Formel berechnet:

$$\text{Wachstumsinhibition (\%)} = \left(1 - \frac{\bar{x}\,(Probe[t_{15} - t_0])}{\bar{x}(K_3[t_{15} - t_0])}\right) x100 \tag{2.2}$$

2.6.2. Bestimmung der zytotoxischen Aktivität

Studien zu zytotoxischen Effekten von *Sepedonium* Extrakten und isolierten Substanzen wurden an humanen HT-29 Zellen durchgeführt. Die Zellen wurden von der deutschen

Tabelle 2.11.: Messparameter

Messparameter	Einstellung
Exzitationswellenlänge λ_{ex}	510 nm (Bandbreite 10)
Emissionswellenlänge λ_{em}	535 nm (Bandbreite 10)
Messmodus	von oben
Verstärkung (manuell)	60
Anzahl der Blitze	10
Verzögerungszeit	0 µs
Integrationszeit	40 µs
Spiegelauswahl (automatisch)	50 %
Messung pro Kavität	3x3 quadratisch
Temperatur	27 °C
Schütteldauer (linear)	10 s
Schüttelintensität	niedrig
Setzzeit	1 s
Einheit	RFU

Sammlung von Mikroorganismen und Zellkulturen (DSMZ) bezogen (Bestellnummer: ACC 299) und entstammen dem Primärtumor einer 44 Jahre alten, kaukasischen Frau mit Kolon-Adenokarzinom [280]. Grundlage des verwendeten Viabilitätsassays ist die kolorimetrische Erfassung der Umsetzung eines gelben Tetrazolium Salzes (XTT) zum entsprechenden (orangefarbenen) Formazan Salz durch metabolisch aktive Zellen [281, 282].

2.6.2.1. Verwendete Lösungen

Zellkultur Medium Die Kultivierung der HT-29 Zellen erfolgt mit supplementiertem RPMI 1640 Medium.

Tabelle 2.12.: Zellkulturmedium

RPMI 1640 (PAA Laboratories)	90 %
fötales Kälberserum (Gibco)	10 %
Penicillin/Streptomycin-Lsg.	1 %
L-Alanyl-L-Glutamin	2 mM
HEPES-Puffer	16 mM (pH 7,25)

Lösungen für den Zellproliferationsassay Die Visualisierung der Zellviabilität erfolgte mit dem Cell Proliferation Kit II (XTT) der Firma Roche Diagonstics, Mannheim. Das XTT-Reagenz und das Elektronen-Kopplungs-Reagenz (ECR) wurden bei jedem Experiment frisch zusammengemischt. Dazu wurden das XTT-Reagenz (1mg/mL) und das ECR (0,383 mg/mL) zunächst im Wasserbad bei 37 °C aufgetaut. Die jeweils einzusetzenden Volumina wurden mit den Formeln (2.3) und (2.4) berechnet. Anschließend wurde das XTT mit RPMI 1640 Medium auf 0,34 mg/mL verdünnt (2.5).

$$\text{V(XTT) [mL]} = \frac{\text{5 mL x Anzahl der Kavitäten}}{\text{96 Kavitäten}} \tag{2.3}$$

$$\text{V(ECR) [mL]} = \frac{\text{0,1 mL x Anzahl der Kavitäten}}{\text{96 Kavitäten}} \tag{2.4}$$

$$\text{V(Medium) [mL]} = \frac{\text{V(XTT) + V(ECR)}}{0,34} \tag{2.5}$$

2.6.2.2. Zellkultivierung

Die Kultivierung der Zellen erfolgte in Zellkulturflaschen der Firma TPP (25 cm^2, 75 cm^2, 150 cm^2). Nach 48 h bis 72 h wurde das alte Medium abgesaugt und durch frisches, im Wasserbad auf 37 °C temperiertes, RPMI 1640 Medium ersetzt. Die Zellen wurden bei 37 °C und 5 % CO_2 in einem CO_2-Inkubator (Binder) inkubiert. Für alle Versuche wurden Zellen zwischen der 5. und 50. Passage verwendet.

2.6.2.3. Passagieren der Zellen

Für den Assay wurden HT-29 Zellkulturen bis maximal 80 % Konflueszenz verwendet, da nur für diesen Bereich ein linearer Zusammenhang zwischen Zellzahl und Inkubationszeit besteht. Die Subkultivierung erfolgte durch Passagieren in eine neue Zellkulturflasche. Dazu wurde zunächst das Medium abgesaugt und die Zellen mit 5 mL Dulbecco's PBS (PAA, H15-002) gewaschen. Um die adherent wachsenden Zellen von der Oberfläche zu lösen, wurden sie mit 1 mL Trypsin EDTA (1:250, PAA, L11-660) ca. 5 Minuten bei 37 °C inkubiert. Anschließend wurden die Zellen in 9 mL RPMI 1640 Medium resuspendiert und im Verhältnis 1:10 in eine neue Zellkulturflasche transferiert.

2.6.2.4. Testvorschrift

Aussäen der Zellen Die HT-29 Zellen wurden zunächst optisch auf ihre Eignung für den Test überprüft (Form, Konflueszenz < 80 %). Anschließend wurde das Kulturmedium unter sterilen Bedingungen abgesaugt. Im nächsten Schritt wurden die Zellen wie unter 2.6.2.3 beschrieben von der Oberfläche abgelöst. D. h. die Zellen wurden mit 5 mL Dulbecco's PBS gewaschen, 5 Minuten lang mit 1 mL Trypsin/EDTA bei 37 °C inkubiert und in 9 mL RPMI 1640 Medium resuspendiert. Die Zellen wurden nun in ein 15 mL Falcon™-Röhrchen transferiert und 5 Minuten bei 150 rpm in einer Zentrifuge (Eppendorf, 5810 R) zentrifugiert. Der Überstand wurde dekantiert und das Sediment in 5 mL Dulbecco's PBS resuspendiert. Im nächsten Arbeitsschritt wurden die Zellzahl mit Hilfe einer Neubauer-Zählkammer bestimmt. Dazu wurden 50 µL Zellsuspension und 50 µL Trypanblau in einem 1,5 mL Eppendorfgefäß vermengt und 10 µL dieser Suspension in die Zählkammer pipettiert. Nach Auszählen der viablen Zellen (werden durch Trypanblau nicht blau gefärbt), wurde die Zellzahl durch Formel (2.6) berechnet.

$$\text{Zellkonzentration [Zellen/mL]} = \frac{\sum Z}{GxFxT} \tag{2.6}$$

Z = Anzahl der viablen Zellen
G = Anzahl der Großquadrate
F = Fläche der Großquadrate (0,01 cm^2)
T = Tiefe der Kammer (0,01 cm)

Die HT-29 Zellen wurden in einer Konzentration von 1500 Zellen pro Kavität ausgesät. Die einzusetzende Zellkonzentration für eine 96-Well Mikrotiterplatte (100 µL /Kavität-> 9 mL/Platte) wurde nach der folgenden Formel berechnet:

$$\text{Stammsuspension [µL]} = \frac{135000 \text{ Zellen}}{\text{Zellkonzentration [Zellen/µL]}} \tag{2.7}$$

Die Stammsuspension wurde anschließend mit RPMI 1640 Medium auf 9 mL (pro Mikrotiterplatte) aufgefüllt und mit einer Mehrkanalpipette je 100 µL in die Kavitäten pipettiert. Da Messungen an den Rändern von Mikrotiterplatten häufig mit einem größeren Fehler behaftet sind als an den inneren Kavitäten, wurden die Zellen nicht in die Kavitäten am Rand der Mikrotiterplatte ausgesät. Die Zellen wurden anschließend für 16-20 h bei 37 °C und 5 % CO_2 inkubiert.

Testung der Sepedonium ampullosporum Extrakte Die aufgereinigten Extrakte wurden in einer Massenkonzentration von 1 mg/mL in MeOH aufgenommen (HAM102, LB05/S.13 ff.). Anschließend wurden jeweils 2 mL in ein austariertes Gefäß überführt und mehrere Stunden im N_2-Strom getrocknet. Nach Auswaage an der Analysewaage wurden die Extrakte in einer Massenkonzentration von 10 mg/mL in DMSO gelöst. Alle Testlösungen wurden als Triplikate getestet. Bei der Herstellung ist zu beachten, dass die Testlösungen zweifach konzentriert angesetzt werden müssen, da sie durch die in der Mikrotiterplatte vorhandene Zellsuspension im Verhältnis 1:2 verdünnt werden. Die in DMSO gelösten Extrakte wurden im Verhältnis 1:1000 mit RPMI 1640 Medium verdünnt und in 3 verschiedenen Konzentrationen (5 µg/mL, 0,5 µg/mL und 0,05 µg/mL) getestet.
Als Positivkontrolle wurde Digitonin verwendet. Dazu wurde eine 250 mM Digitonin-Stammlösung (in DMSO) 1:1000 mit RPMI 1640 Medium verdünnt.
DMSO dient als Referenz- und Wachstumskontrolle. Das DMSO wird 1:1000 mit RPMI 1640 Medium verdünnt. Die Endkonzentration von DMSO entspricht einer 0,05 %igen DMSO-Lösung.
Nachdem je 100 µL der vorbereiteten Lösungen (Testlösungen der Extrakte, DMSO, Digitonin) in die Kavitäten der Mikrotiterplatte pipettiert wurden, wurde diese für 72 h bei 37 °C und 5 % CO_2 inkubiert. Zur Ermittlung der Wachstumsinhibition wurden insgesamt drei Passagen auf diese Weise getestet, so dass für die Auswertung schließlich 9 Datenpunkte pro Testkonzentration zur Verfügung stehen.

Testung von Verbindung 61 Alle Testlösungen wurden als Triplikate getestet. Bei der Herstellung ist zu beachten, dass die Testlösungen zweifach konzentriert angesetzt werden müssen, da sie durch die in der Mikrotiterplatte vorhandene Zellsuspension im Verhältnis 1:2 verdünnt werden. Um mögliche chemische Transformationen zu verhindern, wurden die Stammlösungen bei -18 °C aufbewahrt und alle Verdünnungen erst kurz vor dem Test frisch hergestellt. Ausgehend von der jeweiligen Stammlösung wurden anschließend mit RPMI 1640 Medium 9 verschiedene Verdünnungen hergestellt.
Zur Herstellung einer 20 mM Stammlösung von **61** wurden 1,91 mg **61** (1,18 µmol) in 59 µL DMSO aufgenommen (HAM121, LB05/S.11 ff.). Vor jeder Messung wurden 1 µL der Stammlösung mit 999 µL RPMI 1640 Medium verdünnt. Hiermit wurden nun Lösungen in den Konzentrationen 50 µM, 20 µM, 10 µM, 7 µM, 5 µM, 2 µM, 0,4 µM, 0,08 µM und 0,016 µM hergestellt und getestet.
Als Positivkontrolle wurde Digitonin verwendet. Eine 250 mM Digitonin-Stammlösung (in DMSO) wurde zunächst 1:1000 mit RPMI 1640 Medium vorverdünnt. Anschließend wurden

neun Verdünnungen hergestellt. Digitonin wurde in den Konzentrationen 250 µM, 83 µM, 28 µM, 9 µM, 3 µM, 1 µM, 0,3 µM, 0,1 µM und 0,038 µM getestet.
DMSO diente als Referenz- und Wachstumskontrolle. Das DMSO wurde 1:1000 mit RPMI 1640 Medium verdünnt. Die Endkonzentration von DMSO entspricht einer 0,05 %igen DMSO-Lösung.
Nachdem je 100 µL der vorbereiteten Lösungen (Testlösungen von **61**, DMSO, Digitonin) in die Kavitäten der Mikrotiterplatte pipettiert wurden, wurde diese für 72 h bei 37 °C und 5 % CO_2 inkubiert. Zur Ermittlung einer IC_{50}-Kurve wurden insgesamt drei Passagen auf diese Weise getestet, so dass für die Auswertung schließlich neun Datenpunkte pro Testkonzentration zur Verfügung stehen.

Bestimmung der Zellviabilität mittels XTT Nachdem die Mikrotiterplatte 72 h inkubiert wurde, wurde das Medium vorsichtig aus den Kavitäten abgesaugt. Auf diese Weise sollte verhindert werden, dass Testsubstanzen mit potenziell reduzierenden Eigenschaften das XTT zum Formazan Salz reduzieren und somit artifiziell eine höhere Zellviabilität vorgetäuscht wird. Zur kolorimetrischen Erfassung der Zellproliferation wurde die benötigte XTT/ECR Lösung für jeden Assay frisch angesetzt (siehe Abschnitt 2.6.2.1) und je 150 µL der Lösung pro Kavität hinzugegeben. Um die Eigenabsorbtion des Mediums (Mediumkontrolle, MK) bestimmen zu können, wurden je 150 µL XTT/ECR Lösung in drei Kavitäten ohne Zellen gegeben. Die Mikrotiterplatte wurde anschließend wieder im Inkubator bei 37 °C gelagert und 4 h lang inkubiert. Anschließend wurde die Platte 1 Minute bei 370 x g zentrifugiert und die Absorption bei λ = 490 nm (600 nm Referenzwellenlänge) im Mikrotiterplattenreader (Molecular Devices) gemessen.

2.6.2.5. Datenauswertung

Zunächst wurde die Eigenabsorbtion des Mediums (MK) von allen gemessenen Extinktionswerten abgezogen. Der Mittelwert der DMSO Kontrolle ($mean(DMSO)$) wurde als Referenzwert für 100 % Zellwachstum definiert. Die prozentuale Inhibierung des Zellwachstums wurde durch die folgende Formel berechnet:

$$\text{Wachstumsinhibtion (\%)} = 100 - (100x\frac{Messwert - MK}{mean(DMSO) - MK}) \tag{2.8}$$

Die Berechnung des IC_{50}-Wertes für **61** erfolgte mit dem R-Paket *nplr* [283]. Das Paket verwendet eine 5-Parameter logistische Regression, die auch als Richards Funktion bekannt

ist [284, 285].

$$y = B + \frac{T - B}{[1 + 10^{b(x_{mid} - x)}]^s} \tag{2.9}$$

Dabei entsprechen B und T der unteren und oberen Asymptote. Die Parameter b, x_{mid} und s sind die Steigung der Hill Gleichung, die x-Koordinate am Inflexionspunkt und ein asymmetrischer Koeffizient, respektive [283]. Das Paket optimiert alle 5 Parameter gleichzeitig über die Newton Methode. Der Optimierungsfaktor ist die Minimierung der gewichteten Summe der Abweichungsquadrate:

$$\sum_{i=1}^{n} w_i(y_i - \hat{y}_i) \quad \text{mit} \quad w_i = (\frac{1}{res_i})^p \tag{2.10}$$

2.7. Massenspektrometrie

2.7.1. Fourier-Transform-Ionencyclotronresonanz-Massenspektrometrie (FT-ICR-MS)

Die Messungen erfolgten mit einem FT-ICR Apex III 70e Massenspektrometer der Firma Bruker Daltonics. Das Gerät ist ausgestattet mit einer Infinity™-Zelle und einem 7.0 Tesla Supraleitermagneten (Bruker, Karlsruhe, Deutschland). Zur Elektrosprayionisation diente eine Apollo Ionenquelle. Die Messungen erfolgten per Direkteinlaß mit einer Spritzenpumpe (Agilent) bei einer Flussrate von 120 µL/h. Nach der Aufnahme der Spektren mit 512k Datenpunkten erfolgte ein 16x Zerofilling und die anschließende Fourier-Transformation. Voruntersuchungen mit Verdünnungsreihen haben gezeigt, dass für die meisten Peaks in MS-Spektren von Sepdonium Extrakten eine lineare Beziehung zwischen Konzentration und Peakintensität besteht, wenn der intensivste Peak im Spektrum eine Intensität von ca. 2 bis $3x10^7$ besitzt. Aus diesem Grund wurden vor jeder Messreihe zunächst einige Testmessungen an drei bis fünf Proben durchgeführt. Dabei wurden die Einstellungen für Ionenakkumulation und Scananzahl solange variiert, bis der intensivste Peak im Spektrum eine Intensität im oben angegebenen Bereich besaß. Die Parameterkombination, die bei den Testmessungen am häufigsten zu dem gewünschten Ergebnis führte, wurde für die gesamte Messreihe verwendet. Alle Extrakte wurden als Triplikate gemessen.

Tabelle 2.13.: Messparameter FT-ICR-MS

Messparameter	Einstellung Positiv-Ionen-Modus (Negativ-Ionen-Modus)
Trockengas (N_2)	150°C
Kapillarspannung	-4200 V (+4200 V)
Kapillarausgangpannung	100 V (-100 V)
Endplatten Spannung	-3700 V
Skimmer 1	15 V (-15 V)
Skimmer 2	6 bis 8 V (-6 bis -8 V)
Ioneninjektion	1800 µs
Fallenspannung	15 V (-15 V)
DC Offset Hexapol	1,5 V (-1,25 V)
Messbereich	*m/z* 97 bis 2000

2.7.2. UPLC-Quadrupol-Time-of-Flight-Massenspektrometrie (QqTOF-MS)

Alle Proben wurden zunächst in einer Massenkonzentration von 1 mg/mL in MeOH aufgenommen und anschließend 1:10 mit MeOH verdünnt. Die chromatographischen Trennungen erfolgten in einer HSS T3 Säule (1.0 x 100 mm, Partikelgröße 1.8 µm, Waters) mit einem Acquity UPLC System (Waters). Je Messung wurden 2 µL der Probe (0,1 mg/mL in MeOH) im Full-Loop-Modus injiziert und bei einer Flußrate von 200 µL/min chromatographisch aufgetrennt. Das verwendete Gradientensystem ist in Tabelle 2.14 wiedergegeben. Um möglichst konstante Messbedingungen zu schaffen, wurden Probenraum und UPLC-Säule auf 8 °C respektive 40 °C temperiert.

Tabelle 2.14.: Gradientensystem UPLC-QqTOF-MS. Laufmittel: A = H_2O/0.1 % CH_3COOH, B = ACN/0.1 % CH_3COOH

Zeit	Flussrate [mL/min]	A (%)	B (%)
0,0	0,2	98	2
1,0	0,2	98	2
9,0	0,2	2	98
14,1	0,2	2	98
15,0	0,2	98	2
10,0	0,2	98	2

Die eluierten Substanzen wurden in einem Massenbereich von *m/z* 100-2000 mit einem MicrOTOF-Q II Hybrid Quadrupol time-of-flight Massenspektrometer (Bruker, Billerica, MA), ausgestattet mit einer Apollo II Elektrospray Ionen Quelle, im Positiv-Ionen-Modus gemessen (HAM129, LB06, S.78 ff.). Zur internen Massenkalibration wurden bei jeder Messung zum Zeitpunkt t = 13 min über ein Umlenkventil 20 µL einer 10 mM Lithiumformiat Lösung in Isopropanol/Wasser (1/1 (v/v)) injiziert. Die Meßparameter sind in Tabelle 2.15 angegeben. Jeder Extrakt (0,5 µg/µL in MeOH) wurde als Triplikat analysiert.

Tabelle 2.15.: Messparameter MicrOTOF Q II

Messparameter	**Einstellung Full-Scan-Modus (MS^2-Modus)**
Zerstäubergas (N_2)	1,4 bar
Trockengas (N_2)	6 L/min, 190 °C
Kapillarspannung	4500 V
Endplatten Offset	-500 V
Funnel 1 RF	200 Vpp
Funnel 2 RF	200 Vpp
ISCID Energie	0 V
Hexapol RF	100 Vpp
Quadrupol Ionen Energie	5 eV
Quadrupol RF	45,6 Vpp (92,3 Vpp)
Kollsionsgas	Argon
Kollisionsenergie	10 eV (15, 30, 60 eV)
Kollisionszellen RF	200 Vpp
Transferzeit	70 µs (99 µs)
Prä-Puls Speicherzeit	5 µs
Isolationsweite	(10 m/z)

2.7.3. UPLC-Ion Trap-Massenspektrometrie (UPLC-IT-MS)

Zur Bestimmung der Primärstruktur der Peptaibole wurden MS^n Experimente mit einem UPLC-Ionenfallen-Massenspektrometer (UPLC-IT-MS) System durchgeführt (HAM129IT, LB06, S. 87 ff.). Die chromatographischen Trennungen erfolgten mit einem Acquity UPLC System (Waters) mit einer HSS T3 Säule (1,0 x 100 mm, Partikelgröße 1,8 µm, Waters). Es wurde das in Tabelle 2.16 aufgeführte Gradientensystem verwendet. Je Messung wurde 1 µL der Probe im „Partial Loop with Needle Overfill"-Modus injiziert und mit einer Fließgeschwindigkeit von 200 µL/min chromatographiert. Der Verlauf der Chromatographie wurde mit einem Photodioden Array Detektor (PDA) bei einer Wellenlänge von λ = 210 nm und λ = 280 nm aufgezeichnet. Um möglichst konstante Messbedingungen zu erzeugen, wurden Probenraum und UPLC-Säule auf 8 °C respektive 40 °C temperiert. Das UPLC-Sytem ist On-line mit einem LCQ Deca XP MAX Ionenfallen Massenspektrometer (Thermo Finnigan) gekoppelt, so dass die eluierten Substanzen hinsichtlich ihres Masse-zu-Ladungsverhältnisses (m/z) analysiert werden konnten. Die verwendeten Geräteparameter sind in Tabelle 2.17 aufgeführt.

Tabelle 2.16.: Gradientensystem UPLC-IT-MS. Laufmittel: A = H_2O/0,2 % CH_3COOH, B = ACN/0,2 % CH_3COOH

Zeit	Flussrate [mL/min]	A (%)	B (%)
0,00	0,2	70	30
1,00	0,2	70	30
6,00	0,2	5	95
9,00	0,2	5	95
9,01	0,2	70	30
10,00	0,2	70	30

Tabelle 2.17.: Messparameter Deca XP Ionenfalle

Messparameter	Einstellung Positiv-Ionen-Modus (Negativ-Ionen-Modus)
Schutzgas (N_2)	40 a.u.
Quellenspannung	4,5 kV (4 kV)
Kapillarspannung	27 V (-47 V)
Kapillartemperatur	275°C
Endplatten Offset	-500 V

2.7.4. ESI-MS

Zur Kontrolle der Aufreinigung der zu isolierenden Verbindungen wurde ein API 150EX Massenspektrometer (Applied Biosystems) verwendet. Das Gerät ist mit einer „Turbo Ion Spray“ Ionenquelle ausgestattet. Die Injektion der Proben (10 µL) erfolgt über einen HTC-PAL Autosampler per Direkteinlaß.

2.8. IR-, UV/Vis-Spektroskopie

UV-Spektren wurden mit einem Jasco V-560 UV/Vis-Spektrometer im Bereich zwischen 190 und 800 nm aufgenommen. Die Messungen der Infrarotspektren erfolgte mit einem Thermo Nicolet 5700 FT-IR Spektrometer. Spezifische Drehwerte wurden mit einem Jasco DIP-1000 Polarimeter gemessen.

2.9. AcorA

2.9.1. Prinzip der Aktivitäts-Korrelations-Analyse

Ein Naturstoffextrakt besteht aus hunderten bis tausenden von Substanzen, von denen häufig nur eine oder einige wenige eine biologische Aktivität aufweisen. In der Regel ist die Bioaktivität proportional zur Konzentration der aktiven Substanz(en). Weiterhin wird jede Substanz in spektroskopischen Messungen durch ein oder mehrere Signale repräsentiert, die Auskunft über Identität und Konzentration der Substanz geben. Häufig ist auch die Signalintensität der Substanz proportional zu ihrer Konzentration in der vorliegenden Matrix. Durch die Proportionalität zwischen Bioaktivität und Konzentration sowie zwischen Signalintensität und Konzentration muss auch ein proportionaler Zusammenhang zwischen Signalintensität und Bioaktivität bestehen.

1. Bioaktivität $\sim$ c(Substanz)
2. Signalintensität $\sim$ c(Substanz)

$\Rightarrow$ Signalintensität $\sim$ Bioaktivtät

Die Aktivitäts-Korrelations-Analyse nutzt nun diese Korrelation zwischen Bioaktivität und Signalintensität um bioaktive Substanzen in einer komplexen Mischung zu identifizieren.
Da von einem Rohextrakt immer nur die Gesamtbioaktivität bestimmt werden kann, ist es aus statistischen Gründen notwendig, eine ganze Reihe von Extrakten herzustellen, die die bioaktive(n) Substanz(en) in möglichst unterschiedlichen Konzentrationen enthalten. Dies kann beispielsweise dadurch erreicht werden, dass man einen aktiven Extrakt auf unterschiedliche Weise modifiziert (z. B. Fraktionierung, chemische Modifizierung etc.). Da die Biosynthese vieler Substanzklassen artspezifisch ist, kann man alternativ auch die natürliche Variation einer bioaktiven Substanz in Extrakten mehrerer verwandter Arten für AcorA ausnutzen. Durch Messung der Bioaktivität mehrerer Extrakte erhält man somit deren Aktivitätsprofil. Das entsprechende Metabolitenprofil wird generiert, in dem man dieselben Extrakte mit spektroskopischen Methoden (beispielsweise MS, LC-MS, NMR, IR) analysiert. Korreliert man nun die Einzelpeaks aus dem Metabolitenprofil mit den gemessenen Bioaktivitäten, sollten die Peaks, die für die biologische Aktivität verantwortlich sind, eine signifikante Korrelation aufweisen.
Zur Bestimmung der Korrelation zwischen Aktivitäts- und Metabolitenprofil wird die Spearman-Rangkorrelation verwendet. Sie bietet u. a. folgende Vorteile:

1. Es muss kein linearer Zusammenhang zwischen Signalintensität und Bioaktivität bestehen. Dieser Aspekt ist insbesondere bei Auftreten von mehreren, möglicherweise additiv oder auch synergistisch wirkenden Substanzen relevant. Ebenso ist bei vielen biologischen Aktivitätsassays nur ein proportionaler, aber nicht unbedingt linearer, Zusammenhang zwischen den Messgrößen gegeben.
2. Durch die Überführung der Messwerte in Ränge ist die Spearman-Rangkorrelation robust, d. h. unempfindlich gegenüber Ausreißern.
3. Die Methode ist skaleninvariant, d. h. kleine und große Peaks werden in ihrer Bedeutung gleich gewichtet.
4. Die Messwerte müssen weder normalverteilt, noch muss die den Messwerten zugrunde liegende Verteilung bekannt sein.

Ausgangspunkt für die Berechnung des Spearman-Rangkorrelationskoeffizienten ist eine Datenmatrix, die in jeder Zeile x_1,...,x_n die Informationen über die Identität des Peaks enthält (*m/z*-Wert, rt/*m/z*-Wert, chemische Verschiebung δ...). In den Spalten der Datenmatrix ist die Intensität der Peaks x_{i1},..., x_{im} über m Extrakte enthalten. Die unterste Zeile der Matrix enthält die Werte für die Bioaktivität B_1,...,B_m in den m Extrakten. Die Berechnung des Spearman-Rangkorrelationskoeffizienten erfolgt nach Formel 2.11:

$$\rho_i = 1 - \frac{6 \sum_{j=1}^{m} {d_{ij}}^2}{m \cdot (m^2 - 1)} \qquad mit \quad d_{ij} = Rang(x_{ij}) - Rang(B_j) \tag{2.11}$$

Dabei werden die Messwerte für Peakintensität und Bioaktivität zunächst in Ränge überführt, d. h. der Größe nach sortiert. Anschließend wird geprüft, inwieweit die Ränge in den einzelnen Extrakten übereinstimmen. Bei vollständiger Übereinstimmung des Verlaufs der Ränge der beiden Kenngrößen erhält man $\rho = 1$ (maximale Korrelation).
Nach Berechnung aller ρ_i-Werte schließt sich die Frage an, ab welchem Rangkorrelationswert von einer statistisch signifikanten Korrelation gesprochen werden kann. Zu diesem Zweck wird unter der Nullhypothese H_0, d. h. es gibt keine Korrelation zwischen Bioaktivität und Signalintensität, ein Permutationstest durchgeführt. Mit Hilfe des Permutationstests wird die Verteilung der Spearman-Rangkorrelationkoeffizienten $P(\rho)$ unter der Nullhypothese H_0 approximiert. Dazu werden die Intensitätswerte in jeder Zeile x_i zufällig permutiert und der Spearman-Rangkorrelationskoeffizient ρ neu berechnet. Dieser Vorgang wird für jede Zeile n-Mal (mit $n \geq 10000$) wiederholt. In Abbildung 2.1 ist exemplarisch ein Histogramm abgebildet, das die Verteilung $P(\rho)_{H_0}$ der Rangkorrelationskoeffizienten

unter der Nullhypothese H_0 widerspiegelt. Entsprechend dieser Verteilung $P(\rho)_{H_0}$ werden die Signifikanzgrenzen derart gewählt, dass man unter H_0 einen Fehler mit der Irrtumswahrscheinlichkeit α machen würde.

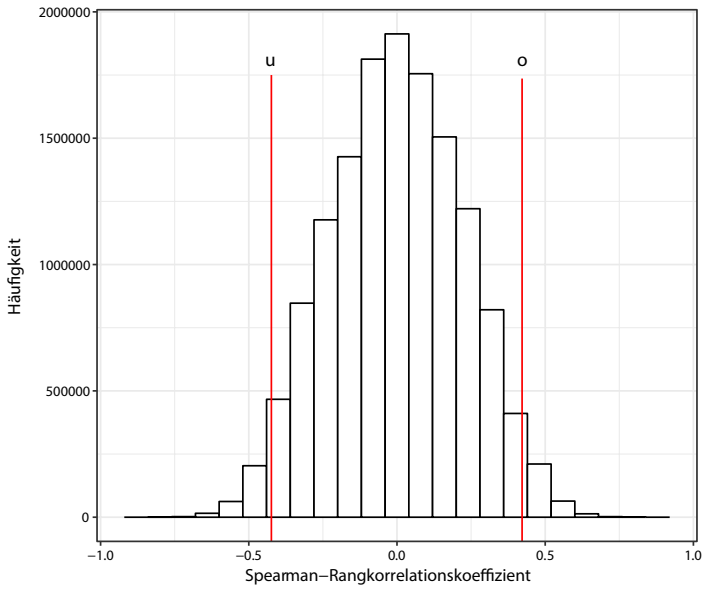


Abbildung 2.1.: Häufigkeitsverteilung $P(\rho)$ unter H_0. Die untere (u) und obere (o) Signifikanzgrenze mit Irrtumswahrscheinlichkeit α sind als rote Linien eingezeichnet. ρ-Werte < u entsprechen einer signifikant negativen Korrelation. ρ-Werte > o entsprechen einer signifikant positiven Korrelation.

Nun wird ein einseitiger oberer Test mit einer Irrtumswahrscheinlichkeit α durchgeführt:

- H_0: es gibt keine Korrelation zwischen Bioaktivität und Peakintensität
- H_1: es gibt eine positive Korrelation zwischen Bioaktivität und Signalintensität, d. h. diese Peaks werden Bestandteil der Hitliste

Für den Grenzwert gilt:

- oberer Grenzwert o: ρ -Wert, für den gilt: $P(\rho < o) = 1\text{-}\alpha$

Das Entscheidungsverfahren für die Signifikanz eines Peaks x_i lautet wie folgt:

$$d(x_i) := \begin{cases} H_1, & falls \quad P(\rho_i \geq o) = \alpha \\ H_0, & sonst \end{cases} \tag{2.12}$$

Alle Peaks, deren Intensitäten signifikant mit der Bioaktivität korrelieren, werden in einer Hitliste zusammengefasst. Diese Hitliste enthält diese Peaks geordnet nach kleiner werdenden Rangkorrelationskoeffizienten.
Um anschließend Hinweise über die der Hitliste zugrunde liegende(n) Substanz(en) bekommen zu können, muss zunächst eine Dekonvolution der signifikanten Peaks durchgeführt werden. D. h. Peaks, die einen gemeinsamen Ursprung haben, müssen entsprechend annotiert werden. Mit Parametern der annotierten Peaks, wie z.B. exakte Masse, Retentionszeit oder gegebenenfalls auch MS^n Daten, kann dann in Substanzdatenbanken wie z. B. Pubchem [286] und Dictionary Of Natural Products [287] oder auch in Spektrendatenbanken wie etwa MassBank [288] und METLIN [91] nach der potentiell aktiven Substanz gesucht werden. Im Hinblick auf die Dereplikation bietet die *in silico* Identifizierung einen enormen Vorteil, da nur noch eine kleine, ausgewählte Anzahl von Peaks identifiziert werden muss.
Da AcorA lediglich statistische Korrelationen liefert, muss der kausale Zusammenhang zwischen der Substanz und ihrer Bioaktivität in einem entsprechenden Assay belegt werden. Dazu muss die Substanz zunächst isoliert werden. AcorA bietet für die Isolierung einen entscheidenden Vorteil. Da die Masse der zu isolierenden Substanz durch die *in silico* Identifizierung bereits bekannt ist, kann die Substanz in den Fraktionen, die während des Aufreinigungsprozesses generiert werden, durch massenspektrometrische Methoden sehr leicht verfolgt werden. Eine zeitraubende aktivitäts-geleitete Fraktionierung, bei der die einzelnen Fraktionen mit einem entsprechenden Assay getestet werden müssten, ist also nicht erforderlich. Der prinzipielle Ablauf eine Aktivitäts-Korrelations-Analyse ist schematisch in Abbildung 2.2 dargestellt.

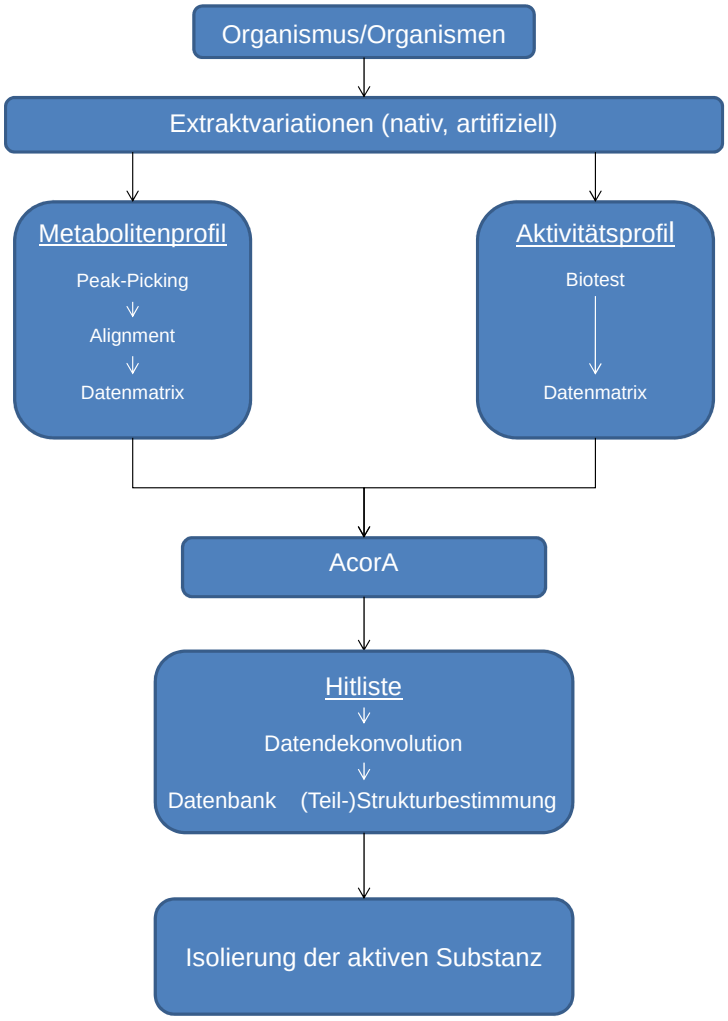


Abbildung 2.2.: AcorA Ablaufschema

2.9.2. Datenprozessierung/Datenanalyse

Datenprozessierung Der Ablauf der Datenauswertung ist in Schema 2.3 dargestellt. Zunächst wurden die Rohdaten mit dem Programm CompassXport (Bruker) in das mzData Format konvertiert. Alle weiteren Datenverarbeitungsschritte wurden mit der Statistik-Software R [289] durchgeführt. Das Peak-Picking der MS-Daten erfolgte mit dem MassSpecWavelet Algorithmus [290] des R-Paketes XCMS [291, 292]. Das Peak-Picking wurde mit den Skalen 1 und 7 durchgeführt, wobei nur Peaks mit einem Signal/Rausch-Verhältnis *snthresh* > 3 und einer minimalen Peakintensität von *peakThr* = 80000 berücksichtigt wurden.

```
xsPos <- xcmsSet(method="MSW",SNR.method='data.mean',winSize.noise=500,
    peakThr=80000, snthresh=3,amp.Th=0.005, scales=c(1,7))
```

Peaks aus Replikatmessungen wurden zusammen aligniert, wenn sie in mindestens zwei von drei Replikatmessungen (*minsamp* = 2) mit einer vorher definierten maximalen Massenabweichung (*mzppm* = 5-10 ppm) beobachtet wurden. Dazu wurden die Replikatmessungen zunächst gruppiert:

```
# Gruppierung #
classes<-sampnames(xsPos)
names<-sampnames(xsPos)
dim(names)<-c(length(names),1)
classes<-apply(names,1,function(x){substr(x,1,nchar(x)-7);})
sampclass(xsPos)<-classes

# Alignment #
xsgPos <- group.mzClust(xsPos, minsamp=2, mzppm=5)
groupmat <- groups(xsgPos)
values <- groupval(xsgPos, "medret", "into")
tab <- cbind(groupmat, values)
```

Vor der Durchführung der Aktivitäts-Korrelations-Analyse wurden zwei weitere Signalfilterungsschritte durchgeführt. Erstens wurden Peaks mit einem Masse zu Ladungsverhältnis kleiner als 150 aus der alignierten Peakliste gelöscht, da dieser Bereich in den FT-ICR-Spektren stark verrauscht ist und generell nur wenig interessante Peaks enthält. Zweitens wurden nur die Peaks in die endgültige Peakliste übernommen, die nicht in der Kontrolle (BW) vorkamen. Auf diese Weise wurden die Hintergrundpeaks der Malz-Pepton Agarkulturen eliminiert.

Datenanalyse Vor der Analyse der Daten wurden zunächst NA-Werte durch einen konstanten Wert von 0,0001 ersetzt (Anteil der NA-Werte im Proof of Concept Gesamtdatensatz: 74 %, Anteil der NA-Werte im *Sepedonium* Datensatz: 76 %). Die Aktivitäts-Korrelations-Analyse erfolgte mit dem R-Paket *AcorA*, das in unserer Arbeitsgruppe von A. Gohr programmiert wurde. Die Verteilung $P(\rho)_{H_0}$ wurde mit 1000 Permutationen approximiert. Die Signifikanzgrenzen wurden so gewählt, dass unter Annahme der Nullhypothese eine Irrtumswahrscheinlichkeit von α = 5 % auftreten kann.

2.9.3. Verfahren zur Auswertung der Hitliste

Nach der Durchführung der Aktivitäts-Korrelations-Analyse erhält man als Ergebnis eine Hitliste, die die signifikant korrelierenden *m/z* Werte sortiert nach Größe der Korrelationskoeffizienten ρ enthält. In Abhängigkeit des verwendeten massenspektrometrischen Verfahrens liegt die Anzahl der Peaks in der Hitliste etwa zwischen 50 (FT-ICR-MS Spektren) und 250 (LC-MS Spektren) [277, 293].
Eine idealisierte Hitliste würde ausschließlich *m/z*-Werte enthalten, die sich auf Substanzen mit biologischer Aktivität zurückführen lassen. Da die alignierten Peaklisten aus mehreren Tausend *m/z*-Werten bestehen, tritt ein multiples Testproblem auf, bei dem jeder einzelne Test mit einer Irrtumswahrscheinlichkeit α behaftet ist. Es ist daher nicht auszuschließen, dass die Hitliste auch falsch positive Elemente enthält, d. h. *m/z*-Werte von Substanzen, die nicht kausal mit der biologischen Aktivität assoziiert sind.
Um nun zwischen richtig positiven und falsch positiven Signalen in der Hitliste differenzieren zu können, kann eine Besonderheit in der Elektrospray-Ionisations-Massenspektrometrie (ESI-MS) ausgenutzt werden. Eine Verbindung erzeugt in der Regel nicht nur ein einziges Massensignal, sondern auch entsprechende Isotopen- und Adduktpeaks. Da diese Signale untereinander hoch korreliert sind, sollten theoretisch alle Massensignale einer biologisch aktiven Verbindung eine hohe Korrelation zur Bioaktivität aufweisen.

Auswertung über die Peakdichte Um die Auswertung der Hitliste zu erleichtern, können die *m/z*-Werte der Hitliste in Abhängigkeit der Peakdichte graphisch darstellt werden. Ein Bereich mit hoher Peakdichte weist auf ein Peakcluster bestehend aus Isotopen- und Adduktpeaks hin. Verfügen die Peaks in diesem Cluster zusätzlich über hohe Korrelationskoeffizienten, kann dies auf eine richtig positive Korrelation zur Bioaktivität hinweisen. Die Peakdichte kann mit dem folgenden R-Skript dargestellt werden.

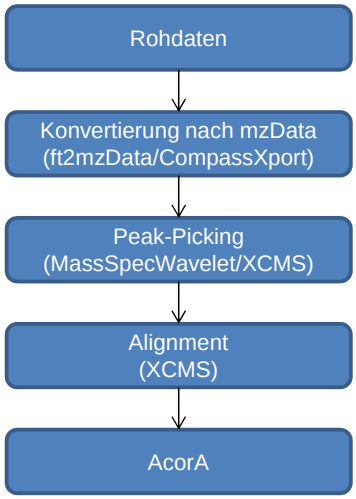


Abbildung 2.3.: Ablaufschema Datenprozessierung/Datenanalyse

```
### Skript zur Darstellung des Peakdichteplots
# data = Datenmatrix aus m/z Werten und korrespondierenden
    Rangkorrelationskoeffizienten aus Hitliste
library(ggplot)
ggplot(data=data,aes(x=m.z)) + geom_density(adjust= 0.2, alpha=0.99,
    fill="grey50")
### Skript Ende
```

Auswertung über die Korrelationsmatrix Wie oben beschrieben besteht zwischen den einzelnen Isotopen- und Adduktpeaks einer Verbindung auf natürliche Weise eine hohe Korrelation. Die graphische Darstellung dieser Korrelationen in einer Pearson-Korrelationsmatrix bietet eine weitere Möglichkeit zur schnellen Identifizierung von Isotopen- und Adduktpeaks einer Verbindung. Die Korrelationsmatrix wurde mit dem R-Paket *corrplot* generiert [294]. Die Daten wurden hierarchisch mit der Ward Methode [295] geclustert.

```
### Skript zur Berechnung und Visualisierung der Pearson
    Korrelationsmatrix der Peaks der Hitliste
# hitcor := Datenmatrix aus alignierter Peakliste mit den Peaks aus der
     Hitliste
library(corrplot)
mz.cor<-cor(hitcor,method = "pearson")
```

```
5 corrplot(mz.cor,method="color",hclust.method="ward",order="hclust",tl.
    srt=45,tl.col = "black", tl.cex = 0.35,type="lower",diag = F)
6 ### Skript Ende
```

2.9.4. AcorA: Proof of Concept Studie

In einer Proof of Concept Studie (Experiment HAM066, Laborbuch (LB) 02, S.92 ff.) sollte untersucht werden ob mithilfe der Aktivitäts-Korrelations-Analyse eine *in silico* Identifizierung von biologisch aktiven Substanzen in Naturstoffextrakten möglich ist. Dazu wurden von Michels fünf schwach aktive methanolische Grundextrakte (Wachstumsinhibition < 20 %) aus fünf Arten der Gattung *Hygrophorus* (siehe Tabelle 2.2) zur Verfügung gestellt. Diese wurden anschließend - wie weiter unten beschrieben - mit Antibiotika in verschiedenen Konzentrationen versetzt. Als Antibiotika wurden Amoxicillin, Erythromycin und Rifampicin gewählt, da alle drei auf einem unterschiedlichen Wirkmechnismus beruhen. Amoxicillin ist ein typisches Penicillin-Derivat, das die Zellwandsynthese der Prokaryonten inhibiert. Während Rifampicin die Transkription stört, blockiert Erythromycin die Proteinbiosynthese an den Ribosomen [296].
Die Grundextrakte wurden in einer Massenkonzentration von 18 mg/mL in MeOH angesetzt. Nun wurde 16x randomisiert einer der 5 Grundextrakte ausgewählt (jeweils 850 µL) und mit je 50 µL Amoxicillin, Erythromycin und Rifampicin versetzt. Die zuzusetzende Stoffmengenkonzentration wurde zufällig aus 4 verschiedenen Ausgangskonzentrationen pro Antibiotikum (200 µM, 2 µM, 0,2 µM, 0 µM, gelöst in MeOH) ausgewählt. Die Zugabe von 0 µM Antibiotikum wurde durch Hinzugabe von 50 µL MeOH simuliert. Auf diese Weise wurden die in Tabelle 2.18 dargestellten Extrakte erhalten.
Die biologische Aktivität der 16 gespickten Extrakte (HEA) sowie der 5 Grundextrakte (HE) wurde anschließend im Wachstumsinhibitionsassay mit *Bacillus subtilis* getestet (HAM066, LB02, S. 112 ff.). Dazu wurden Aliquots der Proben jeweils 1:10 und 1:100 mit MeOH verdünnt und entsprechend 2.6.1 analysiert. Die FT-ICR-MS Messungen erfolgten wie unter 2.7.1 beschrieben (HAM066, LB02, S. 112 ff.).

Tabelle 2.18.: Endkonzentrationen der gespickten *Hygrophorus* Extrakte des Proof of Concept Experiments

Extrakt	Hygrophorus Extrakt	Amoxicillin [µM]	Erythroymcin [µM]	Rifampicin [µM]	Wachstumsinhibition (%)
HE1	1	0	0	0	25 ± 4
HE2	2	0	0	0	8 ± 5
HE3	3	0	0	0	19 ± 6
HE4	4	0	0	0	17 ± 3
HE5	5	0	0	0	12 ± 6
HEA1	4	0,01	0,01	10	85 ± 3
HEA2	5	0,01	0,1	10	92 ± 1
HEA3	5	0	0,01	0	22 ± 13
HEA4	5	0,01	10	10	86 ± 5
HEA5	3	0,01	0	0,1	28 ± 2
HEA6	2	10	0,01	0,01	13 ± 3
HEA7	3	0,1	0,01	0	19 ± 6
HEA8	5	0	0	0,01	5 ± 10
HEA9	1	0,1	10	0,01	92 ± 3
HEA10	2	10	0	0,01	33 ± 5
HEA11	4	0	0,01	0	19 ± 5
HEA12	2	10	0,01	0,01	10 ± 11
HEA13	3	0,1	0,1	0,1	32 ± 6
HEA14	1	10	0,1	0,1	21 ± 1
HEA15	1	0,1	0,01	10	88 ± 3
HEA16	3	0	0	10	88 ± 2

2.9.5. AcorA mit Sepedonium ampullosporum Extrakten

In diesem Versuch sollte gezeigt werden, dass AcorA nicht nur unter artifiziellen Bedingungen, sondern auch unter realen Laborbedingungen genutzt werden kann, um biologisch aktive Verbindungen in Naturstoffextrakten *in silico* zu identifizieren. Für die Untersuchungen wurden die in Tabelle 2.1 aufgeführten Stämme wie in 2.4.1 beschrieben extrahiert (HAM087, LB04, S. 79 ff.). Für die Messungen mit dem FT-ICR-Massenspektrometer wurden alle Extrakte in einer Konzentration von 1 mg/mL in MeOH gelöst und anschließend im Verhältnis 1:10 mit MeOH verdünnt. Das Metabolitenprofil wurde durch Triplikatmessungen der Methanolextrakte im Positiv- und Negativ-Ionen-Modus erhalten (Abschnitt 2.7.1, MAA027, LB04, S. 164 ff.). Das Bioaktivitätsprofil der Extrakte wurde wie in Kapitel 2.6.2 beschrieben generiert (HAM102, LB05, S. 13 ff.). Datenprozessierung und Datenanalyse erfolgten wie in Abschnitt 2.9.2 (HAM114, LB05, S.109 ff.) dargestellt.

2.10. Isolierung von Verbindung 61

Für die Isolierung von **61** wurden 50 Erlenmeyerkolben, wie unter 2.3 beschrieben, mit dem *Sepedonium*-Stamm KSH533 inokuliert (HAM115, LB05, S. 141 ff.).
Nach einer Inkubationszeit von 25 Tagen wurden Myzel (HAM115y) und Kulturmedium (HAM115f) durch Vakuumfiltration (Filterpapier, Grade 6, ⊘ 24 cm) voneinander separiert. Das Kulturfiltrat (HAM115fW) wurde anschließend *in vacuo* auf 2 L eingeengt und 1x mit 2 L Ethylacetat extrahiert. Im Anschluß wurde das Volumen der wässrigen Phase am Rotationsverdampfer weiter auf 1 L eingeengt und 2x mit je 1 L Ethylacetat extrahiert. Die drei Ethylacetat-Phasen (HAM115fE) wurden vereinigt und zur Trockne eingeengt.
Das Myzel wurde mit einem Mixer zerkleinert und 3x mit je 1 L Ethylacetat auf einem Schüttler (105 rpm) extrahiert. Die drei Ethylacetat-Phasen des Myzelextrakts (HAM115yE) wurden mit der Ethylacetat-Fraktion HAM115fE vereinigt. Nach Entfernung des Lösungsmittels durch Vakuumdestillation wurden 989 mg Rohextrakt (HAM115E) erhalten. Messungen mit FT-ICR-MS bestätigten die Anwesenheit von Verbindung (**61**)
Durch Größenausschlußchromatographie an Sephadex LH-20 (SEC1) wurden die relativ großen Peptaibole von den meisten weiteren Metaboliten separiert. Nach Evaporation des Lösungsmittels in einem IR-Dancer (Hettlab AG) wurden Aliquots der aufgefangenen Fraktionen mittels ESI-MS analysiert. Die Fraktionen 30 bis 34 enthielten die gesuchte Verbindung mit der Masse m/z = 812. Die Fraktionen wurden vereinigt (HAM115_ES_30-34, 134 mg) und durch eine präparative HPLC (Gradientensystem HPLC1) weiter aufgereinigt.

Der Hauptpeak (HAM115_ES_30-34P2, 40 mg) wurde anschließend mit einer präparativen HPLC (Gradientensystem HPLC2) isokratisch weiter fraktioniert. Auf diese Weise konnten 24 mg von Verbindung **61** (HAM115_ES_30-34P2.2) isoliert werden.

2.11. Vergleich von AcorA mit multivariaten Methoden

In diesem Abschnitt werden multivariate Analysemethoden vergleichend zu AcorA dargestellt. Als Grundlage für diesen Vergleich dienten die Daten aus dem Proof of Concept Experiment. Durch Messungen der Antibiotikastammlösungen mit dem FT-ICR-Massenspektrometer (HAM061, LB02, S. 83 ff.) und anschließendem Peak-Picking (visuell) wurde die Grundmenge der für die verschiedenen Datenanalysemethoden zu findenden Peaks (True Positive, TP) festgelegt (HAM070, LB03, S. 82 ff.). Da während des automatischen Peak-Picking und Alignment Prozesses mithilfe von XCMS einige Peaks nicht annotiert wurden, standen für die Suchalgorithmen insgesamt 47 *m/z* -Werte der Antibiotikapeaks zur Verfügung (siehe Tabelle A im Anhang). Die Beurteilung der Prädiktionsgüte der verschiedenen Datenanalysemethoden erfolgte mithilfe von ROC-Kurven.

2.11.1. Receiver Operating Characteristics Analysen

Theoretischer Hintergrund Receiver Operating Characteristics Analysen (ROC-Analysen) werden zur Beurteilung der Qualität von Modellen und insbesondere im Bereich Metabolomics auch zur Klassifizierung von Biomarkern verwendet [297]. In dieser Arbeit dienten ROC-Analysen zur Beurteilung und Vergleich der Prädiktionsqualität verschiedener Analysemethoden. Basis für die Analyse ist die klassische Konfusionsmatrix:

- RP = richtig positiv, ein Antibiotikumpeak wird als solcher erkannt
- RN = richtig negativ, ein Nicht-Antibiotikumpeak wird als solcher erkannt
- FP = falsch positiv, ein Nicht-Antibiotikumpeak wird fälschlicherweise als Antibiotikumpeak erkannt
- FN = falsch negativ, ein Antibiotikumpeak wird fälschlicherweise nicht erkannt

Für die Beurteilung der Qualität einer Analysemethode sind Sensitivität (Richtig-Positiv-Rate, Trefferquote), Spezifität (Richtig-Negativ-Rate) und Genauigkeit (Precision) von

Tabelle 2.19.: Konfusionsmatrix

		Tatsächlich	
		Aktiv	Nicht-aktiv
Vohergesagt	Aktiv	TP	FP
	Nicht-Aktiv	FN	RN

besonderer Bedeutung:

$$Sensitivität = \frac{RP}{RP + FN}, Spezifität = \frac{RN}{RN + FP}, Genauigkeit = \frac{RP}{RP + FP} \quad (2.13)$$

Zur Klassifikation kann auch das F_1-Maß herangezogen werden [110, 298, 299, 300]. Es ergibt sich aus dem harmonischem Mittel von Genauigkeit und Trefferquote:

$$F_1 = \frac{2 \cdot Genauigkeit \cdot Trefferquote}{Genauigkeit + Trefferquote} \quad (2.14)$$

Die Falsch-Positiv-Rate ergibt sich als Komplement der Spezifität (1-Spezifität). Nach Sortierung der Daten anhand der jeweiligen Methode (z.B. Loadings, Regressionskoeffizienten, Gini-Index) und Berechnung von Sensitivität und Spezifität werden Richtig-Positiv-Rate und Falsch-Positiv-Rate in einer ROC-Kurve gegeneinander aufgetragen. Die Fläche unter der Kurve (Area Under the Curve (AUC)) gibt an, wie gut die Analysemethode zwischen Peaks von aktiven und inaktiven Verbindungen differenzieren kann. Je mehr Antibiotikapeaks in den oberen Bereichen der Hitliste angereichert werden, desto steiler verläuft die ROC-Kurve, da in diesem Fall die Sensitivität bei gleichzeitig geringer Falsch-Positiv-Rate (d. h. hoher Spezifität) zunimmt. Im Idealfall nähert sich die AUC einem Wert von 1. In dem hier betrachtetem Fall würde dies bedeuten, dass die ersten 47 Peaks in der Hitliste aus den gesuchten Antibiotikapeaks bestehen würde. Ein Wert von 0,5 stellt einen Zufallsprozess dar und entspricht somit dem Raten.
Für den Anwender sind insbesondere die ersten 50 bis 100 Peaks in einer nach Größe der jeweiligen Methode sortierten Ergebnisliste von entscheidender Bedeutung. Je mehr Peaks der aktiven Verbindungen in diesem Bereich angereichert werden, desto eher wird der Anwender in der Lage sein, die Informationen (z. B. mehrere Addukt- oder Isotopenpeaks einer Verbindung) zu amalgamieren und eine aktive Substanz als solche zu erkennen. In einer ROC-Kurve ist daher der Bereich niedriger Falsch-Positiv-Rate (= hohe Spezifität) von besonderem Interesse. Aus diesem Grunde wurde in dieser Arbeit für jede Analysemethode auch der partielle AUC-Wert (pAUC) im Bereich zwischen 0 und 5 % Falsch-Positiv-Rate

berechnet. Je höher der pAUC-Wert, desto spezifischer werden Peaks aktiver Verbindungen erkannt und oben in der Ergebnisliste angereichert.

Durchführung Sensitivität und 1-Spezifität wurden mit dem R-Paket ROCR [301] berechnet. Die Berechnung der AUC und pAUC-Werte sowie deren 95 % Konfidenzintervalle erfolgte mit dem R-Paket pROC [302]. Zur Ermittlung der Konfidenzintervalle der AUC-Werte verwendet das R-Paket eine Näherung auf Grundlage der Methode von DeLong [303]. Die 95 % Konfidenzintervalle der pAUC-Werte wurden durch 1000 Bootstrap Replikate berechnet. P-Werte, als Maß für den signifikanten Unterschied zwischen zwei ROC-Kurven, wurden mit einer Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ anhand von 2000 Bootstrap Replikaten berechnet. Die Erstellung der ROC Kurven erfolgte mit dem R-Paket ggplot2 [304].

2.11.2. Hauptkomponentenanalyse

Theoretischer Hintergrund Die Hauptkomponentenanalyse (Principal Component Analysis (PCA)) ist ein wichtiges Verfahren, das zur Reduktion hochdimensionaler Datensätze verwendet wird. Die Grundannahme der PCA ist, dass Variablen (z. B. Signale in Spektren) mit hoher Varianz einen vergleichsweise hohen „Informationsgehalt“ repräsentieren, wohingegen Variablen mit geringer Varianz - beispielsweise Rauschsignale in Spektren - einen vergleichsweise niedrigen „Informationsgehalt“ besitzen. Der Algorithmus der PCA sortiert einen vorliegenden Datensatz nach Informationsgehalt (Varianz) und fasst diese dann in den s. g. Hauptkomponenten (engl.: prinicpal components, PC) zusammen.
Dabei wird die Originaldatenmatrix $\mathbf{X}$ mit n Objekten und m Variablen durch eine Linearkombination aus a Hauptkomponenten, jeweils bestehend aus dem Produkt von Score- ($\mathbf{T}$) und Loadings-Matrix ($\mathbf{P}^\top$), ersetzt:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1 + \mathbf{t}_2\mathbf{p}_2 + \mathbf{t}_a\mathbf{p}_a \ldots \mathbf{t}_n\mathbf{p}_n \tag{2.15}$$

Anschaulich geben die Loadings $\mathbf{P}^\top$ den Beitrag einer Variablen x_i zur Richtung der Hauptkomponente an. D. h. je größer der Loadings-Wert einer Variable ist, desto stärker ist ihr Beitrag an der Gesamtvarianz der entsprechenden Hauptkomponente.
Die Scores $\mathbf{T}$ erhält man durch Projektion der x-Variablen auf die Achse der jeweiligen Hauptkomponente. Sie geben somit die Lage eines Objektes in dem neu konstruierten Koordinatensystem an.
Mathematisch führt die Berechnung der Score- und Loadings-Matrix auf die Bestimmung

der Eigenvektoren $\mathbf{v}$ und Eigenwerte $\boldsymbol{\lambda}$ der Matrix $\mathbf{X}$:

$$\mathbf{Xv} = \lambda\mathbf{v} \tag{2.16}$$

oder mit Hilfe der Einheitsmatrix $\mathbf{I}$ formuliert als lineares Gleichungssystem:

$$(\mathbf{X} - \lambda\mathbf{I})\mathbf{v} = 0 \quad \text{mit} \quad \mathbf{v} \neq 0 \tag{2.17}$$

Ein häufig genutztes Lösungsverfahren zur Berechnung der Score- und Loadings-Matrix ist die Singulärwertzerlegung (engl.: singular value decomposition, SVD):

$$\mathbf{X} = \mathbf{UDV}^{\mathsf{T}} = (\mathbf{UD})\mathbf{V}^{\mathsf{T}} = \mathbf{TP}^{\mathsf{T}} \tag{2.18}$$

Die Matrix $\mathbf{U}$ ist eine orthonormale Transformationsmatrix und enthält die orthogonalen, normierten Eigenvektoren der Kovarianzmatrix $\mathbf{XX}^{\mathsf{T}}$. Die Matrix $\mathbf{D}$ ist eine Diagonalmatrix und enthält auf ihrer Diagonalen die Quadratwurzel der Eigenwerte (Singulärwerte) der Kovarianzmatrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ der Größe nach geordnet. Je größer der Eigenwert bzw. Singulärwert ist, desto größer ist die Varianz des korrespondierenden Eigenvektors. Das Produkt aus $\mathbf{U}$ und $\mathbf{D}$, d. h. die Wichtung der normierten Eigenvektoren mit der in den Singulärwerten enthaltenen Varianz, ergibt die Score-Matrix $\mathbf{T}$. Diese Operation entspricht der Projektion der Objekte in das neue Koordinatensystem, dessen Achsen über die Loadings-Matrix determiniert werden. $\mathbf{T}$ enthält in ihren Spalten die Scores der einzelnen Hauptkomponenten. Die Matrix $\mathbf{V}^{\mathsf{T}}$ enthält die orthonormalen Eigenvektoren der Matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$. Sie ist identisch mit der Loadings-Matrix $\mathbf{P}^{\mathsf{T}}$.
Die in $\mathbf{D}$ enthaltenen Eigenwerte entsprechen der Varianz einer jeden Hauptkomponente:

$$Var_i = \frac{d^2}{n-1} \tag{2.19}$$

Der Anteil der erklärten Varianz für eine Hauptkomponente a_i lässt sich somit berechnen über:

$$EV_i = \frac{Var_i}{\sum_{j=1}^{a} Var_j} \tag{2.20}$$

Eine weitere Möglichkeit zur Berechnung der Hauptkompontenen ist der NIPALS Algorithmus (**N**onlinear **I**terative **Par**tial **L**east **S**quares), der von Herman Wold eingeführt wurde. Sukzessiv werden die einzelnen Hauptkomponenten mit dem Ziel berechnet, die maximale

Varianz der x-Variablen auszuschöpfen, wobei die Hauptkomponenten zueinander orthogonal sein müssen. D. h. zunächst wird die erklärte Varianz für die Hauptkomponente PC1 maximiert und eine Restvarianz $\mathbf{E}_1$ berechnet:

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1 \tag{2.21}$$

Anschließend wird für $\mathbf{E}_1$ wieder die maximale Varianz in den x-Variablen gesucht, mit der Bedingung, dass diese Hauptkomponente PC2 orthogonal zu PC1 ist.

$$\mathbf{E}_2 = \mathbf{E}_1 - \mathbf{t}_2\mathbf{p}_2 \tag{2.22}$$

Dieser Prozess wird fortgeführt, bis maximal m-1 Hauptkomponenten berechnet sind.
Praktisch wird zunächst aus der Datenmatrix $\mathbf{X}$ die Spalte mit der höchsten Varianz als erste Schätzung für den Scorevektor $\mathbf{t}_a$ verwendet. Nun wird $\mathbf{X}$ auf den Scorevektor $\mathbf{t}_a$ projiziert (2.23) und normiert (2.24):

$$\mathbf{p}'_a = \frac{\mathbf{X}_a^\top \mathbf{t}_a}{\mathbf{t}_a^\top \mathbf{t}_a} \tag{2.23}$$

$$\mathbf{p}_a = \frac{\mathbf{p}'_a}{\| \mathbf{p}'_a \|} \tag{2.24}$$

Da $\mathbf{t}_a$ anfangs nur geschätzt war, wird nun $\mathbf{X}$ auf den zuvor berechneten Loadingsvektor $\mathbf{p}_a$ projiziert:

$$\mathbf{t}_a = \frac{\mathbf{X}\mathbf{p}_a}{\mathbf{p}_a^\top \mathbf{p}_a} \tag{2.25}$$

Nun wird der neue Scorevektor $\mathbf{t}_a$ mit dem alten Scorevektor verglichen. Ist $\mathbf{t}_{aneu} \neq \mathbf{t}_{aalt}$ wird die Iteration ab 2.23 weitergeführt. Sind beide identisch (oder konvergieren gegen einen festgelegten Wert), ist die Iteration für diese Hauptkomponente abgeschlossen. Die entsprechenden Scores und Loadings werden gespeichert und wie oben beschrieben aus der Datenmatrix $\mathbf{X}_a$ entfernt:

$$\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}_a\mathbf{p}_a^\top \tag{2.26}$$

Durchführung Für die Hauptkomponentenanalyse wurden für den POC Datensatz zwei verschiedene Präprozessierungsmethoden verwendet. Zum einen wurde eine in der Metabolomics häufig verwendete z-Transformation (Gl. 23) der Daten durchgeführt. In der zweiten Variante wurde lediglich eine Mittenzentrierung (Gl. 18) der Daten durchgeführt. Die Hauptkomponentenanalyse erfolgte mit dem R-Paket pcaMethods [305] unter Verwendung des NIPALS Algorithmus.

```
result.pca <- pca(data[,start:end], method = "nipals", nPcs = 3 ,scale
    = scaling.method, center = TRUE, na.rm = TRUE)
```

2.11.3. Hauptkomponentenregression (PCR)

Theoretischer Hintergrund Die Hauptkomponentenregression (engl.: principal component regression, PCR) kombiniert die PCA mit der multiplen linearen Regression. Dazu wird zunächst eine Hauptkomponentenanalyse durchgeführt und die Score- und Loadings-Matrizen der jeweiligen Hauptkomponente bestimmt. Sind Scores und Loadings bekannt, kann Gleichung 2.15 wie folgt umgestellt werden:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^\top + \mathbf{E} \\ \mathbf{T}\mathbf{P}^\top(\mathbf{P}^T)^{-1} &= \mathbf{X}(\mathbf{P}^T)^{-1} + \mathbf{E} \\ \mathbf{T} &= \mathbf{X}\mathbf{P} + \mathbf{E} \end{aligned} \tag{2.27}$$

D. h. die Score Matrix $\mathbf{T}$ kann als Repräsentant für die ursprüngliche Datenmatrix $\mathbf{X}$ verwendet werden. Analog zur Regressionsgleichung (Gl. 30) können die Regressionskoeffizienten $\mathbf{q}$ für die Scores berechnet werden [306]:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon} = \mathbf{T}(\mathbf{P}^\top\mathbf{b}) + \boldsymbol{\epsilon} = \mathbf{T}\mathbf{q} + \boldsymbol{\epsilon} \tag{2.28}$$

Durch die Orthogonalität - d. h. lineare Unabhängigkeit - der Scores, werden Multikollinearitäten vermieden. Zudem werden nur Hauptkomponenten mit hohem Informationsgehalt, d. h. mit einem hohen Anteil an erklärter Varianz in $\mathbf{X}$, für die Regression verwendet. Die Anzahl der zu verwendenden Hauptkomponenten muss durch Kreuzvalidierung bestimmt werden. Die Berechnung der Regressionskoeffizienten für die Scores erfolgt, analog zur linearen Regression, über die Methode der kleinsten Fehlerquadrate:

$$\mathbf{q} = (\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top\mathbf{y} \tag{2.29}$$

Mit Hilfe von Gleichung 2.30 kann anschließend auf die Regressionskoeffizienten der Ausgangsmatrix $\mathbf{X}$ zurückgerechnet werden:

$$\mathbf{b} = \mathbf{Pq} = \mathbf{P}(\mathbf{T}^\top\mathbf{T})^{-1}\mathbf{T}^\top\mathbf{y} \tag{2.30}$$

Normalerweise wird bei der Berechnung der PCR von einer mittenzentrierten Matrix ausgegangen. Dadurch wird das Regressionsmodell ohne den Ordinatenabschnitt berechnet und somit fehlt $\mathbf{b}_0$ im Koeffizientenvektor $\mathbf{B}$. Die Berechnung von $\mathbf{b}_0$ erfolgt mit Gleichung 2.31:

$$\mathbf{b}_0 = \bar{\mathbf{y}} - \mathbf{XB} \tag{2.31}$$

Wie oben beschrieben, werden durch Verwendung der Hauptkomponenten die x-Variablen mit der höchsten Varianz für die Vorhersage von $\mathbf{y}$ verwendet. Diese Variablen beschreiben zwar in idealer Weise die Varianz in $\mathbf{X}$. Es gibt jedoch *a priori* keinen Grund, weshalb gerade diese Variablen für die Vorhersage von $\mathbf{y}$ geeignet sein sollten, da bei der Berechnung der Hauptkomponenten kein Bezug auf $\mathbf{y}$ genommen wird. Eine Lösung dieses Problems bietet die Partial-Least-Squares Regression.

Durchführung Die Modellbildung- und Validierung erfolgte über eine 100-fache doppelte Kreuzvalidierung nach Filzmoser [4]. Die Prozedur ist in Algorithmus 1 wiedergegeben. Für die Kreuzvalidierung wurde das R-Paket *Chemometrics* verwendet [307]. Zunächst wurde der Datensatz nach Gleichung 18 zentriert. Anschließend erfolgte die Bestimmung der optimalen Anzahl der Hauptkomponenten über eine 4-fache Kreuzvalidierung in der inneren Schleife (*segments0*). Die Abschätzung des Testfehlers erfolgte über eine äußere Schleife mit 7 Segmenten (*segments*). Diese Prozedur wurde 100 mal wiederholt. Für die Berechnung von jeweils *ncomp* = 10 Hauptkomponenten wurde das Verfahren der Singulärwertzerlegung verwendet.

```
1 res.mvrPCR <- mvr_dcv(y ~ ., data = data.set, ncomp = 10, method = "
    svdpc", repl = 100, segments = 7, segments0 = 4, sdfact=1, plot.opt
    = F)
```

Algorithmus 1 r-fache doppelte Kreuzvalidierung [4]

Require: Datensatz $data$ mit n Objekten und p Variablen

$r_{max} \leftarrow$ Anzahl der Wiederholungen
$a_{max} \leftarrow$ Anzahl der zu testenden Hauptkomponenten
for r = 1 to r_{max} **do**
    Generiere aus dem Datensatz $data$ zufällig k Testdatensätze $test_1, test_i, ...test_k$
    for i = 1 to k **do**
        Kalibrationsdatensatz $kalib$:= $data$ - $test_i$
        Generiere aus $kalib$ zufällig m Validierungsdatensätze $valid_1, valid_j, ...valid_m$
        for j = 1 to m **do**
            Trainingsdatensatz $train$:= $kalib$ - $valid_j$
            Trainiere das Modell mit 1... a_{max} Hauptkomponenten auf $train$
            Berechne den MSECV auf $valid_j$
        end for
        a_{opt} = a[min(MSECV) + 1 Standardfehler]
        Erstelle das Modell mit a_{opt} auf $kalib$
        Berechne MSEP des Modells auf $test_i$
    end for
    Berechne den Standard Error of Prediction (SEP)
end for
a_{final} = max (Häufigkeitsverteilung(a_{opt}))

2.11.4. Partial-Least-Squares Regression (PLSR)

Theoretischer Hintergrund Die Partial-Least-Squares Regression von Hermann Wold ist eine weitere Methode zur multivariaten Regression hochdimensionaler Datensätze. Ähnlich der PCR wird die Matrix der Originalvariablen (**X**) in der Regressionsgleichung (30) durch eine Score-Matrix **T** ersetzt. Im Gegensatz zur PCR fließen jedoch Informationen aus der y-Variablen (Bioaktivität) in die Berechnung der Scores und Loadings mit ein.
Die Berechnung der Hauptkomponenten erfolgt häufig nach dem NIPALS-Algorithmus (**N**onlinear **I**terative **P**artial **L**east **S**quares) von H. Wold. Weitere Algorithmen wie z.B. SIMPLS [308] und Kernel-PLS [309] sind jedoch ebenfalls verfügbar. Für den Fall eines einzigen Regressandenvektors zeigte de Jong, dass NIPALS und SIMPLS identische Ergebnisse liefern [308].
Im allgemeinen Fall (PLS2), in dem sowohl die abhängigen als auch die unabhängigen Variablen aus multivariaten Datensätzen bestehen, wird eine Hauptkomponentenanalyse sowohl der **X**- (Gl. 2.32) als auch der **Y**-Matrix (Gl. 2.33) durchgeführt. Über einen Wichtungsvektor wird dabei die Beziehung zu der jeweils anderen Matrix hergestellt:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\top} + \mathbf{E} = \sum \mathbf{t}\mathbf{p}^{\top} + \mathbf{E} \tag{2.32}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\top} + \mathbf{F} = \sum \mathbf{u}\mathbf{q}^{\top} + \mathbf{F} \tag{2.33}$$

Der Wichtungsvektor **w** wird dabei so gewählt, dass die Kovarianz zwischen X und Y maximiert wird:

$$\mathbf{W} = max(cov(\mathbf{X}, \mathbf{Y})) \tag{2.34}$$

Für den in dieser Arbeit betrachteten Fall, dass die abhängige Variable lediglich aus einem univariaten Variablenvektor **y** besteht, ist die Hauptkomponentenanalyse von **y** (Gl. 2.33) nicht notwendig. Analog zu Kessler [310] kann daher der s. g. PLS1 NIPALS-Algorithmus wie folgt beschrieben werden:
Ausgehend von einer mittenzentrierten Matrix **X** erfolgt zunächst eine Regression der **X**-Matrix auf den **y**-Vektor. Der Index a beschreibt die Nummer der PLS-Komponente:

$$\mathbf{X}_a = \mathbf{y}_a \mathbf{w}_a^{\top} + \mathbf{E} \tag{2.35}$$

Die Least Squares Lösung der Gleichung 2.35 lautet:

$$\mathbf{w}'_a = \mathbf{X}_a^\top \mathbf{y}_a \tag{2.36}$$

Anschließend wird der Wichtungsvektor $\mathbf{w}$ normiert:

$$\mathbf{w}_a = \frac{\mathbf{w}'_a}{\|\mathbf{w}'_a\|} \tag{2.37}$$

Um den Score-Vektor $\mathbf{t}$ zu erhalten, erfolgt im nächsten Schritt eine Regression von $\mathbf{X}$ auf den soeben berechneten Wichtungsvektor $\mathbf{w}_a$:

$$\mathbf{X}_a = \mathbf{t}_a \mathbf{w}_a^\top + \mathbf{E} \tag{2.38}$$

Die Berechnung erfolgt wieder über das Least Squares Verfahren:

$$\mathbf{t}_a = \frac{\mathbf{X}_a \mathbf{w}_a}{\mathbf{w}_a^\top \mathbf{w}_a} \tag{2.39}$$

Nun können die Loadings $\mathbf{p}_a$ von $\mathbf{X}$ mit Hilfe der Scores $\mathbf{t}_a$ nach folgendem Modell berechnet werden:

$$\mathbf{X}_a = \mathbf{t}_a \mathbf{p}_a^\top + \mathbf{E} \tag{2.40}$$

Die Least Squares Lösung von Gleichung 2.40 lautet:

$$\mathbf{p}_a = \frac{\mathbf{X}_a^\top \mathbf{t}_a}{\mathbf{t}_a^\top \mathbf{t}_a} \tag{2.41}$$

In ähnlicher Weise werden anschließend die Loadings $\mathbf{q}_a$ des $\mathbf{y}$-Vektors berechnet:

$$\mathbf{q}_a = \frac{\mathbf{t}_a^\top \mathbf{y}_a}{\mathbf{t}_a^\top \mathbf{t}_a} \tag{2.42}$$

Die zuvor berechneten Vektoren $\mathbf{w}$, $\mathbf{p}$, $\mathbf{q}$ und $\mathbf{t}$ werden nun in den Matrizen $\mathbf{W}$ (W-Loadings von $\mathbf{X}$), $\mathbf{P}$ (P-Loadings von $\mathbf{X}$), $\mathbf{Q}$ (Y-Loadings) und $\mathbf{T}$ (Scores von $\mathbf{X}$) gespeichert. Anschließend erfolgt - ähnlich der PCA - die Deflation der Matrix $\mathbf{X}_a$ sowie des Vektors $\mathbf{y}_a$. Die oben beschriebenen Schritte werden wiederholt bis die gewünschte Anzahl von PLS-Komponenten berechnet ist (maximal m-1 PLS-Komponenten):

$$\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}_a \mathbf{p}_a^\top \tag{2.43}$$

$$\mathbf{y}_{a+1} = \mathbf{y}_a - \mathbf{q}_a\mathbf{t}_a \tag{2.44}$$

Die Berechnung der Regressionskoeffizienten erfolgt über

$$\mathbf{b} = \mathbf{W}(\mathbf{P}^\top\mathbf{W})^{-1}\mathbf{q} \tag{2.45}$$

Durchführung Die Modellbildung- und Validierung erfolgte über eine 100-fache doppelte Kreuzvalidierung nach Filzmoser [4]. Das Verfahren ist in Algorithmus 1 (Seite 78) wiedergegeben. Die Kreuzvalidierung erfolgte mit dem R-Paket *Chemometrics* [307]. Zunächst wurde der Datensatz nach Gleichung 18 zentriert. Die Bestimmung der optimalen Anzahl der latenten Variablen erfolgte über eine 4-fache Kreuzvalidierung in der inneren Schleife (*segments0*). Die Abschätzung des Testfehlers erfolgte über eine äußere Schleife mit 7 Segmenten (*segments*). Für die Berechnung von jeweils *ncomp* = 10 latenten Variablen wurde das Verfahren der „SIMPLS“ Algorithmus verwendet [308].

```
1 res.mvrPLSR<- mvr_dcv(y~., data=data.set, ncomp=10, method="simpls", repl
    =100, segments = 7, segments0 = 4, sdfact=1, plot.opt = F)
```

2.11.4.1. Variable Importance in Projection (VIP)

Theoretischer Hintergrund Eine verbreitete Methode zur Variablenselektion in der PLSR ist die von Wold eingeführte Variable Importance in Projection (VIP) [151]. Die VIP spiegelt den Anteil der erklärten Varianz in $\mathbf{X}$ ($SS(q_a t_a) = \mathbf{q}^2\mathbf{t}_a^\top\mathbf{t}_a$) gewichtet durch die Kovarianz zwischen $\mathbf{X}$ und $\mathbf{y}$ ($\mathbf{w}_{aj} = \frac{w_{aj}}{||w_a||}$) für jede Variable j im Verhältnis zur Gesamtvarianz wider:

$$VIP_j = \sqrt{\frac{m\sum_{a=1}^{h}\left(SS(q_a\mathbf{t}_a)\left(\frac{w_{aj}}{||w_a||}\right)^2\right)}{\sum_{a=1}^{h}SS(q_a\mathbf{t}_a)}} \tag{2.46}$$

Die Variable b_a repräsentiert dabei die Orthogonalprojektion von $\mathbf{y}$ auf den Scorevektor $\mathbf{t_a}$ für jede latente Variable a:

$$q_a = \frac{\mathbf{t}_a^\top\mathbf{y}_a}{\mathbf{t}_a^\top\mathbf{t}_a} \tag{2.47}$$

Im Mittel ist der VIP-Score gleich eins. Daher werden Variablen mit einem VIP-Score > 1 als relevant betrachtet. Zudem haben Thévenot *et al.* einen Zusammenhang zwischen der

univariaten False Discovery Rate (FDR) und dem multivariaten VIP-Score in der OPLS-Regression aufgezeigt [311]. Demzufolge entspricht ein VIP-Score > 1 approximativ einer FDR < 0.05.

Durchführung Nach erfolgter Partial-Least-Squares Regression wurde die VIP für die ersten beiden latenten Variablen mit dem folgenden Skript aus dem R-Paket *pls* [312] durchgeführt:

```
VIP <- function(object) {
    SS <- c(object$Yloadings)^2 * colSums(object$scores^2)
    Wnorm2 <- colSums(object$loading.weights^2)
    SSW <- sweep(object$loading.weights^2, 2, SS / Wnorm2, "*")
    sqrt(nrow(SSW) * apply(SSW, 1, cumsum) / cumsum(SS))
}
```

Die Berechnung des mittleren Vorhersagefehlers des PLSR-Submodells (nach Selektion der relevanten Variablen mit einem VIP-Score > 1) erfolgte wieder mit der in Algorithmus 1 beschriebenen 100-fachen doppelten Kreuzvalidierung.

2.11.5. Quantitative Pattern Activity Relationship (QPAR)

Theoretischer Hintergrund Die Quantitative Pattern Activity Relationship (QPAR) Methode wurde 2009 von Chau publiziert [160]. Basis dieser Methode ist die von Kvalheim eingeführte Target Projection (TP) [153, 156, 157]. Ähnlich der orthogonalen Partial-Least-Squares Regression (OPLS) [141] soll durch die Target Projection im Idealfall ausschließlich die mit $\mathbf{y}$ assoziierte Variation in den x-Variablen aufgedeckt werden. Durch beide Verfahren wird die für $\mathbf{y}$ relevante Information aus $\mathbf{X}$ in einer einzelnen latenten Variable gebündelt. Da die relevanten Informationen nun nicht mehr über mehrere latente Variablen verteilt werden, ist das Modell auf diese Weise leichter interpretierbar.
Während bei der OPLS die Originalmatrix in eine Matrix mit der orthogonalen Information $\mathbf{X}_{\mathrm{Ortho}}$ und in eine Prädiktionsmatrix $\mathbf{X}_{\mathrm{Pred}}$ dekomponiert wird (Gleichung 2.48), verwendet die Target Projection den Regressionskoeffizientenvektor $\mathbf{b}$ aus der PLS Regression zur Berechnung der TP-Scores $\mathbf{t}_{\mathrm{TP}}$ und TP-Loadings $\mathbf{p}_{\mathrm{TP}}^{\mathrm{T}}$ (Gleichung 2.49).

$$\mathbf{X}_{\mathrm{OPLS}} = \mathbf{X}_{\mathrm{ortho}} + \mathbf{X}_{\mathrm{pred}} + \mathbf{E}_{\mathrm{OPLS}} = \mathbf{T}_{\mathrm{ortho}}\mathbf{P}_{\mathrm{ortho}}^{\mathrm{T}} + \mathbf{t}_{\mathrm{pred}}\mathbf{p}_{\mathrm{pred}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{OPLS}} \tag{2.48}$$

$$\mathbf{X}_{TP} = \mathbf{t}_{TP} + \mathbf{p}_{TP}^{T} + \mathbf{E}_{TP} \tag{2.49}$$

In einem ersten Schritt der Target Projection werden zunächst die Regressionskoeffzienten normalisiert:

$$\mathbf{w}_{TP} = \frac{\mathbf{b}_{PLS}}{\|\mathbf{b}_{PLS}\|} \tag{2.50}$$

Um die Target Projected Scores $\mathbf{t_{TP}}$ zu erhalten, werden anschließend die normalisierten Regressionskoeffizienten auf die Matrix $\mathbf{X}$ projiziert:

$$\mathbf{t}_{TP} = \mathbf{X}\mathbf{w}_{TP} = \frac{\mathbf{X}\mathbf{b}_{PLS}}{\|\mathbf{b}_{PLS}\|} = \frac{\hat{\mathbf{y}}}{\|\mathbf{b}_{PLS}\|} \tag{2.51}$$

Der TP-Score Vektor ist proportional zu $\hat{\mathbf{y}}$ und somit maximal mit der vorhergesagten Bioaktivität korreliert [158, 156]. Die Target Projected Loadings $\mathbf{p_{TP}}$ werden anschließend über die Projektion der $\mathbf{t_{TP}}$ auf die Matrix $\mathbf{X}$ erhalten:

$$\mathbf{p}_{TP}^{T} = \frac{\mathbf{t}_{TP}^{T}\mathbf{X}}{\mathbf{t}_{TP}^{T}\mathbf{t}_{TP}} = \frac{\frac{\mathbf{b}_{PLS}^{T}}{\|\mathbf{b}_{PLS}\|}(\mathbf{X}^{T}\mathbf{X})}{\|\mathbf{t}_{TP}\|^{2}} \tag{2.52}$$

Die TP-Loadings sind proportional zu den normalisierten Regressionkoeffizienten und der Varianz/Kovarianz der Daten im $\mathbf{X}$-Raum. Ein hoher TP-Loading Wert wird somit erhalten, wenn eine Variable sowohl über eine hohe Varianz in $\mathbf{X}$ als auch über einen großen Regressionskoeffizienten, d. h. einen großen Beitrag zur Erklärung der Variation in $\mathbf{y}$, verfügt [158]. Da für die Berechnung der TP-Loadings die Varianz/Kovarianz Matrix verwendet wird, sind die TP-Loadings nicht skaleninvariant. D. h. Peaks mit hoher Intensität und geringer Korrelation mit $\mathbf{y}$ werden gegenüber Peaks geringerer Intensität und hoher Assoziation mit der biologischen Aktivität bevorzugt. Um dieses Problem zu entschärfen hat Rajalahti das s. g. Selectivity Ratio eingeführt [159]. Das Selectivity Ratio beschreibt das Verhältnis der erklärten Varianz (SS_{expl}) zur Restvarianz (SS_{resid}) einer Variable x_i in der Target Projection:

$$SR_i = \frac{SS_{expl,i}}{SS_{resid,i}} \tag{2.53}$$

TP-Loadings mit einer hohen erklärten Varianz weisen somit einen höheren Bezug zur Bioaktivität auf, als Variablen mit niedriger erklärter Varianz. Erklärte Varianz und Restvarianz

können wie folgt berechnet werden:

$$SS_{expl,i} = \|\mathbf{t}_{TP}\mathbf{p}_{TPi}^{T}\|^2 \qquad SS_{resid,i} = \mathbf{e}_{TPi} \tag{2.54}$$

Um eine kritische Grenze für die über die Selectivity Ratios gewichteten TP-Loadings zu erhalten, verwendet Rajalahti einen F-Test mit n-3 und n-2 Freiheitsgeraden [159]:

$$F_{calc} = SR_i > F_{crit} = F_{\alpha,N-2,N-3} \tag{2.55}$$

Durchführung Als Grundlage für QPAR dienten die Ergebnisse der PLSR. Die Berechnung der Target Projection und der Selectivity Ratios erfolgte somit mit der zweiten latenten Variablen LV2 anhand des folgenden R-Skripts. Das Skript wurde von A. Gohr geschrieben und anhand des Originaldatensatzes aus [160] validiert.

```
rts<-as.numeric(colnames(X))
# Target Projection
# X = X-Matrix
# b = Regressionskoeffizient aus PLSR

t_TP <- ( X %*% b ) / sqrt(sum(b*b))
p_TP <- ( t(X) %*% t_TP ) / sum(t_TP*t_TP)

X_TP <- t_TP %*% t(p_TP)
E_TP <- X - X_TP

# selectivity ratios
# explained variance / residual variance
explained_variances <- apply((t(t(X_TP) - apply(X,2,mean)))^2,2,sum)
residual_variances  <- apply((E_TP)^2,2,sum)
total_variances     <- apply((t(t(X) - apply(X,2,mean)))^2,2,sum)
selectivity_ratios  <- explained_variances / residual_variances
selectivity_ratios_signed <-selectivity_ratios * sign(p_TP)
```

Zur Bestimmung eines kritischen Grenzwertes F_{crit} wurde ein F-Test mit $\alpha = 5\%$ verwendet.

```
F.crit<-qf(0.95,df1=nrow(X)-2,df2=nrow(X)-3)
```

Die Berechnung des mittleren Vorhersagefehlers des QPAR-Submodells (nach Selektion der relevanten Variablen mit einem Selectivity Ratio > F.crit) erfolgte mit der in Algorithmus 1 beschriebenen 100-fachen doppelten Kreuzvalidierung.

2.11.6. Regularisierungsmethoden

Theoretischer Hintergrund Unter der Annahme orthogonaler (linear unabhängiger) x-Variablen ist der Least-Squares-Schätzer der beste unverzerrte Schätzer für β [2]. Insbesondere bei massenspektrometrischen Datensätzen können jedoch aufgrund der vielen Massensignale zufällige Kollinearitäten auftreten. Im Extremfall, d. h. sind zwei oder mehrere Variablen zu 100 % untereinander korreliert, wird die Kovarianzmatrix singulär, sodass dessen Inverse nicht berechnet werden kann. Weiterhin führen hohe Korrelationen innerhalb der x-Variablen dazu, dass die Regressionskoeffizienten nur sehr ungenau, d. h. mit einer hohen Varianz, sowie tendenziell zu hoch geschätzt werden (siehe auch Gl. 38, [165]). Nach Gleichung 43 kann eine Reduktion des mittleren Vorhersagefehlers (MSE) erreicht werden, indem die Modellvarianz auf Kosten der Erwartungstreue (erhöhter Bias) reduziert wird. In der Praxis kann dies durch Schrumpfung der Regressionkoeffizienten in Richtung Null erreicht werden. Dieses Konzept ist in den s. g. Regularisierungsmethoden (englisch auch Shrinkage methods) verwirklicht, zu denen die Ridge Regression, Lasso und Elastic Net zählen.

Ridge Regression Die Ridge Regression wurde 1970 von Hoerl und Kennard eingeführt, um $(\mathbf{X}^\top\mathbf{X})^{-1}$ auch im Falle hoher Multikollinearität berechnen zu können [165, 313]. Dies geschieht durch Einführung eines konstanten Terms $\lambda\mathbf{I}$ zum LS-Schätzer:

$$\mathbf{b}_{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \tag{2.56}$$

Die Addition eines Terms $\lambda \geq 0$ zur Hauptdiagonalen verhindert, dass die Matrix $\mathbf{X}^\top\mathbf{X}$ auch bei nicht vollem Rang singulär wird. Auf diese Weise bleibt $\mathbf{X}^\top\mathbf{X}$ invertierbar. In der Lagrange Form lässt sich Gleichung 2.56 folgendermaßen schreiben:

$$b_{ridge} = \min_b \left\{ \sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{p} x_{ij}b_j)^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\} = RSS + \lambda \|b\|^2 \tag{2.57}$$

Man erkennt, dass RSS mit der Nebenbedingung eines quadratischen Penalisierungsterms (ℓ_2 Norm) minimiert wird. Je größer der Faktor λ, desto größer wird der Penalisierungsterm und desto stärker werden die Regressionskoeffizienten in Richtung Null reduziert. In Form

der Gleichung 2.58 wird dies noch deutlicher:

$$\min_b \left\{ \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} x_{ij} b_j)^2 \right\} \quad \text{in Hinblick auf} \quad \sum_{j=1}^{p} b_j^2 \leq t \tag{2.58}$$

Der Parameter t kann als Budget-Parameter angesehen werden, der eine Begrenzung für $\sum_{j=1}^{p} b_j^2$ zur Minimierung der RSS darstellt [314]. Bei großem t ist die Beschränkung permissiv, d. h. die Regressionskoeffzienten werden nur wenig penalisiert. Wird t sehr klein gewählt, müssen die Regressionskoeffizienten ebenfalls klein sein, um innerhalb des Budget-Parameters t zu bleiben. Eine interessante Eigenschaft der Ridge Regression ist, dass die Regressionskoeffizienten untereinander hoch korrelierter x-Variablen aufeinander zu geschrumpft werden [315, 316]. Übertragen auf die Situation von Massenspektren bedeutet dies, dass Isotopen- und Adduktpeaks einer Verbindung ähnlich hohe Regressionskoeffizienten erhalten sollten. Da die Regressionskoeffizienten in der Ridge Regression niemals vollständig auf Null gesetzt werden, verfügt die Ridge Regression über keine intrinsische Variablenselektion.

Lasso Der Least Absolute Shrinkage And Selection Operator (Lasso) wurde 1996 von Tibshirani eingeführt und stellt eine Modifikation der Ridge Regression dar [167]. Gleichung 2.59 zeigt die Minimierung der RSS anhand des Lasso Operators in der Lagrange Form:

$$b_{lasso} = \min_b \left\{ \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} x_{ij} b_j)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\} \tag{2.59}$$

Anstatt der ℓ_2 Norm verwendet Lasso die s.g. ℓ_1 Norm $\lambda \sum_{j=1}^{p} |b_j| = \lambda \|b\|_1$ zur Penalisierung der Regressionskoeffizienten. Gleichung 2.59 kann äquivalent zu Gl. 2.58 auch wie folgt formuliert werden:

$$\min_b \left\{ \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} x_{ij} b_j)^2 \right\} \quad \text{in Hinblick auf} \quad \sum_{j=1}^{p} |b_j| \leq t \tag{2.60}$$

Ähnlich der Ridge Regression werden die Regressionskoeffizienten in Abhängigkeit des Budget-Parameters t (bzw. in Abhängigkeit von λ in Gleichung 2.59) in Richtung Null geschrumpft. Lasso hat jedoch zusätzlich die Eigenschaft, dass Regressionskoeffizienten direkt auf Null gesetzt werden, sobald sie einen kritischen Wert unter- bzw. überschreiten. Dies hat den Effekt, dass Variablen mit geringer Erklärungskraft in Hinblick auf den

Regressanden aus dem Modell ausselektiert werden. Das Modell wird somit leichter interpretierbar.
Im Gegensatz zur Ridge Regression fehlt der Lasso Analyse der Gruppierungseffekt, d. h. die Koeffzienten untereinander hoch korrelierter Variablen werden nicht aufeinander zu geschrumpft. Stattdessen selektiert Lasso aus einer Gruppe von korrelierten Variablen willkürlich eine Variable aus und setzt die anderen auf Null [316, 317]. Übertragen auf die Situation in Massenspektren, bedeutet dies, dass aus einer Gruppe von Isotopen- und Adduktpeaks, jeweils nur ein Peak willkürlich selektiert würde. Maximal können mit der Lasso Analyse n Variablen im Model verbleiben [128]. Es werden somit sehr sparsame Modelle erhalten.

Elastic Net Das Elastic Net von Zou und Hastie [316] kann als ein Hybrid aus Ridge Regression und Lasso angesehen werden. Wie Gleichung 2.61 zeigt, enthält der Penalisierungsterm im Elastic Net eine Lasso (ℓ_1 Norm) und eine Ridge Regression (ℓ_2 Norm) Komponente. Über einen Parameter $\alpha \in [0, 1]$ können ℓ_1 und ℓ_2 Norm in einem zu optimierenden Verhältnis zur Penalisierung des KQ-Terms verwendet werden.

$$b_{elnet} = \min_b \left\{ \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} x_{ij} b_j)^2 + \lambda \sum_{j=1}^{p} \frac{1}{2}(1-\alpha) b_j^2 + \alpha |b_j|) \right\} \tag{2.61}$$

Die Elastic Net Analyse verknüpft die Vorteile aus Ridge Regression und Lasso. Zum einen verfügt es - ähnlich der Ridge Regression - über einen Gruppierungseffekt, d. h. die Differenz der Regressionskoeffizienten zweier hoch korrelierter Variablen x_j und x_k konvergiert gegen Null [316, 317]). Zum anderen verfügt das Elastic Net über eine intrinsische Variablenselektion, so dass leicht interpretierbare Modelle entstehen. Im Gegensatz zum Lasso besteht jedoch keine Einschränkung gegenüber der maximalen Anzahl an Variablen, die im Modell verbleiben können.

Durchführung Für alle drei Regularisierungsmethoden wurde das R-Paket *glmnet* verwendet [315]. Da nach Gleichung 2.61 der Anteil von Ridge Regression ($\alpha = 0$), Lasso ($\alpha = 1$) und Elastic Net ($0 < \alpha < 1$) über die Größe von α determiniert wird, können alle drei Regularisierungsmethoden in einem Durchlauf mit Hilfe einer Schleife, in der α von 0 bis 1 variiert wird, ausgeführt werden. Als Resultat erhält man eine Datenmatrix, die die Ergebnisse aller drei Methoden enthält. Der optimale Verhältnis zwischen Ridge Regression

und Lasso in Elastic Net wurde, ebenso wie die Optimierung des Penalisierungsparameters λ, über eine 100-fache doppelte Kreuzvalidierung bestimmt. Der Ablauf dieser Prozedur ist in Algorithmus 2 wiedergegeben. Zunächst wurde der Datensatz z-transformiert. Die Bestimmung des optimalen Penalisierungsfaktors λ erfolgte für jedes α über eine 6-fache Kreuzvalidierung in der inneren Schleife (*nfolds* = 6). Die Abschätzung des Testfehlers erfolgte über eine äußere Schleife mit k = 10 Segmenten. Das Paket *glmnet* erlaubt über die Option „lower.limits“ eine untere Grenze für die Regressionskoeffizienten festzulegen. Da für die Korrelation zwischen Peakintensität und Bioaktivität nur positive Korrelationen relevant sind, wurde „lower.limits=0“ gesetzt.

Algorithmus 2 r-fache doppelte Kreuzvalidierung für Shrinkage Methoden

Require: Datensatz $data$ mit n Objekten und p Variablen
1: $r_{max} \leftarrow$ Anzahl der Wiederholungen
2: $\lambda \leftarrow$ Vektor mit 100 Werten für Penalisierungsfaktor λ
3: $\alpha \leftarrow$ Vektor mit $\alpha = 0, 0.1, ...1.$
4: **for** α = 0 to 1 **do**
5: **for** r = 1 to r_{max} **do**
6: Generiere aus dem Datensatz $data$ zufällig
7: k Testdatensätze $test_1, test_i, ...test_k$
8: **for** i = 1 to k **do**
9: Kalibrationsdatensatz $kalib$:= $data$ - $test_i$
10: Generiere aus $kalib$ zufällig
11: m Validierungsdatensätze $valid_1, valid_j, ...valid_m$
12: **for** j = 1 to m **do**
13: Trainingsdatensatz $train$:= $kalib$ - $valid_j$
14: Trainiere das Modell mit $\lambda_1, \lambda_l \cdots \lambda_{100}$ auf $train$
15: Berechne für jedes λ_l den MSECV auf $valid_j$
16: **end for**
17: λ_{optCV} = λ[min(MSECV) + 1 Standardfehler]
18: Erstelle das Modell mit λ_{optCV} auf $kalib$
19: Berechne MSEP des Modells auf $test_i$
20: **end for**
21: **end for**
22: **end for**

Als Ergebnis aus Algorithmus 2 erhält man ein Array, das zu jedem α eine Menge aus λ_{optCV}-Werten inklusive deren korrespondierenden MSEPs enthält. Das Verfahren zur Ermittlung der optimalen λ-Werte für Ridge Regression, Lasso und Elastic Net, sowie zur Ermittlung

des optimalen Wertes für α in Elastic Net ist in Algorithmus 3 angegeben.

Algorithmus 3 Ermittlung von α_{opt} und $\lambda_{optFinal}$

Require: Array bestehend aus Datenmatrix aller λ_{optCV} und Datenmatrix mit korrespondierenden MSEPs für alle α aus Kreuzvaldierung

INPUT: r_{max} x $test_k$ Matrix für jedes $\alpha_1, \alpha_i, \cdots, \alpha_m$ bestehend aus λ_{optCV}
INPUT: r_{max} x $test_k$ Matrix für jedes $\alpha_1, \alpha_i, \cdots \alpha_m$ bestehend aus MSEP

(1.) Ermittle λ_{opt} für jedes α
for $i = \alpha_1$ to α_m **do**
    λ_{opt_i} = max(Häufigkeitsverteilung von λ_{optCV_i})
end for

(2.) Berechne für jedes λ_{opt_i} den Mittelwert des MSEP
for i = 1 to α_m **do**
    $MSEP_{\lambda_{opt_i}} = \frac{1}{Anzahl\lambda_{opt_i}} * \sum MSEP[\lambda_{opt_i}]$
end for

(3.) Finde α_{opt} für Elastic Net
α_{opt} (Elastic Net) = $\alpha[min(MSEP_{\lambda_{opt_i}})]$

(4.) Ermittle $\lambda_{optFinal}$ für Ridge Regression, Lasso und Elastic Net
$\lambda_{optFinal}$ (Ridge Regression) = $\lambda_{opt}[\alpha = 0]$
$\lambda_{optFinal}$ (Lasso) = $\lambda_{opt}[\alpha = 1]$
$\lambda_{optFinal}$ (Elastic Net) = $\lambda_{opt}[\alpha = \alpha_{opt}$ (Elastic Net)]

```
### R-Skript zur Ermittlung von $\lambda_{optCV}$ und Berechnung des zugehören MSEP
# training := Trainingsdatensatz
# test := Testdatensatz
# Kreuzvalidierung zur Bestimmung von $\lambda_{optCV}$ mit 100 Werten
  cv.out<-cv.glmnet(X[training,],y[training], alpha=alpha.test,
      standardize = T, nfolds = 6, intercept = F, lower.limits=0)
#  Ermittle das größte λ innerhalb minMSE+ 1 SE
  best.lambda<-cv.out$lambda.1se
#  Vorhersage der y-Werte im Testdatensatz mit optimiertem $\lambda_{optCV}$
  shrink.pred<-predict(shrink.mod,s=best.lambda, newx = X[test,],lower.
      limits=0) # lower.limits
# Berechnung des Vorhersagefehlers MSEP
  msep<-mean((y[test.rows]-shrink.pred)^2)
### Ende R-Skript
```

2.11.7. Random Forest Analyse

Theoretischer Hintergrund Die Random Forest Analyse ist ein Verfahren auf der Basis von Entscheidungsbäumen [107]. Die Methode kann sowohl zur Klassifizierung (kategoriale Variablen) als auch zur Regression (metrische Variablen) verwendet werden. In letzterem Fall spricht man auch von Regressionsbäumen. Basis für die Random Forest Analyse ist der CART-Algorithmus (Classification and Regression Trees (CART)), der 1984 erstmals von Breiman publiziert wurde [318]. Anhand des CART Algorithmus wird der X-Variablenraum X_1, X_2, ...X_p durch binäre, rekursive Regression in Regionen R_1, R_2, ...R_J aufgetrennt, sodass bei jedem Verzweigungsschritt (Split) s die Restvarianz RSS der beiden Hälften R_1 und R_2 minimiert wird:

$$RSS = \sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{2.62}$$

wobei $\hat{y}_{R1}$ der Mittelwert aller Beobachtungen in Raum R_1 und $\hat{y}_{R2}$ der Mittelwert aller Beobachtungen in Raum R_2 ist. An jedem Knoten s wird die RSS aller Variablen berechnet und der Split erfolgt mit der Variablen mit der geringsten RSS. Der so partitionierte Datenraum wird anschließend mit dem obigen Verfahren immer weiter aufgeteilt, bis ein bestimmtes Abbruchkriterium erfüllt ist. Das Verfahren nennt sich daher auch rekursives, binäres Splitting. Auf diese Weise wird der Datenraum im Hinblick auf die Regressanden homogen partitioniert.
Ein Nachteil der rekursiven Partitionierung ist die hohe Instabilität und somit hohe Modellvarianz. Wird der Datensatz z.B. durch Aufteilung in Trainings- und Testdatensatz nur geringfügig geändert, ändert sich u. U. auch der gesamte Verlauf des Baumes. Um dieses Problem zu umgehen, werden durch Bootstrapping (n mal ziehen ohne Zurücklegen) viele verschiedene Datensätze erzeugt, Bäume erstellt und anschließend gemittelt. Dieses Verfahren wird als Bagging bezeichnet. Da jedoch bei jeder Bootstrap Probe alle Variablen verwendet werden, besteht hier die Gefahr korrelierter Bäume, d. h. die Struktur der einzelnen Bäume ist untereinander ähnlich, und somit ist die Modellvarianz immer noch vergleichsweise hoch.
Random Forest geht daher noch einen Schritt weiter. Während beim Bagging bei jedem Bootstrap Baum alle zur Verfügung stehenden Variablen verwendet werden, wird bei einem Random Forest Modell nur eine randomisierte Auswahl mit $k < p$ Variablen verwendet. Auf diese Weise werden die einzelnen Bootstrap Bäume dekorreliert und die Vorhersage damit verlässlicher. Um die Bedeutung eines Prädiktors für die Regressanden abschätzen

Algorithmus 4 Random Forest [319]

Lege die Anzahl der Modelle m fest ▷ ntree
for $i = 1$ to m **do**
    Generiere eine Bootstrap Probe aus dem Originaldatensatz
    Trainiere ein Baummodell für dieser Probe
    for each split **do**
        Ziehe randomisiert k ($< p$) X-Variablen ▷ mtry
        Selektiere die X-Variable aus den k Prädiktoren, die den Datenraum
        mit der geringsten RSS partitioniert und teile den Datenraum
    end for
    Lasse den Baum wachsen bis ein spezifisches Stoppkriterium erreicht ist
end for

zu können, verwendet die Random Forest Analyse das Konzept der Variable Importance (VI). Basis für die VI sind die out-of-bag (OOB) Proben, d. h. die Proben die beim Ziehen der Bootstrap Proben ausgelassen wurden (ca. 1/3 aller Proben). Um die Bedeutung einer Variable zu erfassen, werden die Werte jeder x-Variable innerhalb der OOB Proben randomisiert permutiert und der MSE berechnet. Durch zufällige Zerstörung der Korrelation zwischen Regressor und Regressand nimmt der MSE bei Variablen mit hoher Bedeutung stärker zu, als bei Variablen, die mit dem Regressanden nicht oder nur sehr gering assoziert sind.

Durchführung Für die Random Forest Analyse wurde das R-Paket *VSURF* verwendet [320]. Das R-Paket *VSURF* erlaubt eine mehrstufige Variablenselektion.
Zunächst müssen jedoch die Parameter *ntree* (Anzahl der Bootstrap Bäume) und *mtry* (Anzahl der Variablen pro Split) optimiert werden. Zur Anpassung von *ntree* wurden Random Forest Modelle mit 10 Werten zwischen 10 und 10000 berechnet und anschließend der OOB-Fehler extrahiert. Jedes Modell wurde 50 mal berechnet (*nforthres* = 50):

```
Grid.ntree<-c(10, 50,100, 200, 500, 800,1000, 2000, 5000,10000)
rF.ntree<-lapply(Grid.ntree,function(x)VSURF(X,y,ntree=x,nfor.thres =
    50,nfor.interp = 3 ,nfor.pred = 0,nmin=1,nsd=1))
OOB-Fehler<-sapply(1:length(rF.ntree),function(x)rF.ntree[[x]]$mean.
    perf)
```

Nach der Optimierung von *ntree* wurde die Anzahl der Variablen pro Split (*mtry*) mit den im folgenden Listing angegebenen Werten optimiert:

```
Grid.mtry<-c(0.01, 0.04, 0.07, 0.08 ,0.1, 0.12, 0.13, 0.15, 0.18, 0.21,
     0.25, 0.33, 0.5, 0.8, 1)
```

```
rF.mtry<-lapply(Grid.mtry, function(x)VSURF(X, y, mtry = x*ncol(X),
    ntree = 2000, nfor.thres = 50 ,nfor.interp =3 , nfor.pred = 0, nmin
    = 1, nsd = 1))
OOB-Fehler<-sapply(1:length(rF.mtry),function(x)rF.mtry[[x]]$mean.perf)
```

Die Variablenselektion mit *VSURF* verläuft in drei Schritten:

1. **Variablenranking**
   Zunächst werden *mforthresh* Random Forest Modelle aufgestellt und für alle x-Variablen die Variable Importance (VI) und die korrespondierenden Standardabweichungen berechnet. Die Daten werden anschließend in absteigender Reihenfolge nach Größe der VI sortiert.

2. **Variablenselektion**
   Anschließend wird ein CART Modell für die Standardabweichungen berechnet und dessen Minimalwert *minthresh* als Schwelle für Schritt 3 verwendet. Der Schwellwert kann mit einem Faktor *nmin* multipliziert werden. Auf diese Weise kann die Variableneliminierung relaxierter ($nmin \leq 1$) oder restriktiver ($nmin > 1$) verfolgt werden. Hier wurde $nmin = 20$ verwendet.

3. **Eliminierung redundanter Variablen**
   Die Variablen aus Schritt 2 werden, angefangen bei der Variable mit dem höchsten VI, sukzessiv in Random Forest Modelle eingefügt und die minimalen OOB-Fehler *err.min* über je *nforinterp* Modelle gemittelt. Nach Berechnung der korrespondierenden OOB-Standardabweichungen *sd.min* wird das kleinste Modell mit einem mittleren $OOB < err.min + nsd \cdot sd.min$ selektiert. Der Faktor *nsd* erlaubt die Feinabstimmung des Modells. Werte >=1 führen zu sparsameren Modellen. Bei $nsd < 1$ werden größere OOB-Fehler zugelassen und entsprechend mehr Variablen selektiert.

Der oben beschriebene Variablenselektionsprozess erfolgte mit dem folgenden R-Code:

```
RF.select<-VSURF(X, y, mtry = floor(0.08*ncol(X)), ntree = 2000, nfor.
    thres = 50, nfor.interp = 25 , nfor.pred = 10, nmin=20, nsd=0.01)
```

Teil III.

Ergebnisse und Diskussion

3. Ergebnisse und Diskussion

3.1. AcorA Proof of Concept

In einer Proof of Concept Studie wurde zunächst untersucht, ob eine Identifizierung von biologisch aktiven Substanzen in komplexen Naturstoffextrakten mithilfe von AcorA möglich ist. Als Grundvoraussetzung für die Aktivitäts-Korrelations-Analyse gilt, dass die Bioaktivitäten in einem weiten Aktivitätsbereich über verschiedene Extrakte verteilt sein müssen. Um diese unterschiedliche Verteilung von bioaktiven Substanzen in den Extrakten zu simulieren, wurden fünf schwach aktive Methanolextrakte (Wachstumsinhibition < 20 %) aus Basidiomyceten der Gattung *Hygrophorus* randomisiert wie unter 2.9.4 auf Seite 68 beschrieben, mit verschiedenen Konzentrationen an Antibiotika (Amoxicillin, Erythromycin, Rifampicin) versetzt. Auf diese Weise wurden insgesamt 16 gespickte *Hygrophorus* Extrakte erhalten. Zusätzlich wurden die 5 ungespickten Grundextrakte als Negativkontrolle verwendet. Insgesamt wurden 21 Extrakte in einem *Bacillus subtilis* Assay auf ihre inhibitorischen Eigenschaften untersucht und mit dem FT-ICR-Massenspektrometer als Triplikate gemessen.
Durch Messung der Antibiotikastammlösungen war bekannt, welche Massensignale aus den einzelnen Antibiotika stammen. Diese Informationen wurden dazu genutzt, die Qualität der AcorA-Methode - und im weiteren Verlauf dieser Arbeit auch weiterer Datenanalysemethoden - abzuschätzen.

3.1.1. Wachstumsinhibition von Bacillus subtilis

Abbildung 3.1 zeigt die Verteilung der Bioaktivität über die verschiedenen Extrakte. Eine vergleichsweise hohe Inhibition wurde mit den Extrakten erzielt, in denen Rifampicin und/oder Erythromycin in der höchsten eingesetzten Konzentration (10 µM) enthalten waren (HEA01, 02, 04, 09, 15,16). Auffällig ist, dass Amoxicillin selbst in der höchsten eingesetzten Konzentration (10 µM) keine signifikante Wachstumsinhibition hervorgerufen hat. So

zeigen die Extrakte, in denen Amoxicillin in hoher Konzentration, die beiden anderen Antibiotika jedoch nicht oder nur in antibiotisch unwirksamen Konzentrationen vorlagen (HEA06, 10), lediglich eine schwach ausgeprägte Aktivität gegenüber *B. subtilis*. Der Grund hierfür ist in dem vergleichsweise hohen IC_{50}-Wert für Amoxicillin zu sehen. So wurde von Michels ein IC_{50}-Wert von $7{,}1 \times 10^{-1}$ mM für Amoxicillin bestimmt [277]. Die IC_{50}-Werte für Erythromycin ($4{,}7 \times 10^{-4}$ mM) und Rifampicin ($6{,}8 \times 10^{-4}$ mM) liegen etwa 3 Zehnerpotenzen darunter. Die eingesetzten Amoxicillinmengen waren offenbar zu gering, um eine signifikante Wachstumsinhibition zu bewirken.
Wachstumshemmungen in moderater Intensität wurden mit den Extrakten HEA07 und HEA13 erreicht. Die inhibierende Wirkung geht in diesen beiden Extrakten vermutlich von Erythromycin aus, da nur dieses in einer ausreichend hohen Konzentration vorlag. Die 5 *Hygrophorus* Grundextrakte HE01-05 sowie die mit Antibiotika versetzten Extrakte HEA03, 05, 06, 08,10, 11 und 12 verursachten eine geringe Wachstumsinhibition bis zu etwa 30 %.

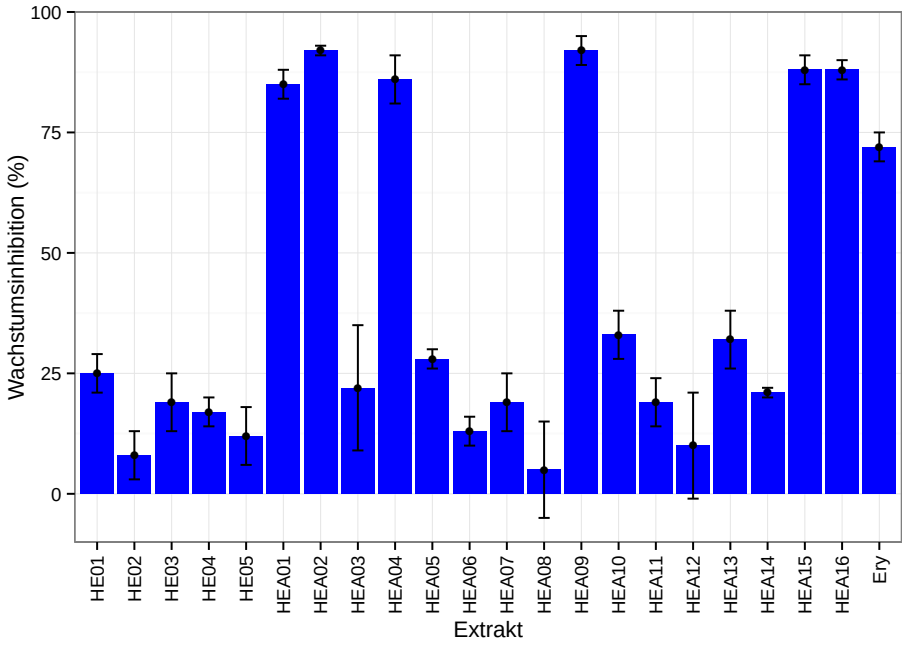


Abbildung 3.1.: Proof of Concept: Ergebnisse des *Bacillus subtilis* Assays für die Konzentration β = 10 mg/mL. HE: *Hygrophorus* Extrakt ohne Antibiotika, HEA: *Hygrophorus* Extrakt mit Antibiotika. Als Positiv-Kontrolle wurde Erythromycin (Ery, 1 µM) verwendet.

Parallel zur Erfassung der Bioaktivitätsdaten wurden Aliquots der 21 Extrakte mit dem FT-ICR-Massenspektrometer im Positiv- und Negativ-Ionen-Modus gemessen. Nach dem Peak-Picking wurden 1393 Peaks im Positiv-Ionen-Modus und 1291 Peaks im Negativ-Ionen-Modus in getrennten Datenmatrizes aligniert. Beispielhaft sind einige Positiv-Ionen-Spektren in Abbildung 3.2 dargestellt.
Insbesondere für Erythromycin (Rot gekennzeichnet) ist die für AcorA benötigte Varianz sehr gut zu beobachten. Die Peakintensität schwankt von 0 im ungespickten Extrakt HE05 über $0,6 \cdot 10^7$ in Extrakt HEA13 zu $1,9 \cdot 10^8$ in den Extrakten HEA09 und HEA04.

Für die Korrelation zwischen Metaboliten- und Aktivitätsprofil wurde das in unserer Arbeitsgruppe entwickelte R-Paket *Acora* verwendet. Das R-Paket kalkuliert für jeden Massenpeak die Spearman-Rangkorrelation zwischen den Signalintensitäten aus MS- und Bioaktivitätsmessung der Extrakte. Die Ergebnisse werden in einer Liste zusammengefasst, in der die *m/z* -Signale absteigend nach Größe ihres Korrelationskoeffizienten sortiert werden.

Die Güte der AcorA-Methode wird in Abbildung 3.3a deutlich. Die Abbildung zeigt die nach Berechnung von Sensitivität (Richtig-Positiv-Rate) und der Falsch-Positiv-Rate (Komplement der Spezifität) erhaltene ROC-Kurve für den Positiv-Ionen-Modus. Sowohl der AUC-(0,99) als auch der partielle AUC-Wert (0,91), der für den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate berechnet ist, zeigen, dass eine Akkumulation der Antibiotikapeaks im oberen Bereich der Ergebnisliste stattgefunden hat. Innerhalb der ersten 50 Peaks der Ergebnisliste befinden sich 38 Peaks, die den Antibiotika zugeordnet werden können. Erweitert man diesen Bereich auf die ersten 100 Einträge in der Ergebnisliste, findet AcorA bereits 44 der insgesamt 47 gesuchten Antibiotikapeaks (3.3b).

Um die Ergebnisliste filtern zu können, ist in dem AcorA R-Paket eine Funktion integriert, die auf Basis des vorliegenden Datensatzes einen Permutationstest durchführt. Dabei wird eine Nullverteilung generiert, mit deren Hilfe eine Signifikanzschwelle für die Korrelationskoeffizienten festgelegt wird. Alle Massensignale deren Korrelationskoeffizient größer ist als die Signifikanzschwelle, werden in einer Hitliste annotiert.

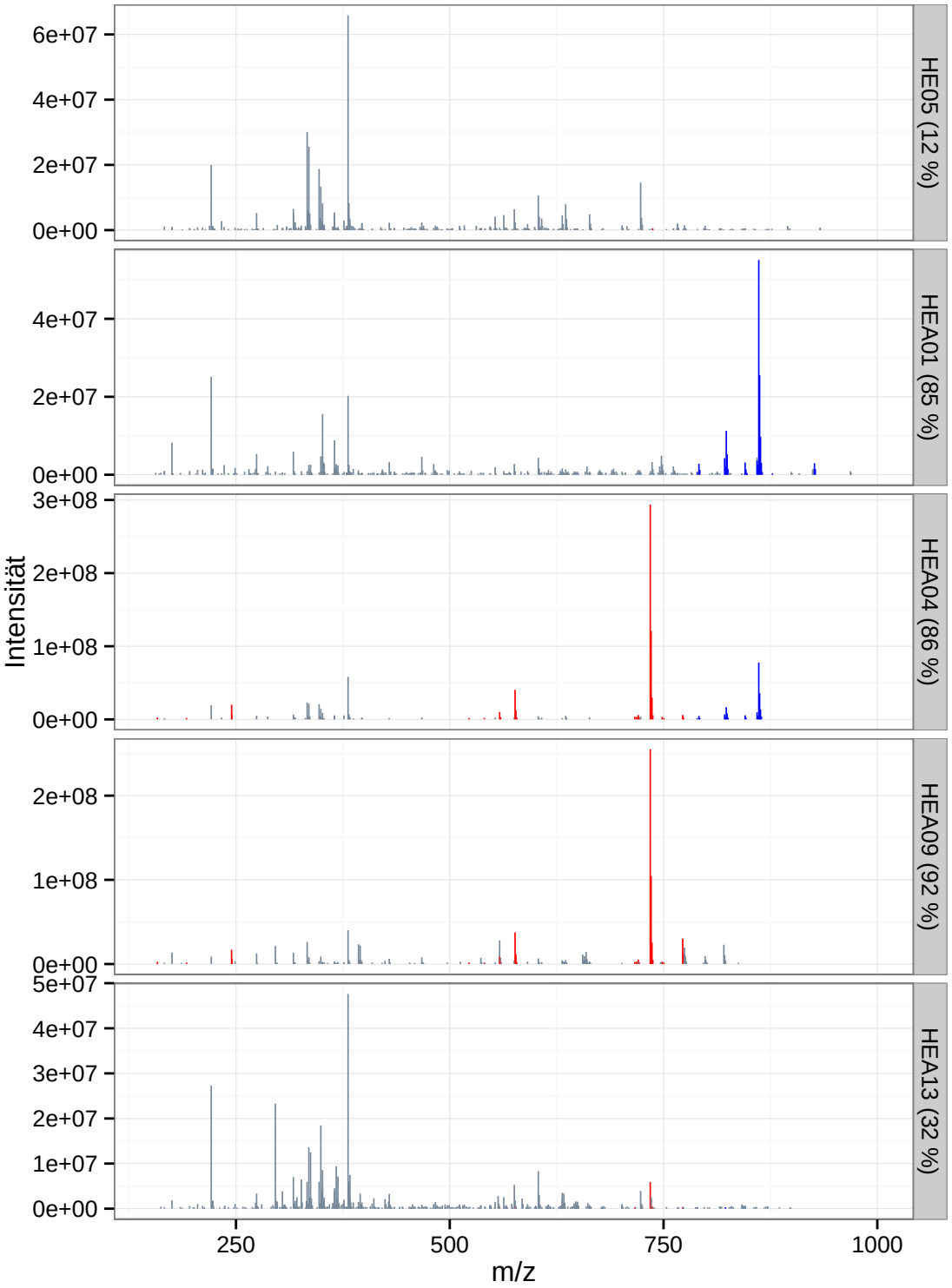


Abbildung 3.2.: FT-ICR-MS Spektren einiger ausgewählter Proben. Peaks aus Erythromycin (Rot), Rifampicin (Blau) sowie der nicht-aktiven Verbindungen (Grau) sind farblich gekennzeichnet. Die Wachsumsinhibition von *B. subtilis* ist in Klammern angegeben. Relevante Antibiotikakonzentrationen sind wie folgt: *HE05:* ungespickt; *HEA01:* 10 µM Rifampicin; *HEA04:* 10 µM Erythromycin, 10 µM Rifampicin; *HEA09:* 10 µM Erythromycin; *HEA13:* 0,1 µM Erythromycin, 0,1 µM Rifampicin.

Für dieses Experiment wurde ein einseitiger, oberer Test mit einer Irrtumswahrscheinlichkeit von $\alpha = 5\ \%$ durchgeführt. Insgesamt wurden 95 signifikant korrelierende Peaks, 63 im Positiv-Modus, 32 im Negativ-Modus ermittelt. Beispielhaft ist die Hitliste für den Positiv-Ionen-Modus in Tabelle 3.1.1 dargestellt. Von den 63 signifikant korrelierenden Peaks konnten 23 dem Rifampicin und 19 dem Erythromycin zugeordnet werden. Interessant ist, dass nicht nur der monoisotopische $[M+H]^+$-Peak von Rifampicin (*m/z* 823,41151) korreliert, sondern in ähnlichem Maße auch dessen Isotopenpeaks und Na^+- und K^+-Addukte. Neben dem Rifampicin findet man in der Hitliste weitere signifikant korrelierende Massensignale, die dem Rifampicin Chinon (Dehydrorifampicin) sowie dem 27-demethoxy-Rifampicin zugeordnet werden können. Beide Derivate sind in der Literatur als Autooxidationsprodukte und Verunreinigungen in Rifampicin Lösungen bekannt [321, 322, 323].

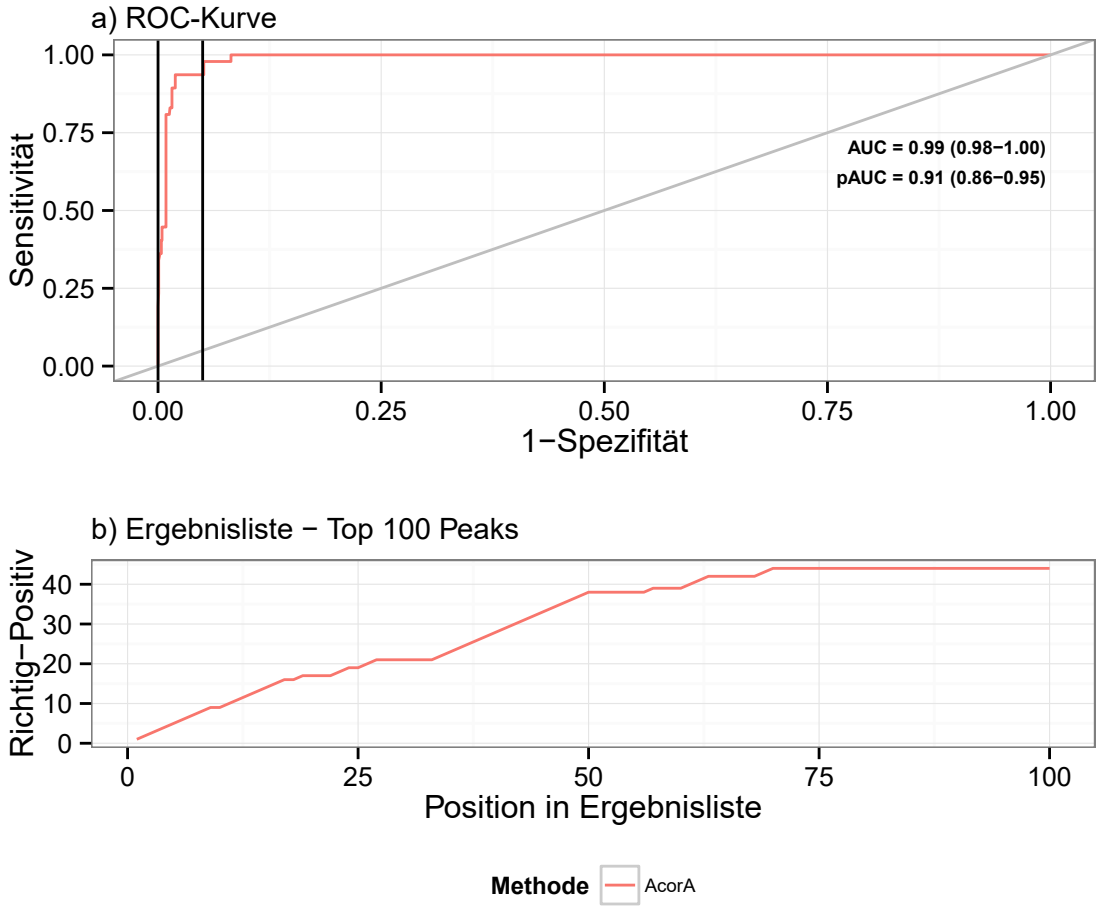


Abbildung 3.3.: a) ROC-Kurve der Ergebnisliste und b) Top 100 Peaks des Positiv-Ionen-Modus.

Tabelle 3.1.: Hitliste Positiv-Modus

Rang	Korrelations-koeffizient	m/z	Probennummer	Annotation	Δ ppm
1	0.687	822.4007	1, 2, 4, 13, 15, 16	[Rif - H_2 + H]$^+$, 1. Isotop	0,7
2	0.674	821.39612	1, 2, 4, 5, 15, 16	[Rif - H_2 + H]$^+$	0,8
3	0.670	862.37169	1, 2, 4, 5, 15, 16	[Rif + K]$^+$, 1. Isotop	0,07
4	0.660	845.39464	1, 2, 4, 15, 16	[Rif + Na]$^+$	0,4
5	0.660	846.39695	1, 2, 4, 15, 16	[Rif + Na]$^+$, 1. Isotop	0,9
6	0.658	791.38599	1, 2, 4, 15, 16	[Rif - MeOH]$^+$	0,3
7	0.658	792.38699	1, 2, 4, 15, 16	[Rif - MeOH]$^+$, 1. Isotop	3,2
8	0.658	859.35289	1, 2, 4, 5, 14, 15, 16	[Rif - H_2 + K]$^+$	0,3
9	0.658	860.35577	1, 2, 4, 5, 14, 15, 16	[Rif - H_2 + K]$^+$, 1. Isotop	0,2
10	0.657	287.15266	1, 2, 4, 15, 16	?	-
11	0.657	863.37055	1, 2, 4, 15, 16	[Rif + K]$^+$, 2. Isotop	5,1
12	0.657	864.37163	1, 2, 4, 15, 16	[Rif + K]$^+$, 3. Isotop	7,8
13	0.655	861.3678	1, 2, 4, 5, 14, 15, 16	[Rif + K]$^+$	0,6
14	0.653	823.41151	1, 2, 4, 15, 16	[Rif + H]$^+$	1,1
15	0.653	824.41474	1, 2, 4, 15, 16	[Rif + H]$^+$, 1. Isotop	1,2
16	0.653	825.41758	1, 2, 4, 15, 16	[Rif + H]$^+$, 2. Isotop	1,8
17	0.603	789.37041	1, 2, 15, 16	[Rif - H_2 - MeOH]$^+$	0,2
18	0.592	287.48765	1, 2, 15, 16	[$C_{12}H_{24}O_6$ + Na]$^+$	-
19	0.592	877.36099	1, 2, 15, 16	NA aus Rif	-
20	0.592	899.32288	1, 2, 15, 16	NA	-
21	0.591	1043.4458	1, 2, 15, 16	NA	-
22	0.581	924.495	1, 2, 5, 15, 16	?	-
23	0.580	926.50933	1, 2, 15, 16	NA aus Rif	-
24	0.580	927.51288	1, 2, 15, 16	NA aus Rif *m/z* 926,50933 , 1. Isotop	-
25	0.554	1041.43219	2, 15, 16	NA	-
26	0.545	790.37557	2, 15 ,15	[Rif - H_2 - MeOH]$^+$, 1. Isotop	2,0
27	0.545	843.37786	2, 15 ,15	[Rif - H_2 + Na]$^+$	1,0
28	0.499	900.32745	1, 2, 16	NA	-
29	0.498	1044.44626	1, 2, 16	NA	-
30	0.463	859.34924	1, 2, 5, 14, 16	NA	-
31	0.458	865.37021	2, 15	NA	-

Tabelle 3.1.: Hitliste Positiv-Modus (Fortsetzung)

Peak	Korrelations-koeffizient	m/z	Probennummer	Annotation	Δ ppm
32	0.458	1041.43823	2, 15	NA	-
33	0.454	925.49872	1, 15, 16	NA	-
34	0.449	865.37584	1, 15, 16	$[Rif + K]^+$, 4. Isotop	6,7
35	0.429	735.46954	2, 4, 7, 9, 13, 14	$[EryA + H]^+$, 1. Isotop	3,3
36	0.420	773.42798	4, 9	$[EryA + K]^+$, 1. Isotop	0,3
37	0.411	158.11733	4, 9	$[Desosamin + H]^+$ aus Ery	1,5
38	0.411	192.13834	4, 9	NA aus Ery	-
39	0.411	244.84432	4, 9	NA aus Ery	-
40	0.411	245.17872	4, 9	NA Ery	-
41	0.411	522,34299[z]	4, 9	$[Fragment\ Ery + H]^+$	-
42	0.411	540,35349[z]	4, 9	$[Fragment\ Ery + H]^+$	-
43	0.411	558.36376	4, 9	$[AEryA - Cladinose + H]^+$	0,2
44	0.411	559.36786	4, 9	$[AEryA - Cladinose + H]^+$, 1. Isotop	1,5
45	0.411	578.38092	4, 9	$[EryA - Cladinose + H]^+$, 2. Isotop	0
46	0.411	718.47376	4, 9	$[EryB + H]^+$	1,0
47	0.411	720.45212	4, 9	$[EryC / NdeMeEryA + H]^+$	0,3
48	0.411	721.45602	4, 9	$[EryC / NdeMeEryA + H]^+$, 1. Isotop	0,5
49	0.411	748.44816	4, 9	$[EryE + H]^+$	0,5
50	0.411	750.46292	4, 9	$[EryF / EryAEO + H]^+$	0,7
51	0.402	195.79219	15, 16	NA	-
52	0.402	286.48032	15, 16	NA	-
53	0.402	891.37364	15, 16	NA	-
54	0.402	1171.62964	15, 16	NA	-
55	0.398	175.1187	HE1, HE2, HE3, HE4, HE5, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16		-
56	0.394	274.49867	1, 2	NA	-
57	0.394	847.40308	1, 2	$[Rif + K]^+$, 2. Isotop	2,3
58	0.389	396.25982	HE1, HE3, 5, 6, 7, 9, 1, 12, 13, 15, 16	NA	-
59	0.387	365.15754	HE1, 5, 9, 14, 15	NA	-

Tabelle 3.1.: Hitliste Positiv-Modus (Fortsetzung)

Peak	Korrelations-koeffizient	m/z	Probennummer	Annotation	Δ ppm
60	0.383	1045.45418	1, 2	NA	-
61	0.381	772.42468	4, 9, 13, 14	$[EryA + K]^+$	0,3
62	0.375	576.37454	4, 9, 13, 14	$[EryA - Cladinose + H]^+$	0,6
63	0.371	734.46701	2, 3, 4, 5, 6, 7, 9, 12, 13, 14	$[EryA + H]^+$	2,1

[a] Diese Fragmente können bei der Fragmentierung verschiedener Erythromycin-Spezies entstehen ([324])

Auch von diesen Verbindungen treten Isotopen- und Adduktpeaks in der Hitliste auf. Es ist nicht bekannt, ob die Verunreinigungen eine identische antibiotische Aktivität besitzen wie die API Hauptsubstanz. Wie Abbildung 3.6 zeigt, sind viele der als NA bezeichneten Signale mit denen des Rifampicins hochkorreliert und weisen daher auch eine Assoziation zur Bioaktivität auf. Möglicherweise stammen sie ebenfalls aus der Rifampicin Stammlösung und wurden bei dessen Messung durch Matrixeffekte supprimiert. Da die *Hygrophorus* Grundextrakte eine Basisaktivität von bis zu 25 % aufweisen, könnten einige der NA Signale auch von (schwach) aktiven Substanzen aus den Grundextrakten verursacht worden sein.

Die Korrelation von Peakclustern, d. h. von verschiedenen Isotopen- und Adduktpeaks, die nach Datendekonvolution einer Substanz zugeordnet werden können, wird auch für Erythromycin beobachtet. Beispielsweise findet man zusätzlich zum $[M+H]^+$-Peak von Erythromycin A (Zeile 63) auch dessen ersten Isotopenpeak (Zeile 35), sowie das entsprechende K^+-Addukt mit dessen ersten Isotopenpeak in den Zeilen 61 und 36 der Hitliste.
Neben dem Hauptantibiotikum Erythromycin A weisen einige weitere Peaks eine signifikante Korrelation auf. Bei diesen Verbindungen handelt es sich mit hoher Wahrscheinlichkeit um Erythromycin B, Erythromycin C, Anhydroerythromycin A und entsprechende Abbauprodukte, bei denen der Cladinoserest fehlt. Diese fermentativen Neben- und Abbauprodukte entstehen bei der biotechnologischen Herstellung von Erythromycin A und sind in geringen Mengen in pharmazeutischen Erythromycinprodukten enthalten [325, 324, 326]. Messungen der Antibiotika-Stammlösungen zeigen, dass diese Verbindungen bereits in der Erythromycinstammlösung vorhanden waren. Ähnlich dem Rifampicin sind auch die Peaks der verschiedenen Erythromycinderivate untereinander hoch korreliert und somit genauso mit der Aktivität assoziert, wie das bioaktive Hauptprodukt (Abbildung 3.6). Insgesamt konnten 67 % der signifikant korrelierenden Peaks im Positiv-Ionen-Modus den Antibiotika

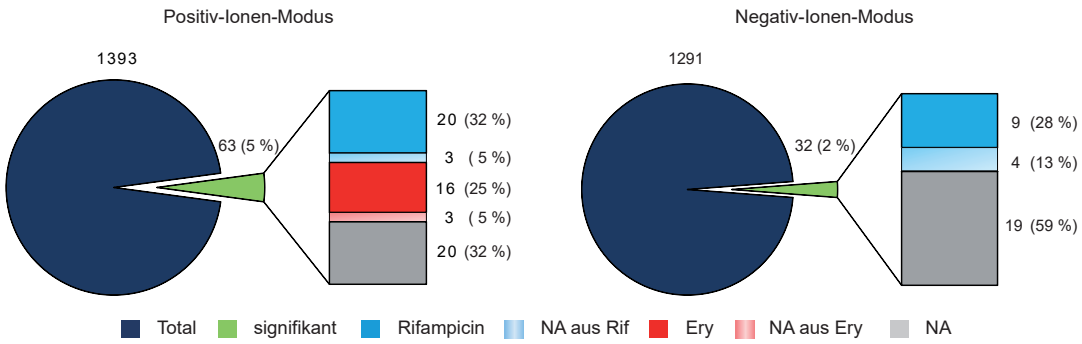


Abbildung 3.4.: Zusammenfassung der Hitlisten.

zugeordnet werden. Im Negativ-Ionen-Modus liegt dieser Anteil mit 41 % deutlich niedriger. Ähnlich dem Positiv-Modus, wird die Hitliste des Negativ-Modus (siehe Tabelle B.2 auf Seite 259 im Anhang) von signifikanten Korrelationen zu Isotopen- und Adduktpeaks des Rifampicin dominiert. Insgesamt wurden 13 der 14 gesuchten Rifampicin Peaks in der Hitliste annotiert. Es wurden jedoch keine signifikanten Korrelationen zu Peaks des Erythromycin gefunden. Im weiteren Verlauf der Ergebnisliste werden bis zu Peak 46 drei der sechs gesuchten Erythromycin Peaks gefunden. Daher weist der AUC-Wert von 0,97 trotzdem auf eine hohe Wiederfindungsrate hin. Eine Übersicht über die Anzahl der signifikant korrelierenden Peaks ist in Abbildung 3.4 gegeben. Da Amoxicillin in den eingesetzten Konzentrationen keine messbare antibiotische Aktivität hervorrufen kann, sind somit auch keine dem Amoxicillin zuzuordnenden Peaks in den Hitlisten vertreten.

Die Ursachen für die gute Performance von AcorA ist in Abbildung 3.5 erkennbar. Die Peaks der bioaktiven Substanzen weisen größtenteils einen deutlich höheren positiven Korrelationskoeffizienten auf als die meisten der nicht-aktiven Verbindungen. Weiterhin ist gut zu erkennen, dass der nicht-parametrische Spearman-Rangkorrelationstest skaleninvariant ist. Kleine Peaks werden somit nicht gegenüber großen Peaks diskriminiert und erhalten ähnlich hohe Korrelationskoeffizienten. Dies ist ein entscheidender Vorteil gegenüber varianzbasierten Methoden (z. B. PCA, PCR) und kann insbesondere bei niedrigkonzentrierten Substanzen relevant sein. Die Tatsache, dass die Peaks aus der Rifampicin Lösung höhere Korrelationskoeffizienten aufweisen als Peaks aus dem Erythromycin, ist vermutlich darauf zurückzuführen, dass Rifampicin in deutlich mehr Proben in hoher Konzentration (und somit gut detektierbar) vorlag als Erythromycin. Auf diese Weise werden die Peaks in mehr Proben detektiert, sodass eine höhere Korrelation kalkuliert wird. Im Negativ-Ionen-Modus wurden

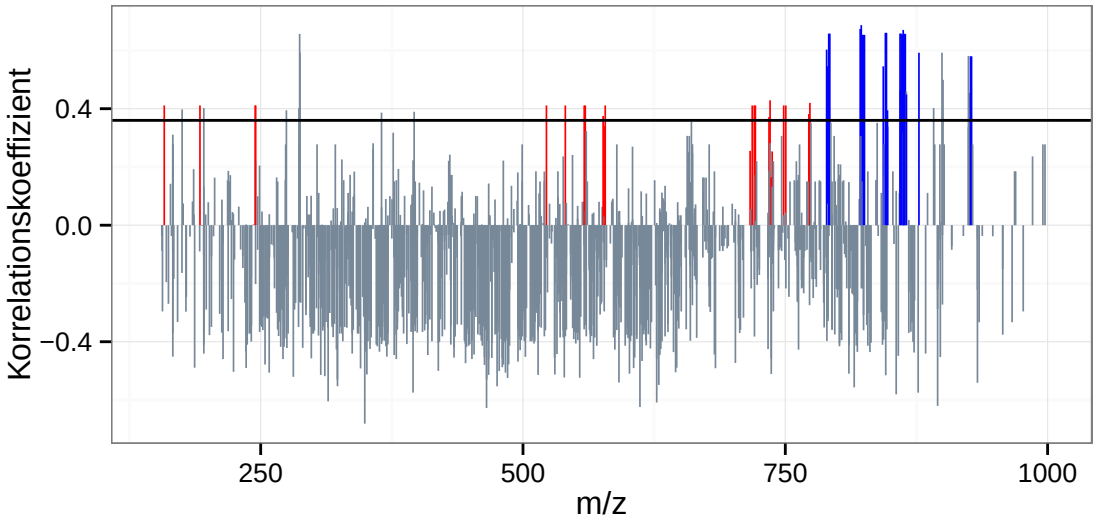


Abbildung 3.5.: Plot der Korrelationskoeffizienten aus der Aktivitäts-Korrelations-Analyse gegen *m/z*. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und den vermutlich nicht-aktiven Substanzen (Grau) sind farblich gekennzeichnet. Die waagerechte schwarze Linie gibt die Signifikanzgrenze an.

die gesuchten Erythromycin Peaks aufgrund schlechter Ionisationseigenschaften sogar nur in einer einzigen Probe detektiert (HEA09). Dies erklärt weshalb Erythomycin nicht in der Hitliste auftauchte und verdeutlicht nochmal, dass es für die Aktivitäts-Korrelations-Analyse wichtig ist, dass die aktive(n) Substanz(en) über möglichst viele Proben in unterschiedlicher Konzentration verteilt werden sollte(n).

3.1.2. Pearson-Korrelation und Korrelationsnetzwerke zur Analyse der Hitliste

Nach Abschluss der Aktivitätskorrelationsanalyse kann eine Hitliste insbesondere bei LC-MS Analysen aus mehreren hundert Signalen bestehen [277]. In diesen Fällen erschwert die hohe Komplexität der Hitliste die Identifizierung zusammenhängender Peakcluster, die, wie oben beschrieben, Hinweise auf einen kausalen Zusammenhang zwischen Metaboliten und Bioaktivität geben können. Gerade die hohe Korrelation innerhalb der Peakcluster kann jedoch dazu genutzt werden, die Beziehungen zwischen einzelnen Peaks aufzudecken und durch anschließende Datendekonvolution die Identifizierung der zugrunde liegenden bioaktiven Metaboliten zu erleichtern. Die Darstellung der Pearson-Korrelationen zwischen den Massensignalen der Hitliste in Abbildung 3.6 verdeutlicht, dass die Hitliste im Wesentlichen aus zwei großen Peakclustern besteht. Beide Cluster sind scharf voneinander abgetrennt und enthalten jeweils nur die Peaks des Rifampicins (oberes Cluster) oder des Erythromycins (unteres Cluster).
Eine andere Option zur Visualisierung der Korrelationen innerhalb der Hitliste ist die Darstellung in einem Korrelationsnetzwerk (Abbildung 3.7 a). Auch hier ist sehr gut erkennbar, dass die Hitliste im Wesentlichen aus zwei Großclustern besteht, deren Peaks untereinander hoch korreliert sind. Unter Verwendung der partiellen Korrelationen in einem Gaussian Graphical Model Netzwerk (Abbildung 3.7 b) werden einzelne Subcluster erkennbar, die eine weitere Feinauflösung des Netzwerks ermöglichen. So bestehen beispielsweise die Einträge in Subcluster Rif_1 aus den Peaks von Rifampicin und 27-demethoxy-Rifampcin. Subcluster Rif_2 besteht aus den Na^+- und K^+-Addukten des Rifampicins. Demhingegen besteht Subcluster Rif_3 überwiegend aus Peaks des Rifampicin Chinons und dessen 27-demethoxy Derivats. Die Peaks in Subcluster Rif_4 wurden in der Hitliste überwiegend als NA annotiert, da sie in den Spektren der „reinen“ Antibiotikastammlösungen nicht beobachtet wurden. Es ist jedoch auffällig, dass diese Peaks in den Extrakten detektiert wurden, die mit Rifampicin versetzt waren. Möglicherweise handelt es sich also um Signale von Rifampicin bzw. Rifampicin Derivaten, die aufgrund von Matrixeffekten in der Antibiotikastammlösung supprimiert und erst in Kombination mit den Pilzgrundextrakten in den Spektren sichtbar wurden.

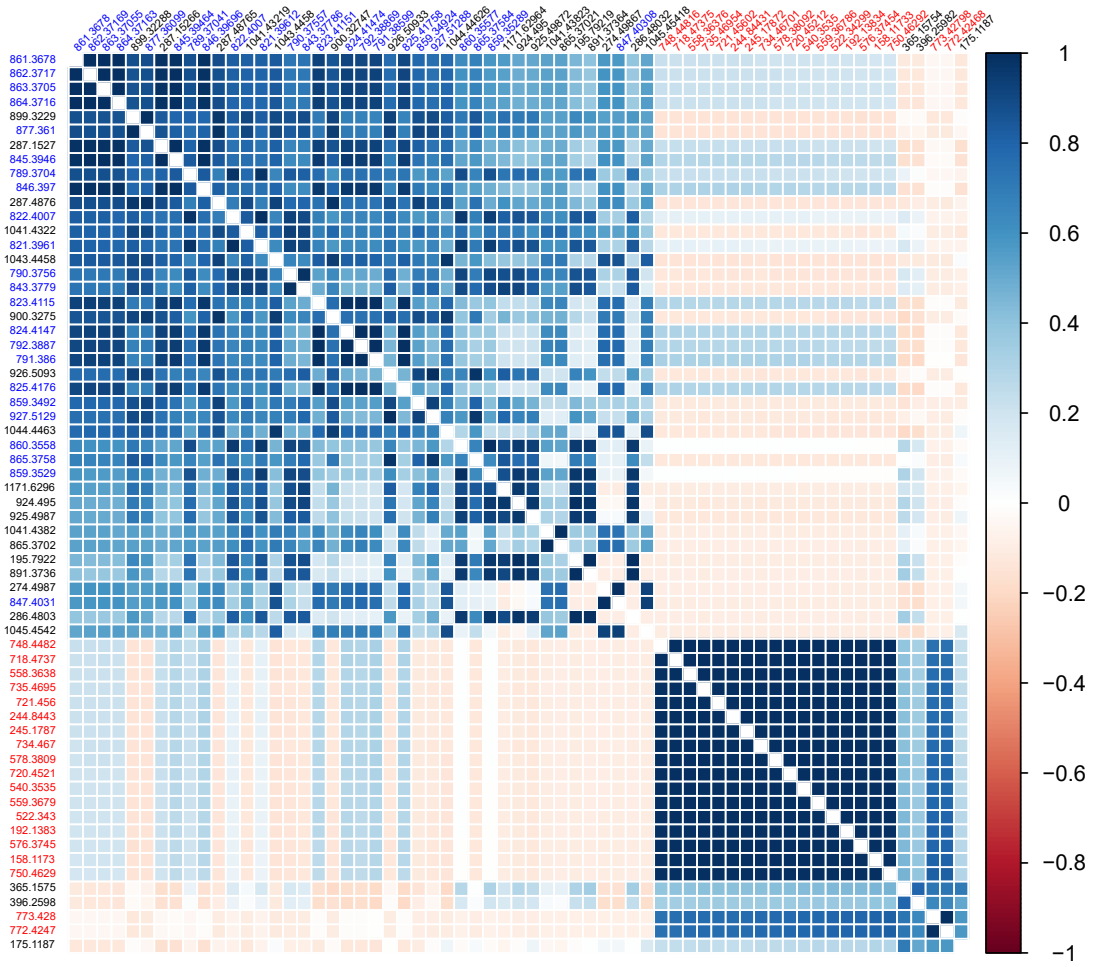


Abbildung 3.6.: Pearson-Korrelationen zwischen den Peaks der Hitliste. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und unbekannter Herkunft (NA, Schwarz) sind farblich gekennzeichnet.

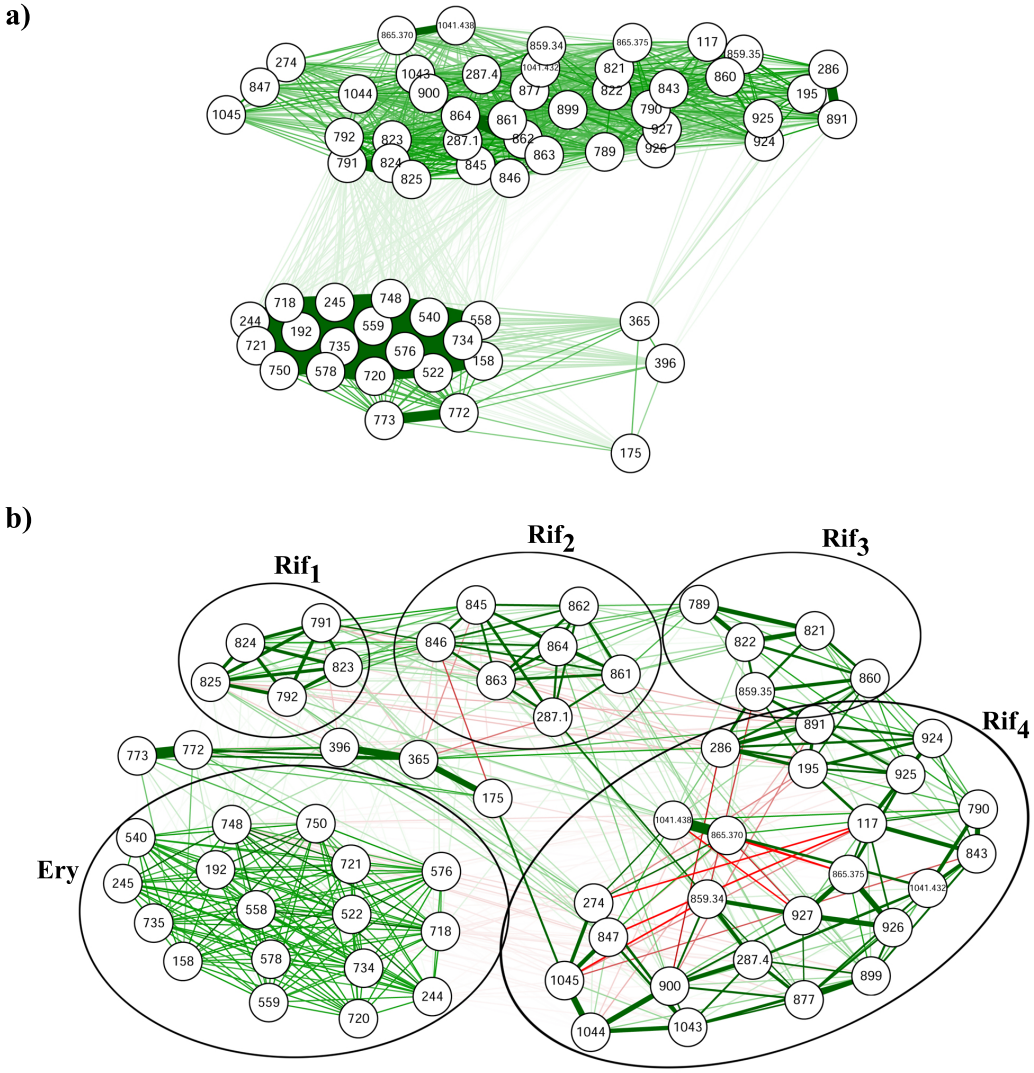


Abbildung 3.7.: a) Korrelationsnetzwerk basierend auf Pearson-Korrelation b) Gaussian Graphical Model Netzwerk basierend auf partiellen Korrelationen. Positive Korrelationen sind in Grün, negative Korrelationen in Rot dargestellt.

3.1.3. Diskussion

Im Rahmen des Proof of Concept Experiments hat die Aktivitäts-Korrelations-Analyse unter Verwendung der Spearman-Rangkorrelation sehr gute Ergebnisse in Hinblick auf die Identifizierung biologisch aktiver Substanzen in komplexen Naturstoffextrakten geliefert. Peaks beider aktiven Antibiotika wurden in hohem Maße in einer Hitliste angereichert. Da das dritte Antibiotikum (Amoxicillin) unterhalb der wirksamen Konzentration eingesetzt wurde, traten somit richtigerweise keine Massensignale des Amoxicillins in der Hitliste auf. Substanzen, die in Extrakten unterhalb der Schwellenwertkonzentration vorliegen, könnten daher mit AcorA nicht identifiziert werden. Das gleiche Problem tritt jedoch auch bei der Bioaktivitäts-geleiteten Fraktionierung auf.
Die Ergebnisse haben gezeigt, dass Peakcluster, bestehend aus Isotopen- und Adduktpeaks einer Substanz, in besonderem Maße auf einen kausalen Zusammenhang zwischen Metaboliten und Bioaktivität hindeuten. Die hohen Korrelationen zwischen Isotopen- und Adduktpeaks können in Form von Korrelationsnetzwerken dazu genutzt werden, die Identifizierung von Peakclustern innerhalb einer Hitliste zu erleichtern. Weiterhin können komplementäre Informationen aus Positiv- und Negativ-Ionen-Modus dazu verwendet werden, den Verdacht einer kausalen Korrelation zu erhärten.
Eine weitere Möglichkeit zur Identifizierung zusammenhängender Peakcluster mit hoher Korrelation zur Bioaktivität bietet ein Kovarianz-Korrelations-Diagramm. Dazu wird die Kovarianz zwischen Massensignal und Bioaktivität gegen m/z aufgetragen. Da die Kovarianz von der Größe der Signale abhängig ist, werden die resultierenden „Peaks“ zusätzlich farblich mit der Korrelation zwischen Massensignal und Bioaktivität gekennzeichnet. Als Ergebnis erhält man ein Pseudo-Spektrum, dass sehr schnell Peakcluster mit hoher Kovarianz und hoher Korrelation erkennen lässt. Abbildung B.3 im Anhang demonstriert das Potenzial dieser Methode am Beispiel der Daten des Proof of Concept Experiments.
Der intuitive Ansatz der AcorA Methode hat den Vorteil, dass sie auch mit basalen Statistik- und Programmierkenntnissen leicht anzuwenden ist. Als nicht-parametrische Methode entfällt bei der Spearman-Rangkorrelation aufwendiges Testen auf Normalverteilung sowie gegebenenfalls eine Transformation der Daten. Zudem werden die Signale nicht nach ihrer Intensität gewichtet. Auf diese Weise wird die Diskriminierung von kleinen Signale verhindert. Im Gegensatz zu den deutlich komplexeren multivariaten Methoden werden die Beziehungen zwischen den einzelnen Signalen bei der Berechnung der Korrelationskoeffizienten jedoch nicht berücksichtigt.
In dem von Gohr programmierten R-Paket „AcorA Version 1.0“ wird die Korrelationsanalyse

mit den exakten Bioaktivitätswerten durchgeführt. Bei vielen Bioaktivitätsassays sind kleine Aktivitätsdifferenzen jedoch qualitativ unerheblich. So unterscheidet sich beispielsweise eine Wachstumsinhibition von 84 % in der Praxis nur marginal von einer Wachstumsinhibition von 82 %. Zudem sind die Bioaktivitätsmessungen immer auch mit Messfehlern behaftet. Als Variante zur bestehenden AcorA Methode könnte die Rangkorrelation daher mit gruppierten (z. B. 0 - 20 % = inaktiv, 21 - 40 % = schwach aktiv, etc.) anstatt mit kontinuierlichen Bioaktivitätswerten durchgeführt werden. Dieses Vorgehen könnte insbesondere bei unpräzisen Bioaktivitätsassays mit hohen Standardabweichungen zu besseren Ergebnissen führen.

Eine weitere Möglichkeit zur Verbesserung der Aktivitäts-Korrelations-Analyse besteht bei der Berechnung der Signifikanzschwelle. In der „AcorA Version 1.0“ wird der kritische Korrelationskoeffizient durch einen Permutationstest festgelegt. Alle Korrelationskoeffizienten, die größer sind als der über den Permutationstest ermittelte Schwellenwert, werden als statistisch signifikant angesehen. Allerdings tritt dabei ein multiples Testproblem (α-Fehler Inflation) auf, das in der „AcorA Version 1.0“ noch nicht berücksichtigt wurde. Durch das n-fache Testen können daher vermehrt falsch positive Signale in der Hitliste auftreten. Dies gilt insbesondere für LC-MS Datensätze, bei denen naturgemäß eine große Anzahl von Datenpunkten pro Messung auftreten.

Durch entsprechende Korrekturverfahren wie beispielsweise die Bonferroni Korrektur [327] (sehr konservativ) oder unter der Festlegung einer vorgegebenen False-Discovery-Rate (Benjamini-Hochberg Methode) [328] könnte die Anzahl der falsch positiven Signale möglicherweise verringert werden. Da der Permutationstest die Nullverteilung des Datensatzes approximiert, kann der p-Wert für ein Signal aus dem Anteil der Permutationen berechnet werden, die einen Korrelationskoeffizienten besitzen, der größer oder gleich dem Korrelationskoeffizienten des Signals ist [329].

Die Erhebung einer Signifikanzschwelle aufgrund des p-Wertes wird jedoch immer wieder kritisiert [330]. U. a. weil die Aussagekraft des p-Wertes massiv von der Teststärke (die Wahrscheinlichkeit mit dem Testsystem einen vorliegenden Effekt entdecken zu können) und der *a priori* Wahrscheinlichkeit abhängt, dass tatsächlich eine echter Effekt bzw. eine echte Korrelation vorliegt [331].

Alternativ könnte die Beurteilung einer signifikanten Korrelation daher auch über die Berechnung von Konfidenzintervallen mithilfe der Bootstrapping Methode erfolgen [332]. Liegt das berechnete Konfidenzintervall eines Korrelationskoeffizienten außerhalb des Nullwertes, kann das Signal als statistisch signifikant eingestuft werden.

Die Ermittlung einer statistisch signifikanten Korrelation bleibt jedoch immer nur ein Hinweis

auf eine kausale Korrelation, die experimentell verifiziert werden muss. Nach der Durchführung von AcorA und der Auswertung der Hitliste muss die potentiell aktive Substanz isoliert und mit einem entsprechenden Assay auf seine Bioaktivität getestet werden.

3.2. Multivariate Methoden zur Datenanalyse

Die Datenanalyse in der vorgestellten AcorA Methode beruht auf einer Spearman Rangkorrelation. Die Funktionsweise der Spearman Rangkorrelation ist vergleichsweise intuitiv und auch die Durchführung der Aktivitäts-Korrelations-Analyse erfordert keine vertieften Kenntnisse in Statistik oder Programmierung.
Deutlich komplexer sind verschiedene Methoden der multivariaten Datenanalyse. Dabei werden Regressionsmodelle aufgrund von Trainigsdatensätzen erzeugt. Die Regressionsmodelle dienen anschließend zur Vorhersage von Bioaktivitäten unbekannter Proben oder auch zur Quantifizierung der Inhaltsstoffe eines pharmazeutischen Produkts [140].
Die während der Regressionsanalyse gewonnenen Regressionskoeffizienten und Faktorladungen (Loadings) geben darüber Aufschluss, wie stark der Beitrag einer bestimmte Variable zur Schätzung eines Regressanden ist. Wird ein Regressionsmodell zur Schätzung einer Bioaktivität y verwendet, lässt sich vermuten, dass Variablen, die einen hohen Beitrag zur Vorhersage der Bioaktivität haben, eben mit dieser selbst auch korreliert sind. Insbesondere bei großen Datenmatrizen mit vielen Variablen nimmt der Anteil der Kollinearität jedoch stark zu. Dadurch steigt die Varianz der Regressionskoeffizienten, so dass die Selektion der Koeffizienten nach Größe u. U. nur bedingt geeignet ist, um einen entsprechenden Zusammenhang zur Bioaktivität herzuleiten [306, 333].
Abhilfe schaffen s. g. Regularisierungsmethoden wie z. B. Ridge Regression, Lasso oder Elastic Net. Hierbei werden für das Modell eher unwichtige, d. h. kleine Regressionskoeffizienten in Abhängigkeit eines Schwellenparameters λ in Richtung Null geschrumpft (Ridge Regression) oder direkt auf Null gesetzt (Lasso). Auf diese Weise werden Variablen, die mit hoher Wahrscheinlichkeit nicht mit der Bioaktivität assoziiert sind, ausselektiert.
In den folgenden Abschnitten sollen nun verschiedene Methoden der multivariaten Datenanalyse zur Identifizierung von unbekannten Substanzen in Naturstoffextrakten qualitativ verglichen werden. Als Grundlage hierzu dient der Datensatz aus dem Proof of Concept Experiment. Insgesamt 47 der 1393 Peaks stammen von den Antibiotika Rifampicin und Erythromycin. Da das dritte Antibiotikum, Amoxicillin, in für *Bacillus subtilis* unwirksamen Konzentrationen eingesetzt wurde, sind dessen Massensignale für die weitere Betrachtung unerheblich.
Die Qualität einer Analysemethode hängt davon ab, wie gut der Experimentator zwischen Signalen von biologisch aktiven und biologisch inaktiven Verbindungen differenzieren kann. Bei den hier vorgestellten Datenanalysemethoden erhält jedes Massensignal einen Zahlenwert, der den Zusammenhang zwischen Bioaktivität und Massensignal repräsentiert (z. B.

Regressionskoeffizient, Loadings). Sortiert man anschließend die Ergebnisliste anhand der Größe des Zahlenwerts, sollten im Idealfall alle 47 gesuchten Antibiotikapeaks unter den ersten 47 Ergebnissen liegen. Dies entspräche in einer ROC-Kurve einem AUC-Wert von 1 (100 % Sensitivität, 100 % Spezifität). Die Akkumulation der Antibiotikapeaks in dem oberen Abschnitt der Ergebnisliste wird im Folgenden als Korrektklassifikationsrate bezeichnet. Je höher die Korrektklassifikationsrate, desto höher sind Sensitivität und Spezifität und desto höher ist die Güte der Analysemethode.

3.2.1. Hauptkomponentenanalyse (PCA)

Die Hauptkomponentenanalyse ist ein Standardverfahren in der Analyse von Metabolomics Datensätzen. Da hier die Regressanden nicht in die Analyse einbezogen werden, wird sie auch als unüberwachte Methode bezeichnet. Üblicherweise wird der Datensatz zunächst z-transformiert, d. h. zentriert und anhand der Standardabweichung skaliert. Abbildung 3.8 zeigt das Ergebnis der Hauptkomponentenanalyse nach z-Transformation. Es sind insgesamt vier Cluster erkennbar, die die vier verwendeten *Hygrophorus* Grundextrakte repräsentieren. So bildet beispielsweise das linke Cluster mit den Extrakten HE2, HEA6, HEA10 und HEA12 den *Hygrophorus* Extrakt 2 (*H. chrysodon*) ab. Extrakt 1 (*H. lucorum*) ist mit den Extrakten HE1, HEA14, HEA15 oben rechts und etwas tiefer mit HEA9 erkennbar. Cluster 3 (HE3, HEA5, HEA7, HEA 13, HEA16) entspricht dem Extrakt aus *H. agathosmus*. Bemerkenswert ist, dass die erste Hauptkomponente den Extrakt 2 (*H.chrysodon)* von den übrigen Extrakten separiert, während die zweite Hauptkomponente die Extrakte der anderen drei Pilzarten auftrennt. Möglicherweise beruhen diese Separierungen auf phylogenetischen Beziehungen, die jedoch hier nicht weiter erörtert werden sollen. Die gewünschte Trennung der Extrakte aufgrund der antibakteriellen Aktivität kann indes nicht beobachtet werden. Dies wird auch durch die entsprechenden ROC-Kurven (Abb. 3.10) bestätigt. Die AUC-Werte von 0,67 (PC1) und 0,48 (PC2) entsprechen einem stochastischen Prozess. Der Grund für die schlechte Korrektklassifikationsrate der z-transformierten Daten ist in den Loadings Plots erkennbar (Abb. 3.9). Durch die Skalierung mit der Standardabweichung wird die für AcorA benötigte Varianz - insbesondere der großen Peaks - stark reduziert, so dass die Antibiotikapeaks im „Rauschen“ der kleineren Peaks verborgen bleiben. Eine Hauptkomponentenanalyse auf Basis der z-transformierten Daten eignet sich daher nicht um die aktiven Komponenten detektieren zu können.
Deutlich bessere Ergebnisse lieferte die PCA der zentrierten Daten. Die Trennung von aktiven und inaktiven Proben ist bereits im Scores-Plot erkennbar (3.8). Extrakte mit hohen

Konzentrationen an Erythromycin (HEA04, HEA09) werden anhand der ersten Hauptkomponente von den übrigen Extrakten getrennt. Extrakte mit hohen Rifampicin Konzentrationen (HEA01, HEA15, HEA02, HEA16) werden über die zweite Hauptkomponente als einzelne Cluster klar von den Extrakten mit niedriger antibakterieller Aktivität separiert. Die bessere Trennung der aktiven Komponenten spiegelt sich auch in den ROC Kurven wider. Der AUC-Wert für die erste Hauptkomponente der mittenzentrierten Daten liegt mit 0,90 deutlich über denen der z-transformierten Daten.

Der Loadings Plot (Abb. 3.9) zeigt, dass die *m/z* Peaks der aktiven Komponenten zumeist sehr deutlich herausragen. Wie bereits im Scores Plot ersichtlich, wird die erste Hauptkomponente sehr stark von den Peaks des Erythromycins dominiert, während in der zweiten Hauptkomponente die Rifampicin Peaks die größten Werte besitzen. Gleichzeitig sind in der zweiten Hauptkomponente die Vorzeichen für die Erythromycin und Rifampicin Peaks entgegengesetzt. Daraus werden verschiedene Probleme der Hauptkomponentenanalyse ersichtlich. Zum einen werden die Informationen über die aktiven Peaks über die verschiedenen Hauptkomponenten verteilt, dies bedeutet in diesem Fall, Erythromycin Peaks werden überwiegend über die erste Hauptkomponente detektiert und Rifampicin Peaks über die

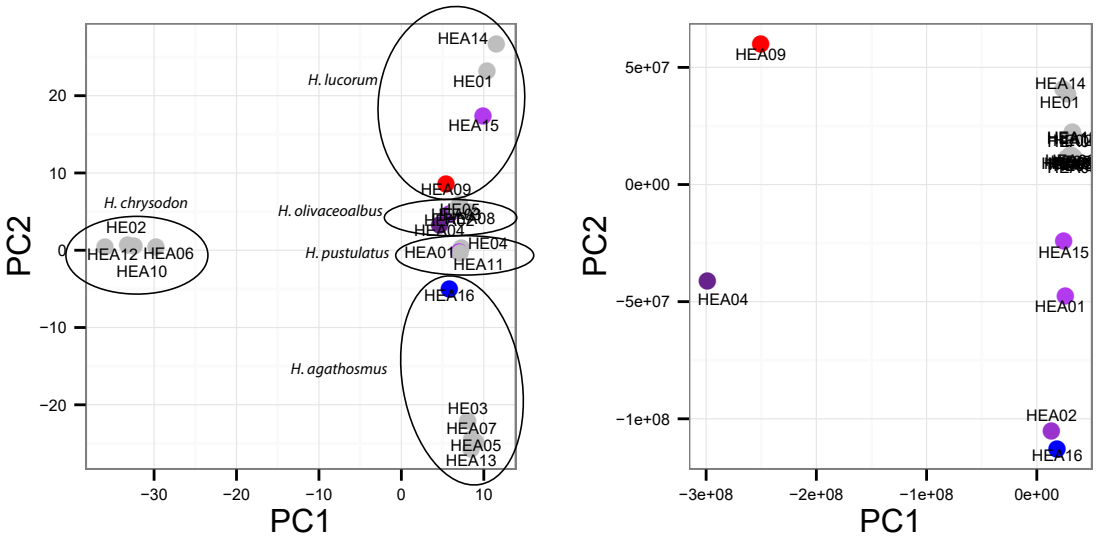


Abbildung 3.8.: Scores Plot der Hauptkomponentenanalyse. Links nach z-Transformation. Rechts nach Zentrierung der Daten. Die Proben mit aktivitätsrelevanten Konzentrationen der Antibiotika sind farblich markiert. Blau: hohe Konzentration Rifampicin, Rot: hohe Konzentration Erythromycin, Violett: hohe Konzentration Rifampicin und Erythromycin.

zweite Hauptkomponente. Der Zuwachs an erklärter Varianz bringt also keinen Vorteil für die Erfassung der bioaktiven Verbindungen. Zum anderen beziehen sich die Vorzeichen der Loadings nur relativ auf den Ursprung des Koordinatensystems und nicht auf die Richtung der Korrelation. Da in der PCA der Bezug zu den Regressanden fehlt, kann eine positive Korrelation daher sowohl durch ein positives als auch ein negatives Vorzeichen der Loadings gekennzeichnet sein. Der Experimentator hat jedoch *a priori* keine Kenntnis darüber, welches Vorzeichen für die Detektion der bioaktiven Komponenten relevant ist.
Im Gegensatz dazu wird bei AcorA die Richtung der Korrelation durch das Vorzeichen bestimmt. Unter anderem auch aus diesem Grund zeigt AcorA eine bedeutend bessere Korrektklassifkationsrate (AUC= 0,99) gegenüber der ersten (AUC = 0,90) und zweiten Hauptkomponente (AUC = 0,67) der mittenzentrierten Daten.

Zusammenfassung Zusammenfassend lässt sich sagen, dass die Hauptkomponentenanalyse aufgrund des fehlenden Bezugs zu den Regressanden nur bedingt für die Selektion von bioaktiven Verbindungen in komplexen Mischungen geeignet ist. In jedem Fall sollte eine übermäßig starke Skalierung der Daten vermieden werden.

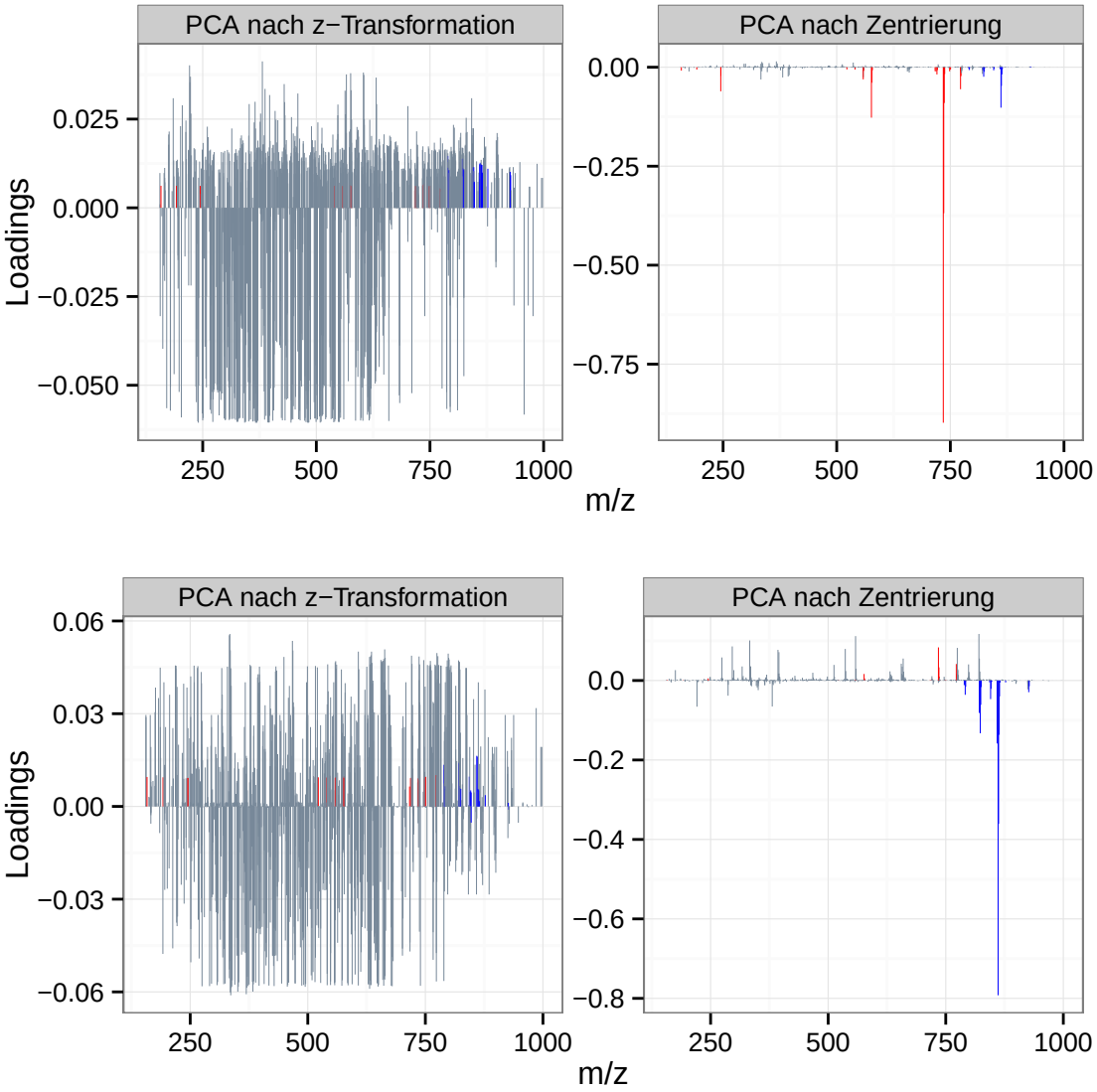


Abbildung 3.9.: Loadings Plot der Hauptkomponentenanalyse. Oben die Loadings der 1. PC, unten die Loadings der 2. PC. Links nach z-Transformation. Rechts nach Zentrierung der Daten. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und den nicht aktiven Verbindungen (Grau) sind farblich gekennzeichnet.

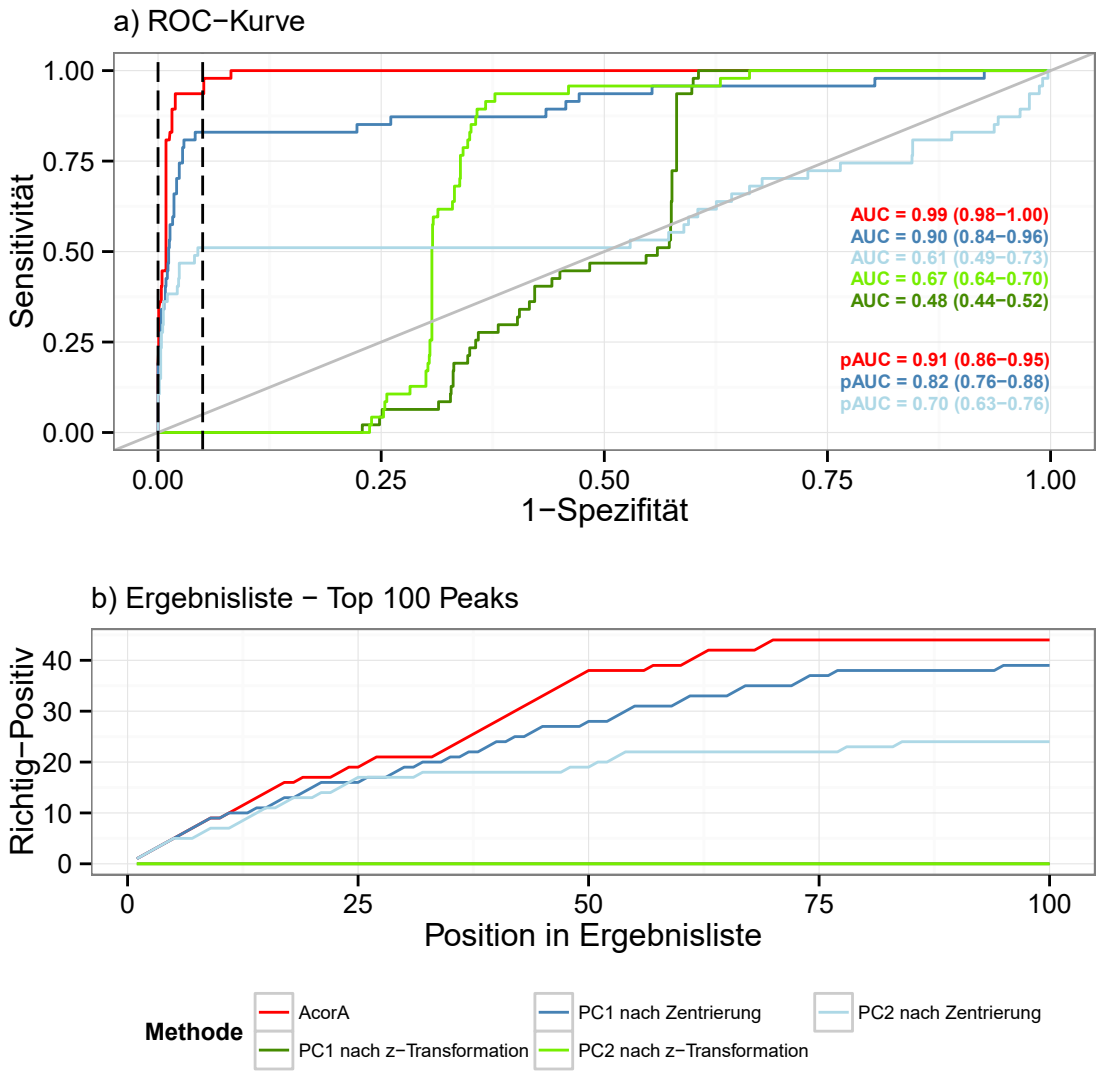


Abbildung 3.10.: a) ROC-Kurven der Hauptkomponentenanalyse nach Zentrierung bzw. z-Transformation der Daten. Als Vergleich ist das Ergebnis von AcorA (Rot) dargestellt. Die 95% Konfidenzintervalle der AUCs und partiellen AUCs sind in Klammern angegeben. Die pAUCs beziehen sich auf den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate, der durch die gestrichelten Linien gekennzeichnet ist. b) Anzahl der annotierten Antibiotikapeaks in Abhängigkeit der Position in der Ergebnisliste.

3.2.2. Hauptkomponentenregression (PCR)

Die Hauptkomponentenregression basiert auf einer Hauptkomponentenanalyse mit anschließender multipler linearer Regression. Anstatt wie in der linearen Regression üblich, die hochdimensionale Originaldatenmatrix **X** für die Schätzung der Regressanden **y** zu verwenden, wird die Score-Matrix **T** als Ausgangpunkt für die Regression genutzt. Auf diese Weise werden Kollinearitäten vermieden und nur die Variablen mit der vermeintlich höchsten „Informationsdichte“ für die Regression verwendet. Die Größe der Regressionskoeffizienten kann als Maß für die Korrelation (bzw. bei nicht z-transformierten Daten als Kovarianz) zwischen den *m/z* Peaks und der Bioaktivität betrachtet werden.
Die Anzahl der relevanten Hauptkomponenten wurde durch repetitive doppelte Kreuzvalidierung bestimmt. Nach 100 Wiederholungen zeigte die zweite Hauptkomponente in den meisten Fällen den geringsten MSECV innerhalb einer Standardabweichung und wurde somit in der anschließenden PCR verwendet. Mit der zweiten Hauptkomponente werden 71 % der Bioaktivität und 81 % der Variation in den x-Variablen erklärt. Die erste Hauptkomponente erklärt deutlich weniger der Varianz in den Daten (y = 29 %, X = 66 %).

Abbildung 3.11 zeigt die Leistungsfähigkeit der PCR für die Identifizierung der bioaktiven Massensignale anhand der ersten und zweiten Hauptkomponente im Vergleich zu AcorA. Mit einem AUC Wert von 0,94 weist die Verwendung der zweiten Hauptkomponente - wie anhand der Kreuzvalidierung erwartet - eine bessere Korrektklassifikationsrate der Antibiotikapeaks auf als die Verwendung der ersten Hauptkomponente (AUC = 0,90). Der Unterschied ist jedoch mit einem p-Wert von 0,47 als nicht signifikant zu bezeichnen. Betrachtet man die partiellen AUC-Werte im Bereich zwischen 95 und 100 % Spezifität, ist der Unterschied zwischen den beiden Hauptkomponenten noch geringer (ΔpAUC = 0,01, p = 0,92). Wie bereits der Scores Plot (Abbildung 3.8) gezeigt hat, findet in der ersten Hauptkomponente hauptsächlich eine Trennung der erythromycinhaltigen Proben statt. In der zweiten Hauptkomponente hingegen werden vor allem mit Rifampicin versetzte Proben von den inaktiven Proben separiert. Dies spiegelt sich auch in der Größe der Regressionskoeffizienten wider (Abbildung 3.12). Während die durch das Erythromycin verursachten Massensignale in der ersten Hauptkomponente die größeren Regressionskoeffizienten besitzen, nimmt in der zweiten Hauptkomponente vor allem die Größe der Regressionskoeffizienten der Rifampicin Peaks zu. Da die Berechnung der Regressionskoeffizienten auf Basis der mittenzentrierten Daten - und somit über die Kovarianz zwischen x- und y-Variablen - erfolgte, hängt die Größe der Regressionskoeffizienten wiederum von der Signalintensität der x-Variablen ab. Aus diesem Grund erhalten größere Massensignale einen größeren Regressionskoeffizienten,

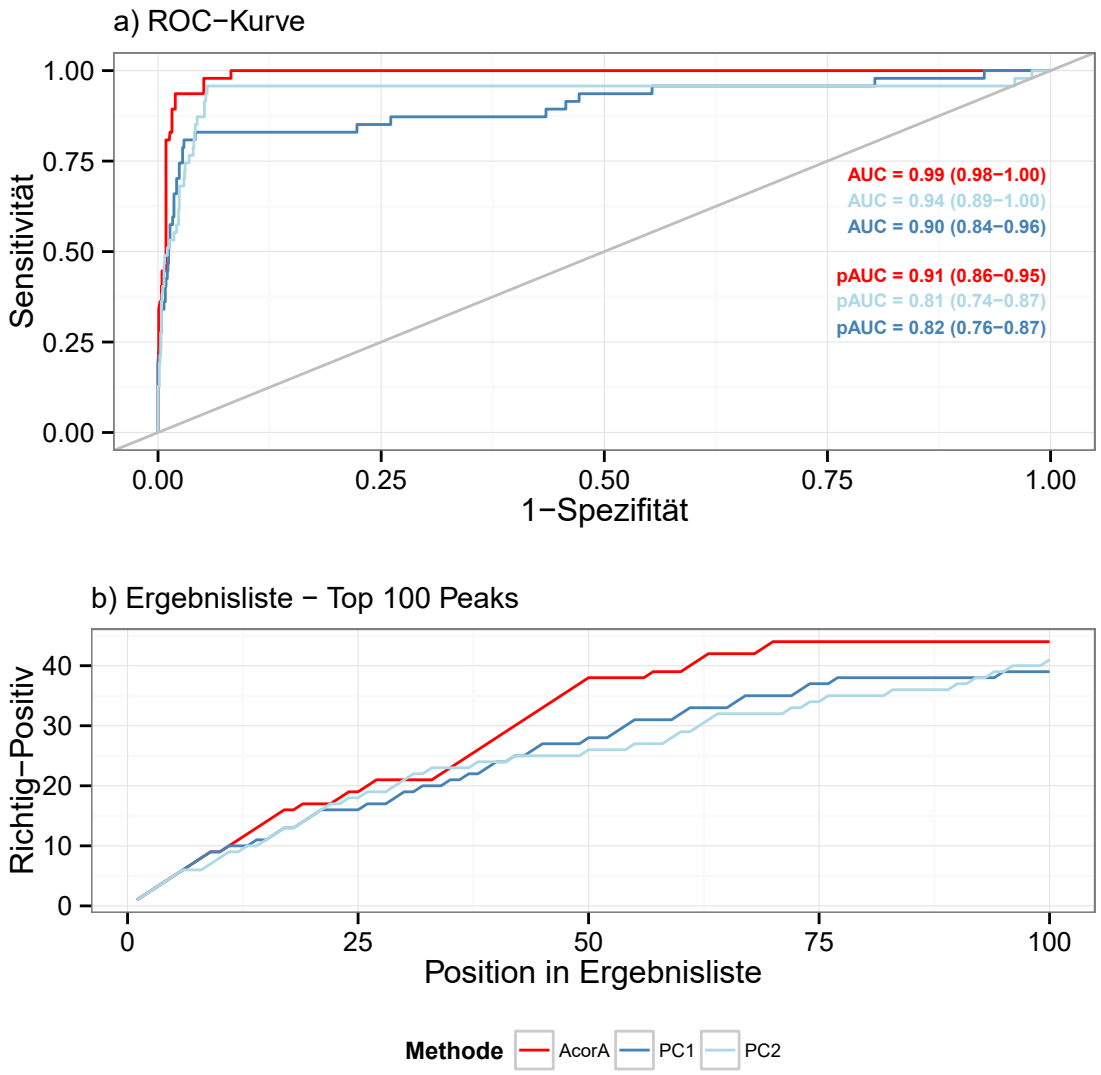


Abbildung 3.11.: a) ROC-Kurven Hauptkomponentenregression. PC1 (Blau) und PC2 (Hellblau) der Daten. Als Vergleich ist das Ergebnis von AcorA (Rot) dargestellt. Die 95% Konfidenzintervalle der AUC- und pAUC-Werte sind in Klammern angegeben. Die pAUC-Werte beziehen sich auf den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate, der durch die gestrichelten Linien gekennzeichnet ist. b) Anzahl der annotierten Antibiotikapeaks in Abhängigkeit der Position in der Ergebnisliste.

ohne das daraus zwingend eine stärkere Korrelation abgeleitet werden kann [140, 159]. Die Ursache für die deutlich bessere Korrektklassifikationsrate der PCR gegenüber der

PCA wird ebenfalls in Abbildung 3.12 klar erkennbar. Durch Verwendung der Bioaktivität als Regressandenvektor wird in der PCR ein positiver Zusammenhang zwischen Bioaktivität und Signalintensität der Antibiotikapeaks hergestellt. Peaks beider Substanzen werden somit nicht wie bei Verwendung der Loadings über die Achsen verteilt, sondern auf dieselbe Achse projiziert. Da - wie erwartet - viele der Antibiotikapeaks eine höhere Kovarianz zur biologischen Aktivität besitzen als die Peaks der meisten nicht-aktiven Verbindungen, werden nach Größensortierung der Koeffizienten, vor allem die Antibiotikapeaks im oberen Abschnitt der Ergebnisliste angereichert. So befinden sich nach Regression auf die zweite Hauptkomponente 26 der gesuchten Antibiotikapeaks unter den ersten 50 Einträgen der Ergebnisliste. Erweitert man diesen Bereich auf die ersten 100 Einträge, werden 41 der 47 gesuchten Antibiotikapeaks gefunden.
Die Klassifizierungsrate der PCR ist mit einem AUC-Wert von 0,94 geringer als bei der von AcorA verwendeten Spearman Rangkorrelation (AUC= 0,99). Im Vergleich zur zweiten Hauptkomponente gibt es keinen signifikanten Unterschied (p = 0,09). In dem für den Experimentator besonders interessanten Bereich zwischen 95 und 100 % Spezifität ist AcorA der PCR überlegen. Der Unterschied von AcorA (pAUC = 0,91) zu PC1 (pAUC = 0,82, p = 0,03) und PC2 (pAUC = 0,81, p = 0,01) ist in beiden Fällen signifikant.

Zusammenfassung Insgesamt scheint die Hauptkomponentenregression zur Identifizierung von biologisch aktiven Substanzen in komplexen Mischungen gute Ergebnisse zu liefern. Die AUC Werte von 0,90 und 0,94 für die erste und zweite Hauptkomponente sprechen für eine gute bis sehr gute Klassifizierungsmethode. Problematisch ist jedoch die Fixierung der PCR auf die Varianz der Daten. Zum einen muss die hohe Varianz eines Massensignals nicht mit der biologischen Aktivität korrelieren. Zum anderen führt die Berechnung der Regressionskoeffizienten über die Kovarianzmatrix ($b = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$) zur Diskriminierung kleinerer Peaks, d. h. biologisch aktive Substanzen in niedriger Konzentration (oder mit geringen Ionisationseigenschaften) könnten unter Verwendung der PCR übersehen werden.

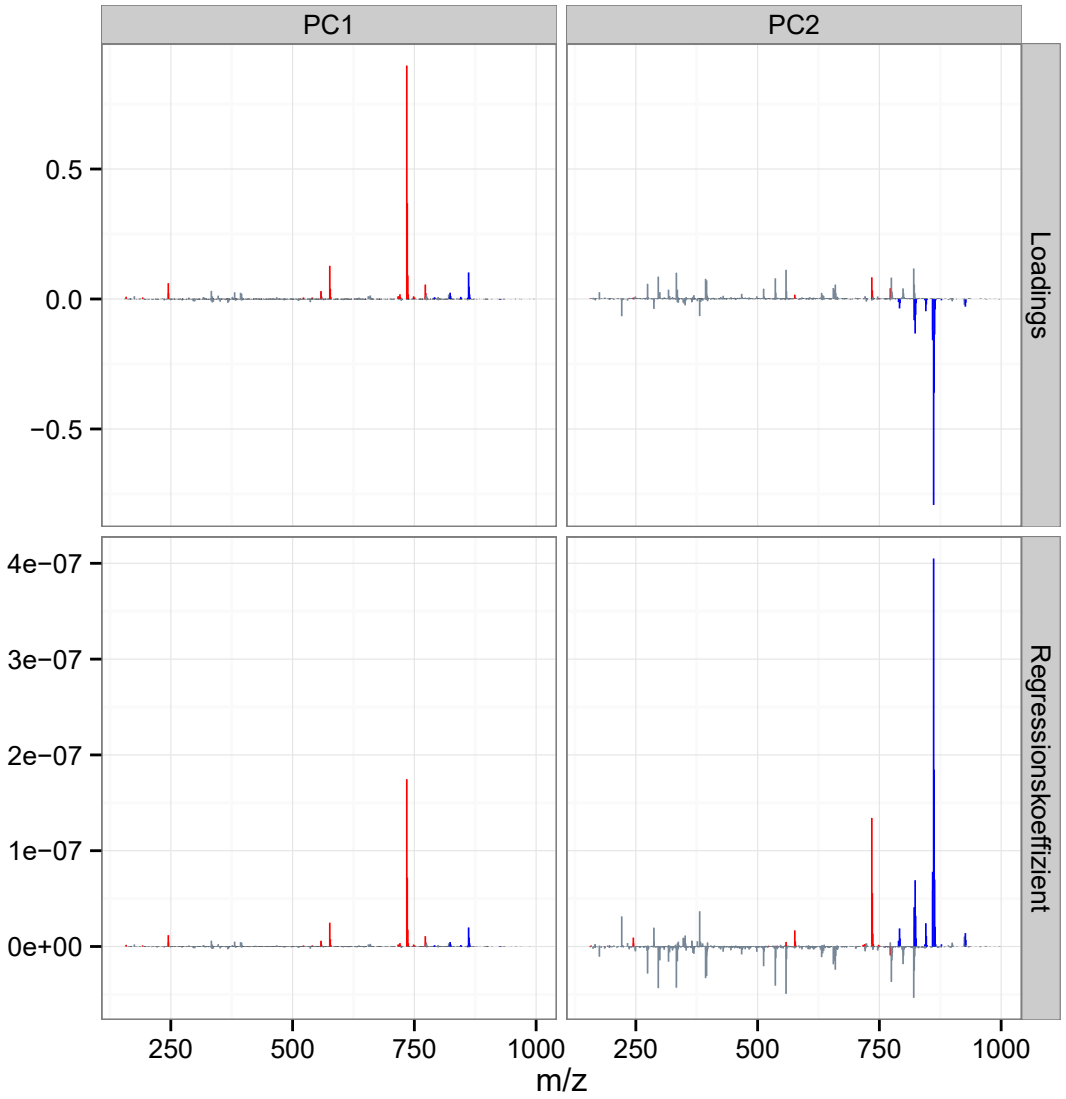


Abbildung 3.12.: Plot der Loadings und Regressionskoeffizienten der *m/z* Werte von PC1 und PC2. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und der nicht aktiven Verbindungen (Grau) sind farblich gekennzeichnet.

3.2.3. Partial-Least-Squares Regression (PLSR)

Die Partial-Least-Squares Regression ist ein weiteres in der Chemometrie und in der Metabolomics vielfach eingesetztes Verfahren zur multivariaten Regression. Ähnlich der PCR wird die Matrix der Originalvariablen (**X**) durch eine Score-Matrix **T** ersetzt. Im Gegensatz zur PCR fließen jedoch über einen Wichtungsvektor **w** Informationen aus der y-Variablen (Bioaktivität) in die Berechnung der Scores und Loadings mit ein. Auf diese Weise sollte die Vorhersage der Bioaktivität verbessert werden.
Bei der PLSR-Beta Methode dient die Größe des Regressionskoeffizienten b einer Variable x als Maß für die Wichtigkeit der Variable in einem Modell. Variablen mit großen Regressionskoeffizienten werden daher als besonders interessant für die Bioaktivtät eingestuft. Bei der Variable Importance in Projection (VIP-Methode) wird zur Abschätzung der Bedeutung einer Variable für das Modell über Gleichung 2.46 der Anteil der erklärten Varianz ihres Wichtungsvektors **w** an der gesamten erklärten Varianz für jede Hauptkomponente miteinbezogen. Im Allgemeinen wird für den VIP ein Schwellenwert zwischen 0,83 und 1,21 (häufig 1) gesetzt, d. h. Variablen größer als der Schwellenwert werden für das Modell als besonders wichtig erachtet und selektiert [334].
In beiden Fällen wurde die Anzahl der relevanten latenten Variablen durch repetitive doppelte Kreuzvalidierung bestimmt. Nach 100 Wiederholungen zeigte die zweite latente Variable in den meisten Fällen den geringsten MSECV innerhalb einer Standardabweichung und wurde somit in der anschließenden PLSR verwendet. Die LV2 erklärt 82 % der Bioaktivität und 81 % der Variation in den massenspektrometrischen Daten. Der Scores Plot der ersten beiden latenten Variablen LV1 und LV2 deutet eine im Vergleich zu PCA/PCR verbesserte Trennung der antibiotikahaltigen Proben an (Abbildung 3.13). Bereits durch die LV1 werden die biologisch aktiven Proben von den inaktiven Proben getrennt. Ähnlich der PCA wird auch hier die erste Komponente von den erythromycinhaltigen Proben bestimmt, während die zweite latente Variable (LV2) zu einer stärkeren Separierung der Rifampicin Proben führt. Der mittlere Fehler zur Vorhersage der antibiotischen Wirkung von unbekannten Extrakten ist mit 28,4 gegenüber der PCR leicht verbessert.

PLSR-Beta Die verbesserte Trennung der antibiotikahaltigen Proben macht sich auch in einer verbesserten Identifizierung der bioaktiven Komponenten bemerkbar (Abbildung 3.14). Die ersten beiden latenten Variablen weisen mit AUC-Werten von 0,99 (LV1) und 0,97 (LV2) eine sehr hohe Klassifizierungsrate auf. Unter den ersten 50 Peaks der Ergeb-

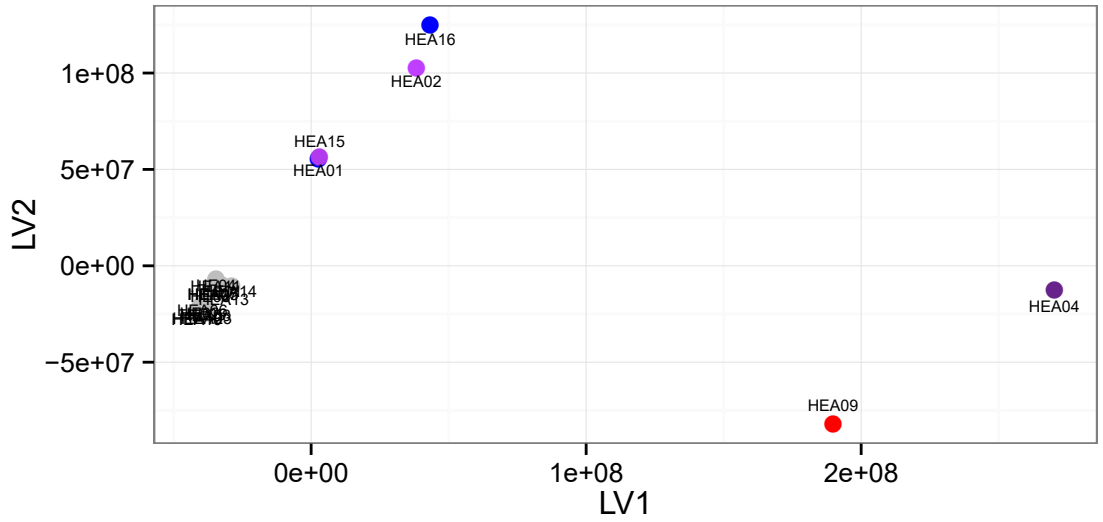


Abbildung 3.13.: Scores Plot der latenten Variablen LV1 und LV2. Die Proben mit aktivitätsrelevanten Konzentrationen der Antibiotika sind farblich markiert. Blau: hohe Konzentration Rifampicin, Rot: hohe Konzentration Erythromycin, Violett: hohe Konzentrationen Rifampicin und Erythromycin.

nistliste befinden sich 30 (LV1) bzw. 25 (LV2) der gesuchten Antibiotikapeaks. Betrachtet man die ersten 100 Peaks der Ergebnisliste, findet die PLSR anhand der LV1 nahezu alle gesuchten Peaks (45 von 47). Entsprechend spiegelt sich die hohe Wiederfindungsrate auch in den partiellen AUC-Werten (LV1 = 0,88, LV2 = 0,87) wider.
Der Grund für die sehr gute Klassifizierungsrate der PLSR ist in Abbildung 3.15 zu erkennen. Bereits in der LV1 erhalten die Massensignale von Erythromycin und Rifampicin relativ große Regressionskoeffizienten und setzen sich deutlich von den Peaks der nicht-aktiven Verbindungen ab. In der LV2 nimmt insbesondere die Größe der Regressionskoeffizienten des Rifampicins zu. Allerdings zeigt sich in der LV2 bereits ein gewisser Grad an Überanpassung (Overfitting), d. h. der Anteil der irrelevanten Regressoren ist angestiegen, was sich an einem parallelen Anstieg der Regressionskoeffizienten der nicht-aktiven Verbindungen äußert und zu einer geringfügig niedrigeren Klassifizierungssrate führt.

Variablenselektion mit der Variable Importance On Projection (VIP) Die VIP ist eine verbreitete Methode zur Wichtung der Variablen im Anschluss an eine Partial-Least-Squares Analyse und eine Alternative zu der oben beschriebenen Verwendung der Regressionskoeffizienten.

Bei dem hier untersuchten Datensatz ist die Klassifizierungsrate der VIP mit AUC-Werten von 0,96 und 0,97 für LV1 und LV2 ähnlich hoch, wie die bei der Beta-Methode. Insbesondere im Bereich hoher Spezifität (95-100 %) liefert die VIP jedoch signifikant schlechtere Ergebnisse als die Beta-Methode ($p = 7,6 \cdot 10^{-7}$) und AcorA ($p = 1,5 \cdot 10^{-3}$). So befinden sich unter den ersten 50 Einträgen lediglich 25 Antibiotikapeaks in der LV1 bzw. 23 in der LV2. Erweitert man diesen Bereich auf 100 Peaks findet man mit der VIP-Methode 36 (LV1) und 33 (LV2) Antibiotikapeaks. Der Grund für die die geringere Klassifizierungsrate wird in Abbildung 3.15 deutlich. Durch die VIP-Methode erhalten alle Variablen, unabhängig davon ob die zugehörigen Loadings, Scores oder Weights ein negatives Vorzeichen besaßen, VIP-Werte mit positiven Vorzeichen (man beachte die Quadrierungen der Y-Loadings **q**, Weights **w** und X-Scores **t** in Gleichung 2.46). Auf diese Weise werden Variablen, die in der PLSR ursprünglich eine negative Kovarianz besaßen mit Variablen positiver Kovarianz auf einer Achse vereinigt. Als Folge werden Variablen, die in negativem - in diesem Fall nicht erwünschtem - Sinne mit der Antwortvariablen korreliert sind, artifiziell in der Ergebnisliste nach oben verschoben, was zu mehr falsch-positiven Resultaten führt.
Insgesamt besitzen 23 Variablen einen VIP-Wert > 1 und können somit für das Modell als signifikant bezeichnet werden. Unter der 23 Variablen befinden sich 10 Peaks, die dem Rifampicin zuzuordnen sind, sowie 6 Peaks des Erythromycins. Die VIP-Methode verfügt somit über eine mittlere Genauigkeit von 69 %. Da lediglich 23 Peaks über dem kritischen Wert liegen, ist die Trefferquote, d. h. die Anzahl der detektierten Peaks im Verhältnis zur Anzahl der gesuchten Peaks, relativ niedrig. Der mittlere Vorhersagefehler RMSEP ist mit 24,9 gegenüber dem vollen PLSR Modell (28,4) verbessert.

Zusammenfassung Zusammfassend lässt sich sagen, dass sich die Partial-Least-Squares Regression sehr gut für die Identifizierung von biologisch aktiven Verbindungen in Mischungen eignet. Durch Verwendung eines Wichtungsvektors, über den Informationen über die Bioaktivität in die Kalkulation der latenten Variablen einfließen, konnte der Anteil der Antibiotikapeaks im oberen Abschnitt der Ergebnisliste gegenüber der PCR vergrößert werden. Wie bereits bei der PCR beobachtet, besteht jedoch auch bei der PLSR das Problem, dass kleinere Peaks aufgrund ihrer geringeren Varianz gegenüber großen Peaks negativ diskriminiert werden. Die daraus resultierenden kleineren Regressionskoeffizienten führen dazu, das diese Peaks erst im hinteren Teil der Ergebnisliste zu finden sind und damit bei der Dateninterpretation möglicherweise übersehen werden.
Die VIP-Methode scheint für die Identifizierung biologisch aktiver Verbindungen eher ungeeignet, da durch diese Methode korrelierende und reziprok korrelierende Signale in der

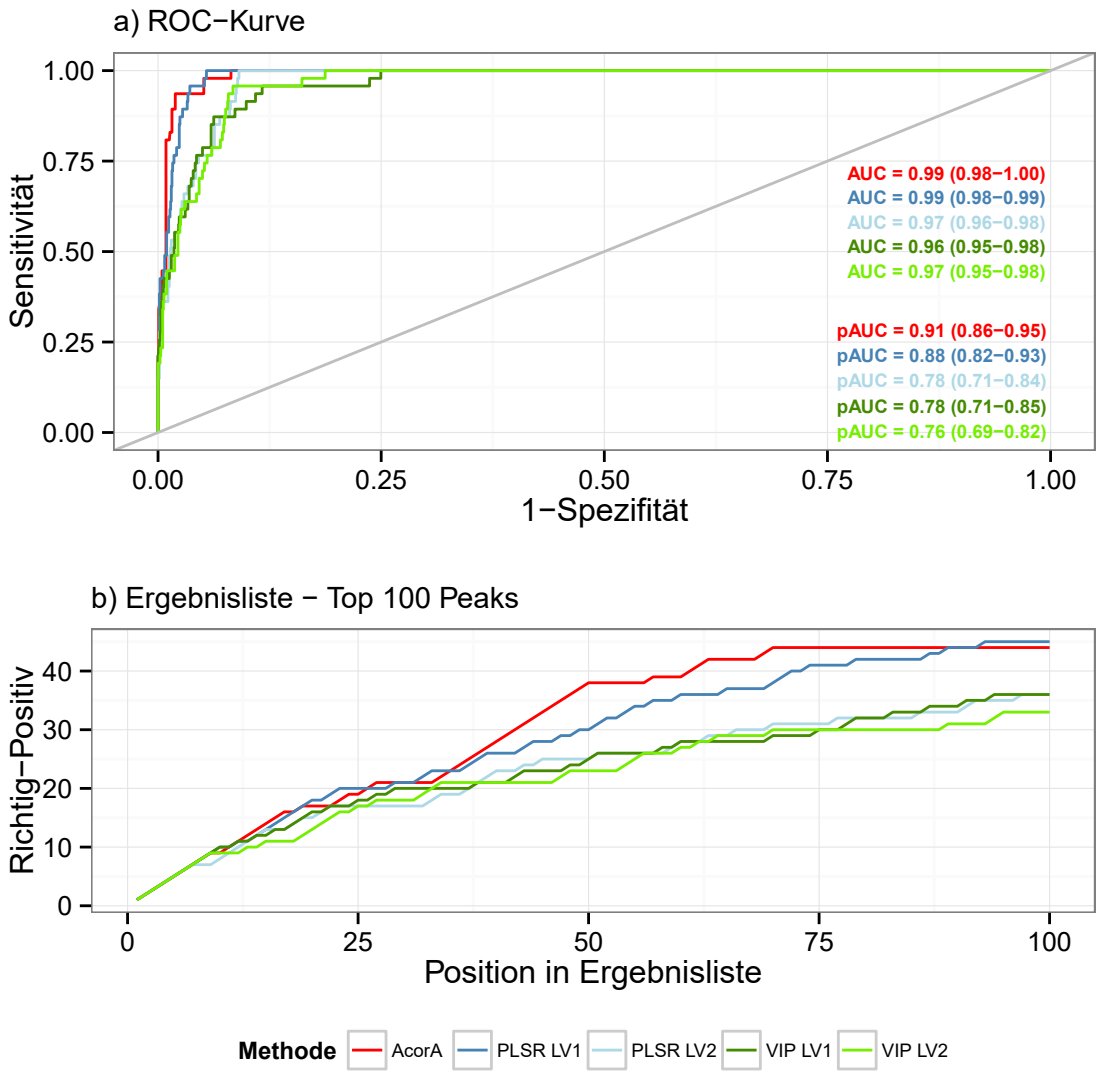


Abbildung 3.14.: a) ROC-Kurven der Partial Least Squares Regression. Als Vergleich ist das Ergebnis von AcorA (Rot) dargestellt. Die 95% Konfidenzintervalle der AUC- und pAUC-Werte sind in Klammern angegeben. Die pAUC-Werte beziehen sich auf den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate, der durch die gestrichelten Linien gekennzeichnet ist. b) Anzahl der annotierten Antibiotikapeaks in Abhängigkeit der Position in der Ergebnisliste.

Hitliste amalgamiert werden.

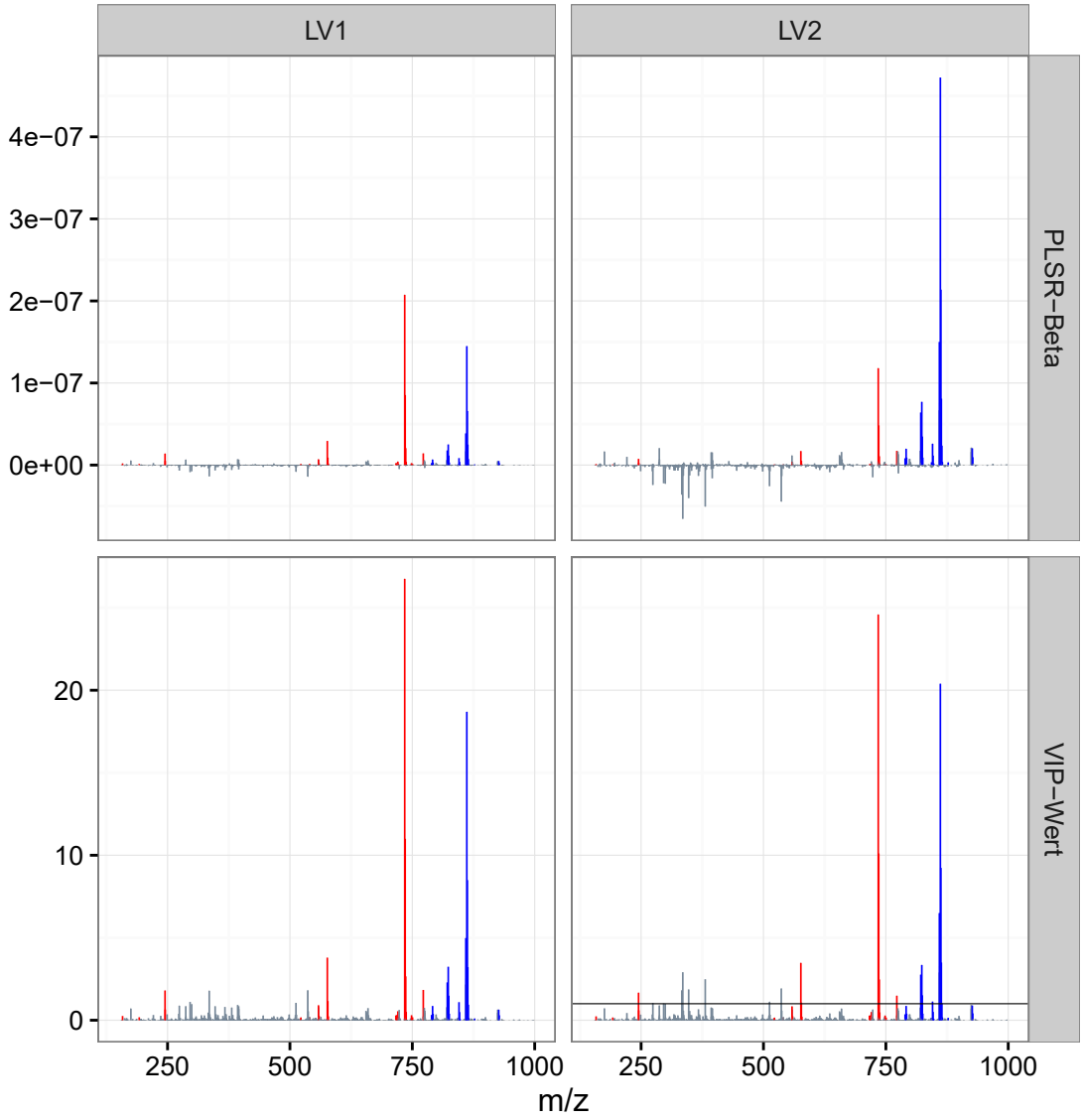


Abbildung 3.15.: Regressionskoeffizienten (Beta-Methode) und VIP-Werte der *m/z* Werte in LV1 und LV2. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und den nicht aktiven Verbindungen (Grau) sind farblich gekennzeichnet.

3.2.4. Quantitative Pattern-Activity Relationship (QPAR)

Die Quantitative Pattern-Activity Relationship (QPAR) Methode wurde 2008 von Chau *et al.* publiziert [160]. Sie verknüpft die von Kvalheim eingeführte Methode der Target Projection (TP) [154] mit der von Rajalahti vorgestellten Variablenselektionsmethode (SR-Plot) [158]. Bei der Target Projection wird das multivariate Regressionsmodell auf eine einzige (target projected) Komponente reduziert, indem versucht wird, die Richtung im multivariaten Spektrenraum zu finden, die am stärksten mit der Bioaktivität verknüpft ist. Durch die Reduktion auf nur eine Komponente gelingt es, die Information über die Bioaktivität, die bei den bisher dargestellten multivariaten Methoden über mehrere Hauptkomponenten/latente Variablen verteilt ist, zu bündeln. In einem zweiten Schritt werden anschließend die s. g. Selectivity Ratios (SR) berechnet, die die einzelnen Target Projected Loadings anhand dem Anteil der erklärten Varianz an der Restvarianz wichtet.

Das Ergebnis der QPAR-Methode ist als exzellent zu bezeichnen. Sowohl mit der Target Projection als auch nach der anschließenden Wichtung durch Berechnung der Selectivity Ratios, wird ein AUC-Wert von 0,99 erzielt. Interessant ist, dass insbesondere im Bereich hoher Spezifität (95-100 %) ein partieller AUC-Wert von 0,94 für Selectivity Ratios und Target Projection erreicht wird, d. h. beide Methoden erzielen bessere Ergebnisse als mit der bei AcorA verwendeten Spearman-Rangkorrelation. Unter den ersten 50 Peaks der Ergebnisliste befinden sich bei der Selectivity Ratio Methode 40, bei der Target Projection Methode 37 und bei AcorA 38 Antibiotikapeaks. Mithilfe der Target Projection und den Selectivity Ratios können innerhalb der ersten 100 Peaks der Ergebnisliste jeweils alle gesuchten Antibiotikapeaks gefunden werden. Bei AcorA sind es lediglich 44 der 47 gesuchten Antibiotikapeaks.

Abbildung 3.17 zeigt den Grund für das gute Ergebnis. Bereits nach der Target Projection sind viele der Antibiotikapeaks deutlich von den Peaks der nicht-aktiven Verbindungen abgesetzt. Diese besitzen entweder sehr kleine oder negative TP-Loadings. Da die Datenmatrix, wie bei den anderen multivariaten Methoden zuvor auch, mittenzentriert wurde, sind die TP-Loadings abhängig von der Größe eines Peaks. Durch die Kalkulation der Selectivity Ratios der TP-Loadings werden diese weitestgehend unabhängig von der Größe der Massensignale. Dies wird in Abbildung 3.17 daran deutlich, dass nahezu alle Peaks, die einer (aktiven) Substanz zugeordnet werden können, in etwa gleich große Werte erhalten. Insbesondere kleine Peaks erhalten dadurch mehr Gewicht und setzen sich von den Signalen der nicht-aktiven Verbindungen besser ab.

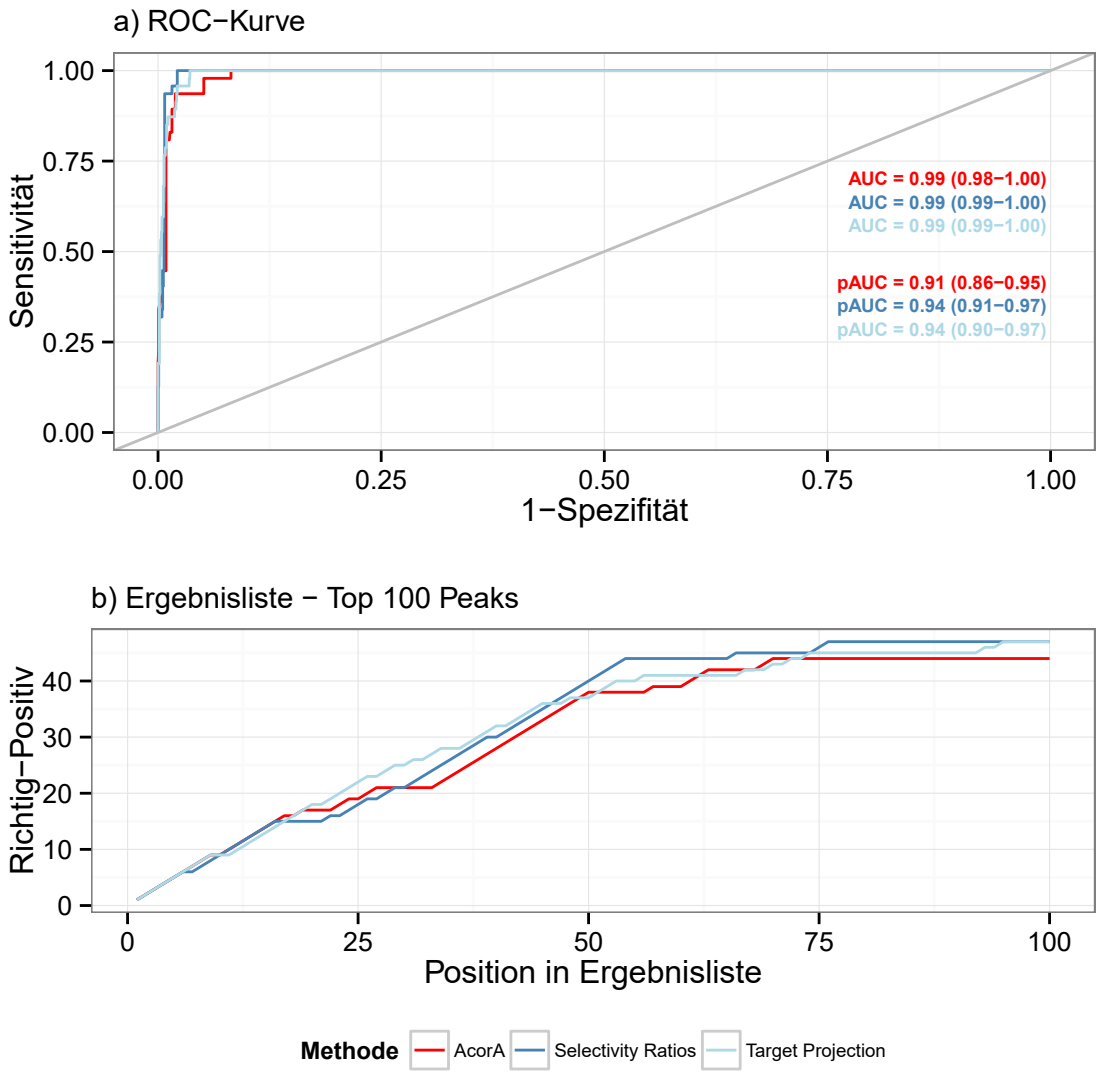


Abbildung 3.16.: a) ROC-Plot von QPAR, aufgeschlüsselt nach Target Projection und Selectivity Ratios. Als Vergleich ist das Ergebnis von AcorA (Rot) dargestellt. Die 95% Konfidenzintervalle der AUC und pAUC-Werte sind in Klammern angegeben. Die pAUC-Werte beziehen sich auf den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate, der durch die gestrichelten Linien gekennzeichnet ist. b) Anzahl der annotierten Antibiotikapeaks in Abhängigkeit der Position in der Ergebnisliste.

Bemerkenswert ist das Massensignal bei *m/z* 287,15266 mit einem SR von 4,33. Dieser hohe Wert deutet auf ein Signal einer bioaktiven Komponente hin. Der Peak konnte jedoch

weder in den *Hygrophorus* Grundextrakten HE1-5 noch in den einzeln gemessenen Antibiotikastammlösungen detektiert werden. Das Signal wird allerdings in den Massenspektren der Extrakte HEA01, HEA02, HEA15 und HEA16 beobachtet, d. h. exakt in jenen Extrakten, die mit der höchsten Konzentration an Rifampicin gespickt wurden. Es wäre daher möglich, dass es sich bei *m/z* 287,15277 um einen Peak aus der Rifampicin Lösung (und somit um einen richtig-positiv Peak) handelt. Dafür spricht auch die Höhe des Selectivity Ratios, das in einer ähnlichen Größenordnung liegt, wie die der anderen Rifampicin Peaks. Möglicherweise wurde der Peak in der Rifampicin Stammlösung durch Matrixeffekte supprimiert.

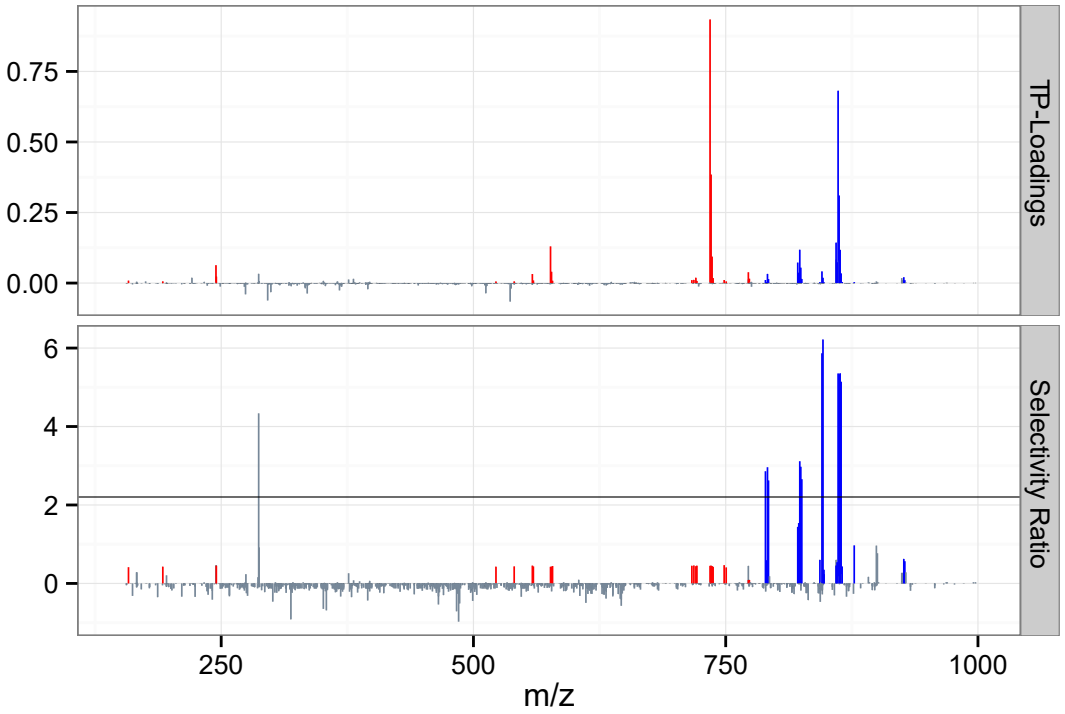


Abbildung 3.17.: Target Projected Loadings (TP-Loadings) und Selectivity Ratios für die *m/z*-Werte im Massenbereich zwischen *m/z* 150 und 1000. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und der nicht aktiven Verbindungen (Grau) sind farblich gekennzeichnet. Der aus dem F-Test berechnete Grenzwert für signifikant korrelierende Peaks F_{krit} = 2,2 ist im SR-Plot als horizontale Linie eingezeichnet.

Variablenselektion Um eine Grenze für die signifikant korrelierenden Peaks zu definieren, verwendeten Chau *et al.* nach Berechnung der Selectivity Ratios einen F-Test [160]. Für den vorliegenden Datensatz mit 21 Proben liegt der kritische Wert einer F-Verteilung mit einem Signifikanzniveau von $\alpha = 5$ % bei 2,2. Der Grenzwert ist in Abbildung 3.17 als schwarze horizontale Linie eingezeichnet. Wie gut zu erkennen ist, wird auf diese Weise nur ein Teil der Rifampicin Peaks als signifikant korrelierend erkannt. Insgesamt liegen 13 Peaks über dem Grenzwert. Davon sind 12 dem Rifampicin zuzuordnen. Der 13. Peak ist der oben erwähnte Peak mit *m/z* 287,15277, der nicht eindeutig dem Rifampicin zugewiesen werden kann. Die verbliebenen 12 Rifampicin Peaks sowie die Peaks des Erythromycins werden aufgrund des F-Tests als nicht signifikant klassifiziert und wären somit für die Datenanalyse verloren. Der F-Test scheint daher zu restriktiv zu sein.

Zusammenfassung Die QPAR Methode hat sich als exzellentes Verfahren zur Identifizierung bioaktiver Komponenten in Naturstoffextrakten erwiesen. Alle 47 gesuchten Antibiotikapeaks konnten sowohl mit der Target Projection als auch mit den Selectivity Ratios innerhalb der ersten 100 Einträge in der Ergebnisliste gefunden werden. Der AUC-Wert von 0,99 ist nahezu optimal und vergleichbar dem der AcorA-Methode. Der partielle AUC-Wert von 0,94 deutet eine sehr hohe Sensitivität bei gleichzeitig hoher Spezifität an und übertrifft damit die Aktivitiäts-Korrelations-Analyse auf Basis einer Spearman-Rangkorrelation (pAUC = 0,91). Die Verwendung des F-Tests zur Definition eines kritischen Grenzwertes scheint jedoch zu restriktiv zu sein.

3.2.5. Random Forest

Die Random Forest Methode ist ein Entscheidungsbaum (engl. decision trees) basiertes Verfahren, das in den vergangen Jahren zur Analyse hochdimensionaler Datensätzen an Popularität gewonnen hat [335, 124, 336]. Die Random Forest Analyse hat gegenüber anderen multivariaten Methoden verschiedene Vorteile. Sie ist nicht-parametrisch, sie erfordert keinen linearen Zusammenhang zwischen Regressor und Regressanden, es besteht keine Gefahr der Überanpassung (Overfitting) und es entstehen keine Komplikationen, wenn deutlich mehr Variablen als Proben vorhanden sind (n << p) [337].
Basis der Datenanalyse ist die binäre, rekursive Partitionierung des x-Variablenraumes in der Gestalt, dass Proben mit ähnlichen Regressanden zusammengruppiert werden. Bei jeder Verzweigung wird die Variable selektiert, die den Variablenraum im Hinblick auf die Restvarianz RSS minimiert. Mithilfe von Bootstrapping werden ganze Ensemble von Entscheidungsbäumen gemittelt, sodass die hohe Varianz einzelner Entscheidungsbäume reduziert und die Genauigkeit der Vorhersage verbessert wird. Durch die zufällige Auswahl einer Teilmenge der Variablen bei jeder Verzweigung (Split) des Entscheidungsbaumes wird dieser Effekt noch verstärkt.
Um die Bedeutung einer Variable zu erfassen, werden die Werte jeder x-Variable innerhalb der out-of-bag (OOB) Proben randomisiert permutiert und der MSE berechnet. Bei Variablen mit hoher prädiktiver Bedeutung nimmt der MSE stärker zu als bei Variablen, bei denen sich der MSE durch Permutation nur wenig ändert.
Zur Optimierung der Random Forest Methode müssen die Parameter *ntree* (Anzahl der Bootstrap Bäume) und *mtry* (Anzahl der Variablen pro Split) angepasst werden. Die Random Forest Methode ist jedoch relativ robust, sodass in der Literatur häufig die von Breiman vorgeschlagenen Standardparameter *ntree* = 500 und *mtry* = Anzahl der Variablen/3 verwendet werden [338, 126].
In dieser Arbeit wurde zunächst der Parameter *ntree* optimiert. Wie [3, 319] zeigen, kann der out-of-bag (OOB) Fehler als gute Näherung für den Testfehler MSEP verwendet werden. Abbildung 3.18 a zeigt die Abhängigkeit der OOB-Fehlerrate von der Anzahl der verwendeten Bootstrap Bäume unter Verwendung der Standardeinstellung für *mtry*. Es ist deutlich erkennbar, dass sich die OOB-Fehlerrate ab ca. 500 Bootstrap Bäumen kaum noch ändert und bei ntree = 2000 ein Minimum besitzt. Die weiteren Analysen wurden daher mit 2000 Bootstrap Bäumen durchgeführt.

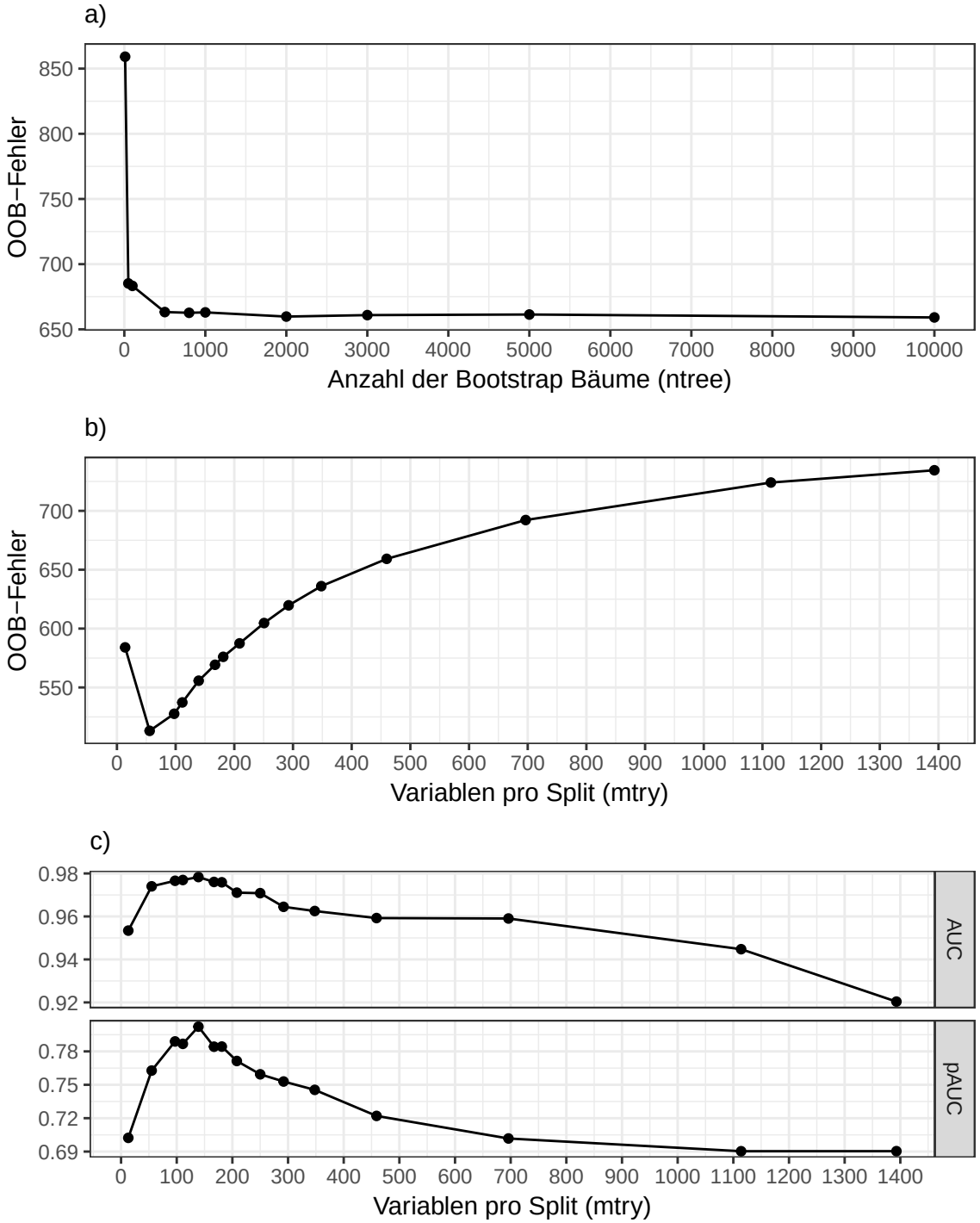


Abbildung 3.18.: a) OOB-Fehler in Abhängigkeit der Anzahl der Bootstrap Bäume (*ntree*).
b) OOB-Fehler in Abhängigkeit der Anzahl der Variablen pro Split (*mtry*).
c) AUC und pAUC bei 95 % Spezifität in Abhängigkeit der Anzahl der Variablen pro Split (*mtry*).

In einem zweiten Schritt wurde versucht über die OOB-Fehlerrate eine Schätzung für die optimale Anzahl der Variablen pro Split (*mtry*) zu erhalten. Abbildung 3.18 b zeigt, dass der OOB-Fehler bei *mtry* = 55 ein Minimum besitzt. Bei kleinerer Splitvariablenanzahl ist der OOB-Fehler größer. Werden mehr als 55 Variablen pro Split verwendet, steigt der OOB-Fehler kontinuierlich an, um bei Verwendung aller Variablen - dies entspricht dem Bagging - ein Maximum zu erreichen.
Anschließend wurden nun mit allen in Abbildung 3.18 b verwendeten *mtry*-Werten Random Forest Modelle erstellt und deren Qualität im Hinblick auf die Identifizierung der Antibiotikapeaks über ROC-Kurven abgeschätzt. Abbildung 3.18 c zeigt die AUC- und pAUC-Werte für die verschiedenen Modelle in Abhängigkeit des Splitparameters *mtry*. AUC und pAUC steigen unter Verwendung kleiner Werte für *mtry* zunächst an, um bei mtry = 139 ein Maximum von 0,98 (pAUC = 0,80) zu erreichen. Wurden mehr als 140 Variablen für jeden Split verwendet, fallen AUC und pAUC kontinuierlich ab. Der Optimalwert für *mtry*, der aus dem OOB-Fehler berechnet wurde (*mtry*= 55), ist nicht identisch mit dem tatsächlichen Optimum, das über den AUC/pAUC Wert bestimmt wurde (*mtry* = 139). Die Unterschiede sind jedoch weder in Bezug auf AUC (p = 0,50) noch auf pAUC (p = 0,44) signifikant. Auffällig ist, dass die Kurvenverläufe für alle verwendeten *mtry* Werte bis Platz 50 in der Ergebnisliste nahezu identisch sind. Bis hierhin wurden im Durchschnitt 20 der 24 Rifampicin Peaks gefunden. Erst auf den Plätzen 51 bis 100 zeigen sich Unterschiede bei der Identifizierung der Antibiotikapeaks. Bis zu Platz 100 wurden mit *mtry*= 55 (Optimum über OOB-Fehler) und *mtry* = 139 (Optimum über AUC/pAUC) 40 der 47 Antibiotikapeaks gefunden. Mit der *mtry* Standardeinstellung (Anzahl der x-Variablen/3 $\approx$ 459) wurden 36 Antibiotikapeaks gefunden. In allen drei Fällen sind die Unterschiede nicht signifikant. Lediglich bei sehr kleinen (*mtry* = 13) und sehr großen Werten (*mtry* = 1393, Bagging) für *mtry* wurden signifikante Unterschiede zum Optimalwert beobachtet. Vergleichbar zu AcorA und den Selectivity Ratios (QPAR) erhalten Massensignale, die von einer Substanz ausgehen, ähnliche VI Werte. Abbildung 3.21 zeigt, dass die Signale der Rifampicin Peaks bei allen verwendeten *mtry* Parametern relativ deutlich von den anderen Massensignalen abgesetzt sind. Weiterhin ist zu erkennen, dass die VI insbesondere für Rifampicin bis zum optimalen *mtry* Wert von 139 zunächst ansteigt und anschließend wieder abfällt, während sich die VI Werte für die Erythromycin Signale nur marginal verändern und in vielen Fällen in der Nähe des Rauschens liegen. Die höhere Präferenz für die Rifampicin Peaks ist möglicherweise darauf zurückzuführen, dass Rifampicin in deutlich mehr Proben in hoher Konzentration (5 Extrakte) vorkommt als Erythromycin (2 Extrakte). Einer der beiden Extrakte mit hoher Erythromycinkonzentration ist äquimolar mit Rifampicin gespickt, sodass

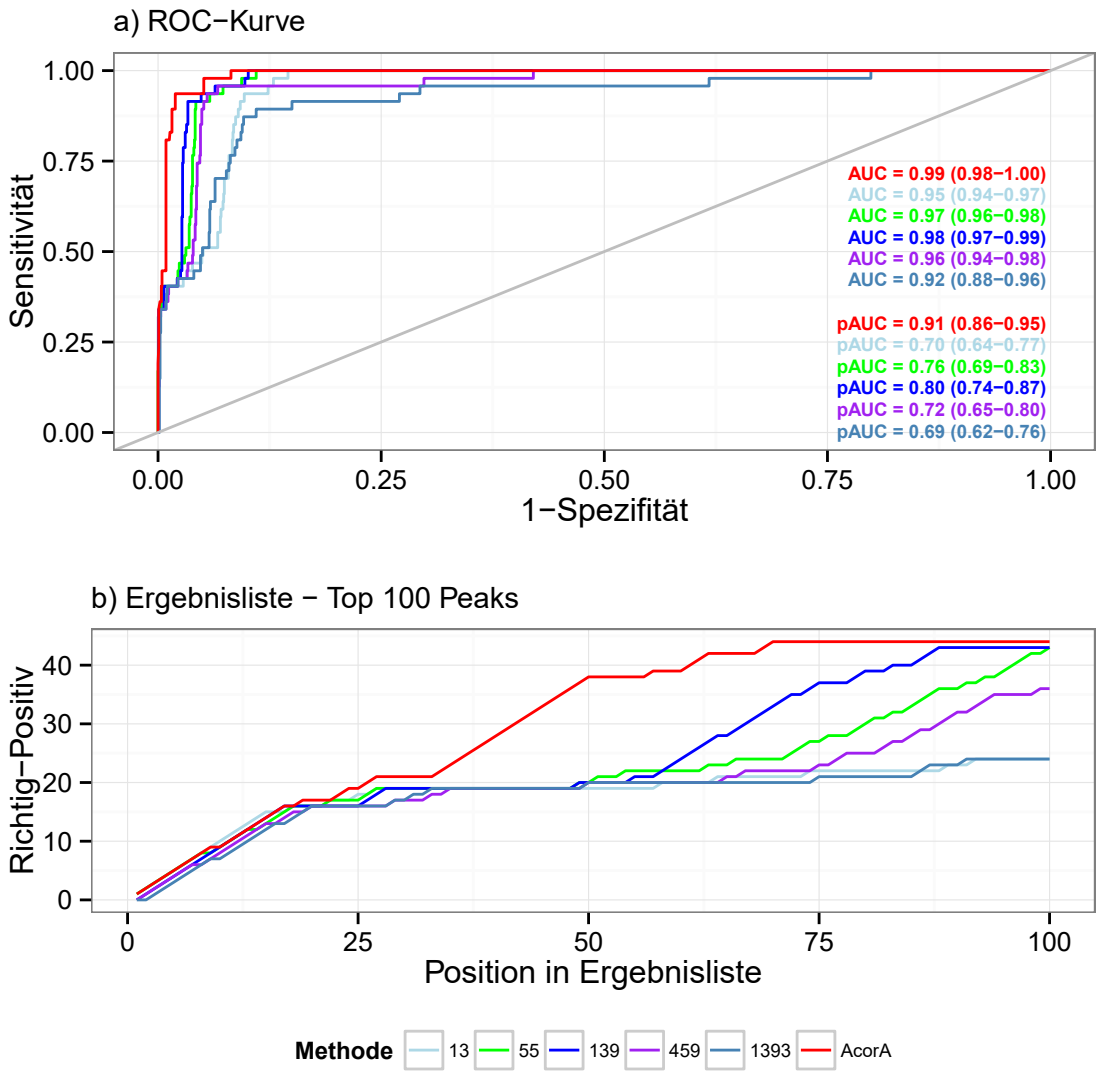


Abbildung 3.19.: ROC-Plot der Random Forest Analyse für verschiedene Werte von *mtry*. Als Vergleich ist das Ergebnis von AcorA (Rot) dargestellt. Die 95% Konfidenzintervalle der AUC und pAUC-Werte sind in Klammern angegeben. Die pAUC-Werte beziehen sich auf den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate, der durch die gestrichelten Linien gekennzeichnet ist. b) Anzahl der annotierten Antibiotikapeaks in Abhängigkeit der Position in der Ergebnisliste.

hier zusätzlich eine Kokorrelation zur Bioaktivität auftritt. Insgesamt werden dadurch Rifampicinhaltige Extrakte bei der Wahl der Bootstrap Replikate häufiger selektiert, als die

Erythromycinhaltige Extrakte.

Variablenselektion Das R-Paket *VSURF* erlaubt eine mehrstufige Variablenselektion. Nach Erstellung von 50 Random Forest Modellen wurden VI und deren Standardabweichungen für alle **X**-Variablen berechnet und nach Größe der VI sortiert (Abbildung 3.20 a). Anschließend wurde ein Grenzwert für die Standardabweichungen der VI berechnet (Abbildung 3.20 b). Für den oben optimierten Fall mit *ntree* = 2000 und *mtry* = 139 wurden durch dieses Verfahren 157 Variablen selektiert. Innerhalb dieses reduzierten Variablensatzes befinden sich 46 der 47 gesuchten Antibiotikapeaks. Anschließend folgte die Eliminierung redundanter Variablen. Dazu wurden anhand der zuvor selektierten und nach VI sortierten Variablen 157 Random Forest Modelle aufgestellt, in dem - angefangen bei der Variable mit dem höchsten VI-Wert - sukzessiv die ersten 1 bis 157 Variablen eingesetzt und die mittleren OOB-Fehler berechnet werden. Bei dem vorliegenden Datensatz bestand das sparsamste Modell mit dem niedrigsten OOB-Fehler aus insgesamt 8 Variablen (Abbildung 3.20 c). 7 der 8 Variablen wurden als Rifampicinpeaks identifiziert. Die Variable *m/z* 483,2876 konnte keinem der eingesetzten Antibiotika zugeordnet werden. Ob es falsch oder richtig positiv ist, wurde in dieser Analyse nicht untersucht.

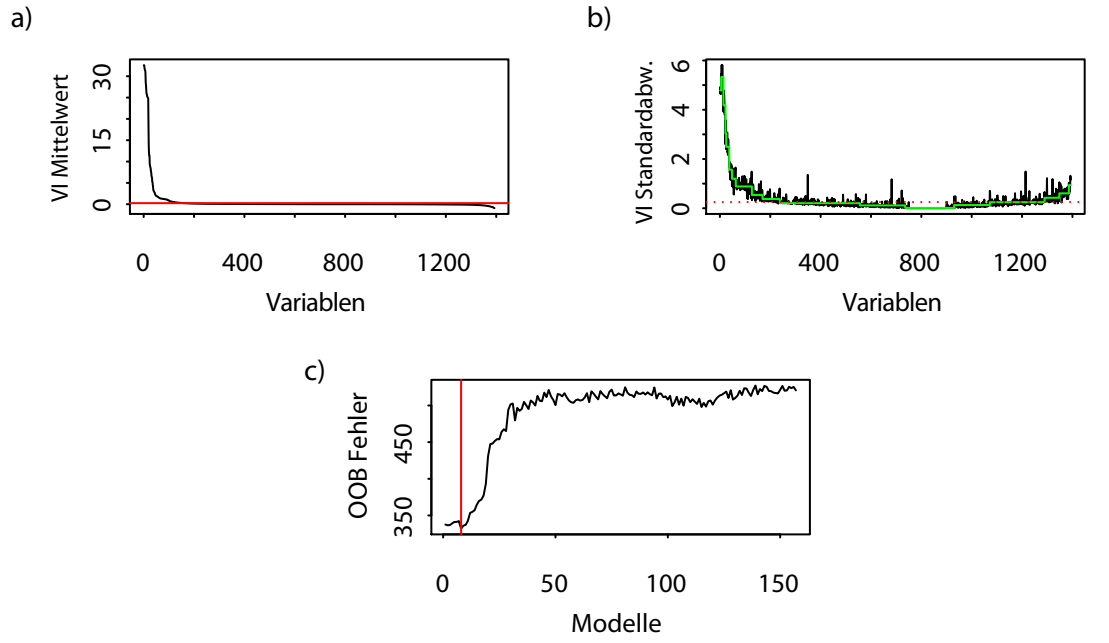


Abbildung 3.20.: Variablenselektion mit VSURF.

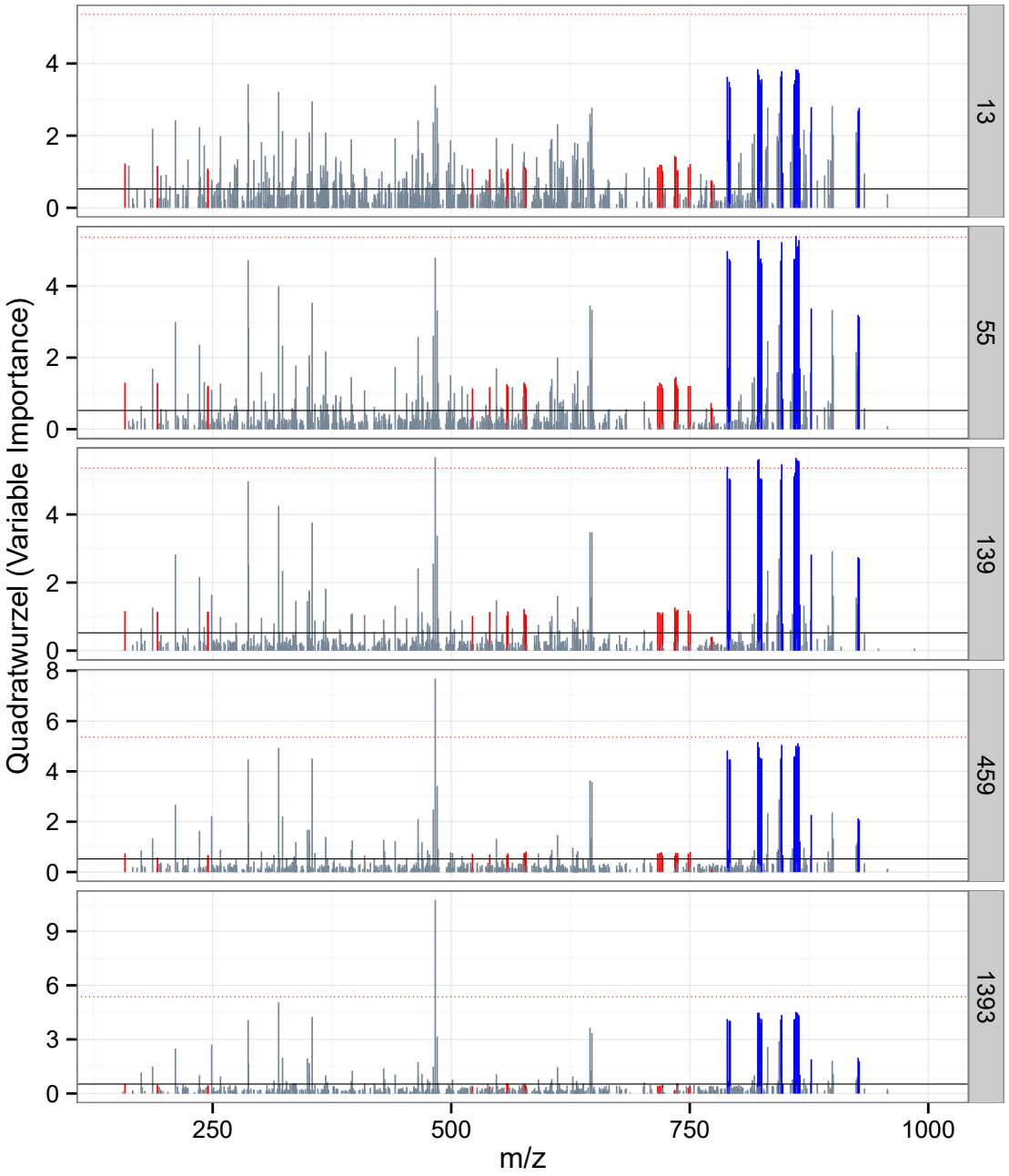


Abbildung 3.21.: Variable Importance der Massensignale im Bereich zwischen *m/z* 150 und 1000 für verschiedene Werte von *mtry* (Anzahl der Variablen pro Split) bei 2000 Bootstrap Replikaten. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und der nicht aktiven Verbindungen (Grau) sind farblich gekennzeichnet. Die schwarze horizontale Linie gibt den Grenzwert im 1. Variablenselektionsschritt an. Die rote horizontale Linie entspricht dem Grenzwert des 2. Selektionsschritts.

Zusammenfassung Die Random Forest Analyse ist in Hinblick auf ihre Fähigkeit zur Identifizierung der Antibiotikapeaks unter Verwendung der optimalen Parametern vergleichbar mit der Hauptkomponentenregression und der Partial-Least-Squares Regression. Befindet sich die bioaktive Komponente nur in sehr wenigen Proben, kann jedoch die Identifizierung, ähnlich wie bei AcorA, erschwert sein. Die Optimalwerte für die Anzahl der Bootstrap Bäume (*ntree*) und die Anzahl der Variablen pro Split (*mtry*) können über den OOB-Fehler recht gut approximiert werden. Allerdings wurden auch mit den von Breimann eingeführten Standardwerten sehr gute Resultate erzielt, sodass mit der Random Forest Analyse eine Methode zur Verfügung steht, die im Vergleich zu anderen multivariaten Methoden sehr robust ist und wenig Aufwand erfordert. Das Paket *VSURF* erlaubt eine mehrstufige Variablenselektion. Für den bearbeiteten Datensatz wurden nach dem ersten Selektionsschritt 46 der 47 gesuchten Antibiotikapeaks gefunden. Der zweite Selektionsschritt entfernt redundante Variablen. Dadurch wurden jedoch lediglich 7 der Rifampicin Peaks als Prädiktoren für die Wachstumsinhibition identifiziert. Für die Analyse von Massenspektren sind die Daten aus dem ersten Selektionschritt von Vorteil, da hier noch die Informationen über Isotopen- und Adduktpeaks zur Verfügung stehen und zur Identifizierung und Verifizierung einer potenziell biologisch aktiven Substanz genutzt werden können.

3.2.6. Regularisierungsmethoden

Ein inhärentes Problem großer Datenmatrizen mit n << p sind Multikollinearitäten innherhalb der Regressoren. In diesem Fall sind die Prädiktoren nicht mehr linear unabhängig, sodass die Varianz der Regressionskoeffizienten ansteigt. Die Selektion der Variablen aufgrund der Größe der Koeffizienten ist dann nur bedingt geeignet, einen Zusammenhang zwischen einer Variable x und der Bioaktivität y herzuleiten [306, 333, 3].
Ein Weg dieses Problem zu mildern sind s. g. Regularisierungsmethoden wie z. B. Ridge Regression, Lasso oder Elastic Net. Durch Verwendung eines Penalisierungsterms werden die Regressionskoeffizienten in Abhängigkeit eines Schwellenparameters λ gruppenweise in Richtung Null geschrumpft (Ridge Regression) oder nach unterschreiten eines gewissen Schwellenwertes direkt auf Null gesetzt (Lasso, Elastic Net). Dabei werden Variablen mit niedriger Varianz stärker penalisiert, als Variablen mit höherer Varianz [3]. Gleichzeitig erhalten untereinander stark korrelierte Variablen Regressionskoeffizienten einen ähnlichen hohen Betrag. Das s. g. Elastic Net ist eine Mischform aus Ridge Regression und Lasso. Es mischt die Eigenschaften der Ridge Regression (gruppenweises Schrumpfen der Regressionskoeffizienten) mit der Variablenselektionsfähigkeit der Lasso Analyse. Alle drei Methoden haben gemeinsam, dass sie die Varianz zum Preis niedrigerer Erwartungstreue (höherer Bias) verringern [3].
Der Penalisierungsfaktor λ bestimmt maßgeblich wie stark die Regressionskoeffizienten in Richtung Null geschrumpft werden. Aus diesem Grunde wurde λ für alle drei Methoden über eine 100-fache doppelte Kreuzvalidierung optimiert. Der in der Kreuzvaldierung am häufigsten erhaltene λ_{opt} -Wert sollte theoretisch den geeignetsten Penalisierungswert darstellen. Die Abbildungen D.3 und D.4 zeigen jedoch, dass insbesondere bei kleinen Werten für α sehr viele λ_{opt}-Werte nahezu singulär auftraten und sich keine klaren Optima finden lassen. In der Ridge Regression (α = 0) besaß die Häufigkeitsverteilung der in der Kreuzvalidierung optimierten λ-Werte zwei deutlich getrennte Extrembereiche (D.3). Tatsächlich wurden anhand der Kreuzvalidierung zwei Maxima bei λ = 1710 und 27535 mit einer Häufigkeit von jeweils fünf erhalten. Der Algorithmus selektiert an dieser Stelle den ersten λ-Wert (RMSEP = 13,0), da dieser mit einem deutlich niedrigeren RMSEP assoziiert ist als der zweite λ-Wert (RMSEP = 47,8). Interessanterweise wurden mit der Einstellung „lower.limits = 0“ (d. h. Regressionskoeffizienten mit Werten < 0 werden automatisch verworfen), ebenfalls zwei deutlich getrennte Extrembereiche mit einem Submaximum bei 733 (RMSEP = 16,1) und einem Maximum bei 27535 (RMSEP = 47,8) erhalten. Hier selektierte der Algorithmus automatisch den zweiten Wert. Das Phänomen, dass keine Gauß-artige Häufigkeitsvertei-

lung der λ-Werte, sondern z. T. bimodale Verteilungen erhalten wurden, ist bei den anderen getesteten α-Werten ebenfalls zu beobachten, allerdings deutlich weniger ausgeprägt als bei der Ridge Regression. Das Problem der variablen Schätzungen von λ ohne klares Optimum wurde auch schon von Houwelingen *et al.* [339] beschrieben. Ursache ist vermutlich die sehr schlecht konditionierte Datenmatrix (n = 21 Proben, p = 1393 Variablen), bei der nur wenig Proben für die doppelte Kreuzvaldierung zur Verfügung stehen.

Ridge Regression Für die Ridge Regression mit den Standardeinstellungen wurde anhand der Kreuzvalidierung ein optimierter λ-Wert von 1710 erhalten. Der RMSEP von 13,0 ist im Vergleich zu den anderen untersuchten Methoden sehr niedrig. Wie erwartet wurden mit der Ridge Regression die Regressionskoeffizienten geschrumpft, jedoch verbleiben alle 1393 Peaks im Regressionsmodell (Abbildung D.2 a, Anhang). Die Ridge Regression ist im Hinblick auf ihre Fähigkeit zur Identifizierung bioaktiver Substanzen mit AcorA und QPAR vergleichbar. Alle drei Methoden besitzen einen nahezu optimalen AUC-Wert von 0,99 (Abbildung 3.22). Im Bereich hoher Spezifität (>= 95 %) erzielt die Ridge Regression mit einem pAUC von 0,89 etwas schlechtere Ergebnisse als AcorA (pAUC = 0,91) und QPAR (pAUC = 0,94). Laut der Theorie erhalten Signale, die untereinander hoch korreliert sind, vergleichbare Regressionskoeffizienten [3]. Dieser Effekt kann sehr gut in Abbildung 3.23 beobachtet werden. Insbesondere die hochkorrelierten Erythromycinpeaks (vgl. mit Abbildung A.1, Anhang) haben nahezu alle die gleichen Regressionskoeffizienten.
Das verwendete R-Paket *glmnet* besitzt eine Option (*lower.limits*), mit der ein unteres Limit für die Regressionkoeffizienten gesetzt werden kann. Da für die Identifizierung der bioaktiven Signale in dem vorliegenden Fall ausschließlich positive Korrelationen von Interesse sind, wurde die Ridge Regression mit *lower.limits* = 0 durchgeführt. Dabei werden in einem ersten Durchlauf zunächst alle Regressionskoeffizienten berechnet. In der darauffolgenden Analyse werden alle Variablen mit Regressionskoeffzienten < 0 auf 0 gesetzt und das Modell auf dem reduzierten Datensatz neu berechnet.

Wie Abbildung D.2 b (Anhang) zeigt, wird das Modell dadurch auf 310 Variablen reduziert. Die resultierende ROC-Kurve ist vergleichbar der der Ridge Regression über alle 1393 Variablen. Der AUC-Wert ist mit 0,99 identisch, der pAUC-Wert ist mit 0,95 sogar noch etwas verbessert und der höchste pAUC-Wert innerhalb der in dieser Arbeit verwendeten Analysemethoden. Interessanterweise ist der RMSEP von 47,8 gleichzeitig der schlechteste der in dieser Studie verwendeten Methoden. Ein schlechter Vorhersagewert des Modells ist also nicht unbedingt gleichbedeutend mit einer schlechten Qualität der Variablenselektion. Ein Grund für die guten Selektionseigenschaften ist in Abbildung 3.23 erkennbar. Im Ver-

gleich zur Standardausführung der Ridge Regression wurden insbesondere Peaks der inaktiven Verbindungen stärker geschrumpft. Auf diese Weise setzen sich die Peaks der Antibiotika noch besser von denen der inaktiven Verbindungen ab. Bis Platz 50 der Ergebnisliste werden bereits 39 der 47 gesuchten Antibiotikapeaks annotiert. Innerhalb der ersten 65 Peaks in der Ergebnisliste sind alle 47 Antibiotikapeaks zu finden. Der Vollständigkeit halber wurde auch das Least-Squares Modell, d. h. eine einfache multivariate Regression, berechnet. Dazu wurde der Schwellenparameter λ auf 0 gesetzt. Das Ergebnis ist wie erwartet deutlich schlechter als bei der Ridge Regression. Unter den ersten 50 Peaks der Ergebnisliste sind 18 Antibiotikapeaks zu finden und auch bis Platz 100 werden lediglich 34 der 47 Antbiotikapeaks annotiert. Weiterhin zeigen AUC (0,93) und insbesondere pAUC (0,67), dass dieses Verfahren vergleichsweise unbrauchbar für die Identifizierung der aktiven Komponenten ist.

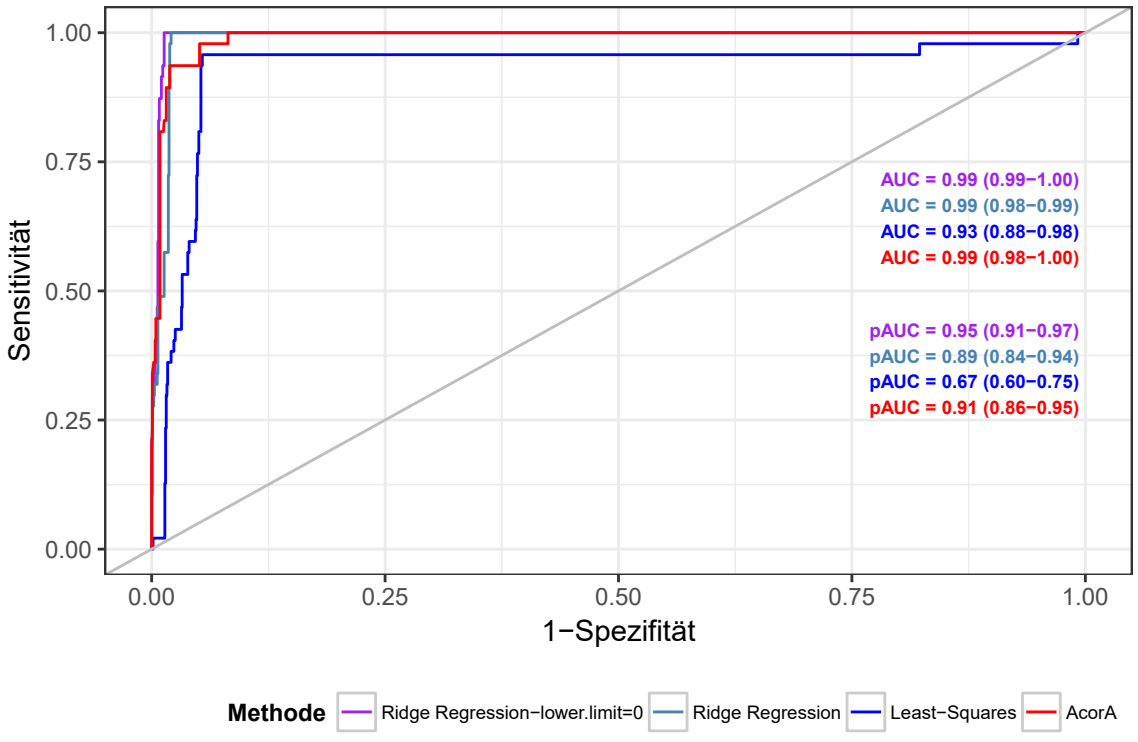


Abbildung 3.22.: ROC-Kurven für Ridge Regression, Ridge Regression mit lower.limit = 0 und Least-Squares. Als Vergleich ist das Ergebnis von AcorA (Rot) dargestellt. Die 95% Konfidenzintervalle der AUC- und pAUC-Werte sind in Klammern angegeben. Die pAUC-Werte beziehen sich auf den Bereich zwischen 0 und 5 % Falsch-Positiv-Rate, der durch die gestrichelten Linien gekennzeichnet ist.

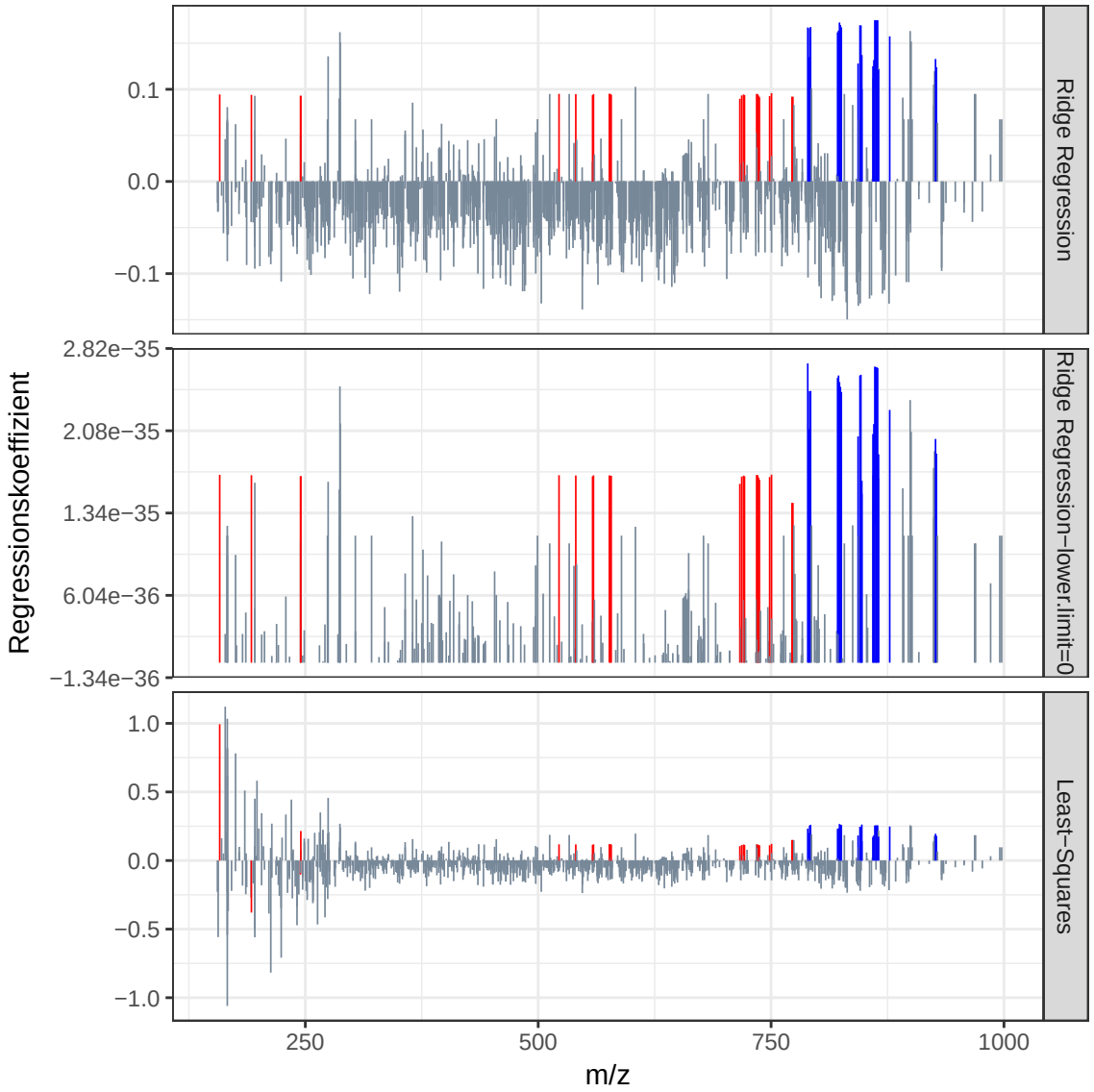


Abbildung 3.23.: Regressionskoeffizienten für die *m/z* Signale im Massenbereich zwischen *m/z* 150 und 1000 ermittelt über Ridge Regression (oben), Ridge Regression mit lower.limit = 0 (mitte) und Least-Squares (unten). Peaks aus Erythromycin (Rot), Rifampicin (Blau) und der nicht aktiven Verbindungen (Grau) sind farblich gekennzeichnet.

Variablenselektion mit Lasso und Elastic Net Ähnlich der Ridge Regression werden bei der Lasso Methode die Regressionskoeffizienten in Richtung Null geschrumpft. Im Gegensatz zur Ridge Regression verwendet Lasso jedoch die s. g. ℓ_1 Norm als Penalisierungsterm.

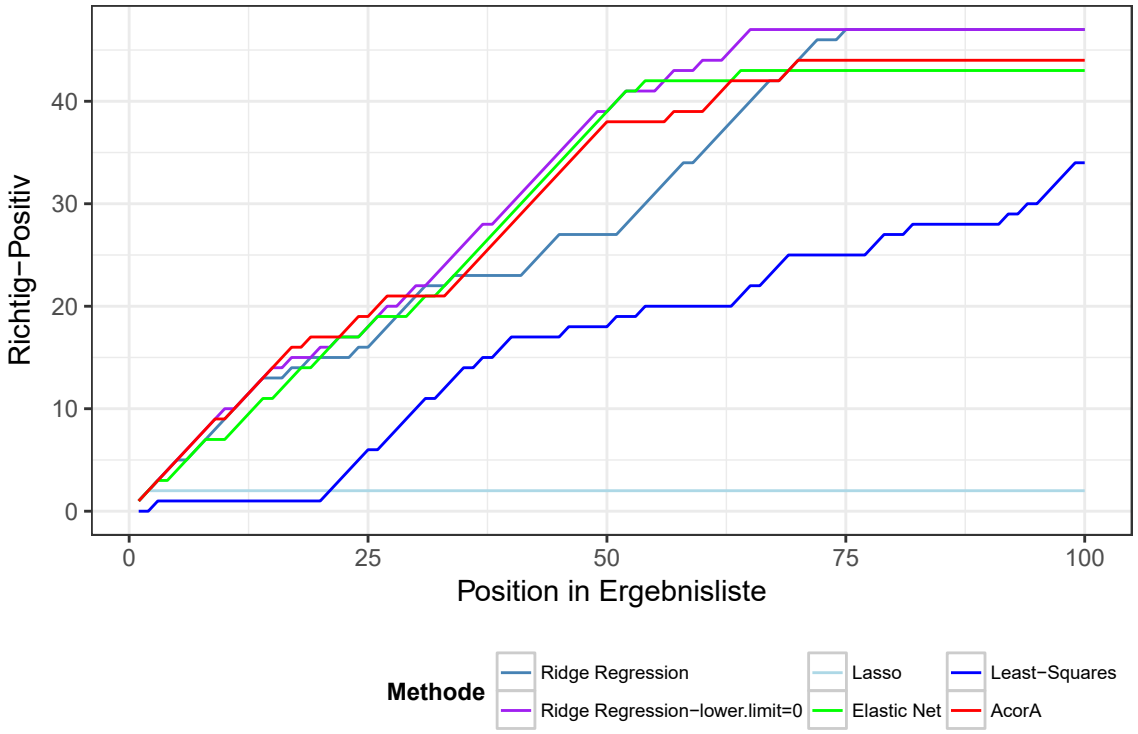


Abbildung 3.24.: Anzahl der Antibiotikapeaks unter den Top 100 Peaks für Ridge Regression, Ridge Regression mit lower.limit = 0, Least-Squares, Lasso, Elastic Net und AcorA.

Dadurch werden Regressionskoeffizienten, die in Abhängigkeit des gewählten Regularisierungsparameters λ unter einen gewissen Schwellenwert fallen, direkt auf Null gesetzt. Weiterhin tendiert die Lasso Analyse dazu aus einer Gruppe von korrelierten Variablen lediglich eine im Modell zu belassen und die anderen zu verwerfen. Insgesamt führt dies zu sehr sparsamen Modellen.

Der RMSEP ist mit 28,42 vergleichbar mit der PCR und PLSR. Abbildung D.2 c zeigt, dass mit λ_{opt} = 19 im finalen Modell lediglich 2 Variablen auftreten, die beide dem Rifampicin zuzuordnen sind ([Rif - H_2 - MeOH]$^+$, [Rif + K]$^+$) (Abbildung 3.25). Peaks aus dem Erythromycin wurden nicht gefunden. Interessanterweise enthält die Ergebnisliste keine redundanten Peaks im Sinne von zusammenhängenden Isotopenpeaks. Lasso selektiert hier gemäß der Theorie jeweils nur einen Peak aus einer Gruppe von hochkorrelierten Peaks. Das Lasso-Modell scheint somit zu restriktiv und es fehlt die Redundanz aus Isotopen- und Adduktpeaks, die eine eindeutige Identifizierung erleichtert.

Eine Mischform aus Ridge Regression und Lasso ist das s. g. Elastic Net. In dem hier

verwendeten R-Paket *glmnnet* kann der Anteil der ℓ_1 Norm (Lasso) und ℓ_2 Norm (Ridge Regression) über einen Faktor α reguliert werden, wobei $\alpha = 0$ der Ridge Regression und $\alpha = 1$ der Lasso Analyse entspricht. Über eine doppelte Kreuzvalidierung wurden anhand des Vorhersagefehlers die optimalen Werte für α und λ bestimmt. Der niedrigste MSEP wurde für $\alpha = 0{,}1$ mit $\lambda_{opt} = 58$ erhalten. Der RMSEP ist mit 18,9 der zweitbeste in dieser Vergleichsstudie. Wie Abbildung D.2 d zeigt, ist das Elastic Net deutlich weniger restriktiv als die Lasso Analyse, sodass insgesamt 64 Variablen im Modell verbleiben. Unter den 64 Variablen befinden sich 22 Peaks, die dem Erythromycin zuzuordnen sind sowie 19 Peaks aus dem Rifampicin. Mit insgesamt 43 richtig positiven Peaks besteht die Hitliste zu 67 % aus den gesuchten Antibiotikapeaks und besitzt die gleiche Genauigkeit wie AcorA. Der starke Anreicherungseffekt wird auch sehr gut in Abbildung 3.25 deutlich. Ähnlich der Ridge Regression erhalten untereinander korrelierte Peaks vergleichbare Regressionskoeffizienten. Aufgrund des erhöhten Anteils der ℓ_2 Norm werden jedoch die meisten der nicht mit der Bioaktivtät assozierten Peaks ausgefiltert.

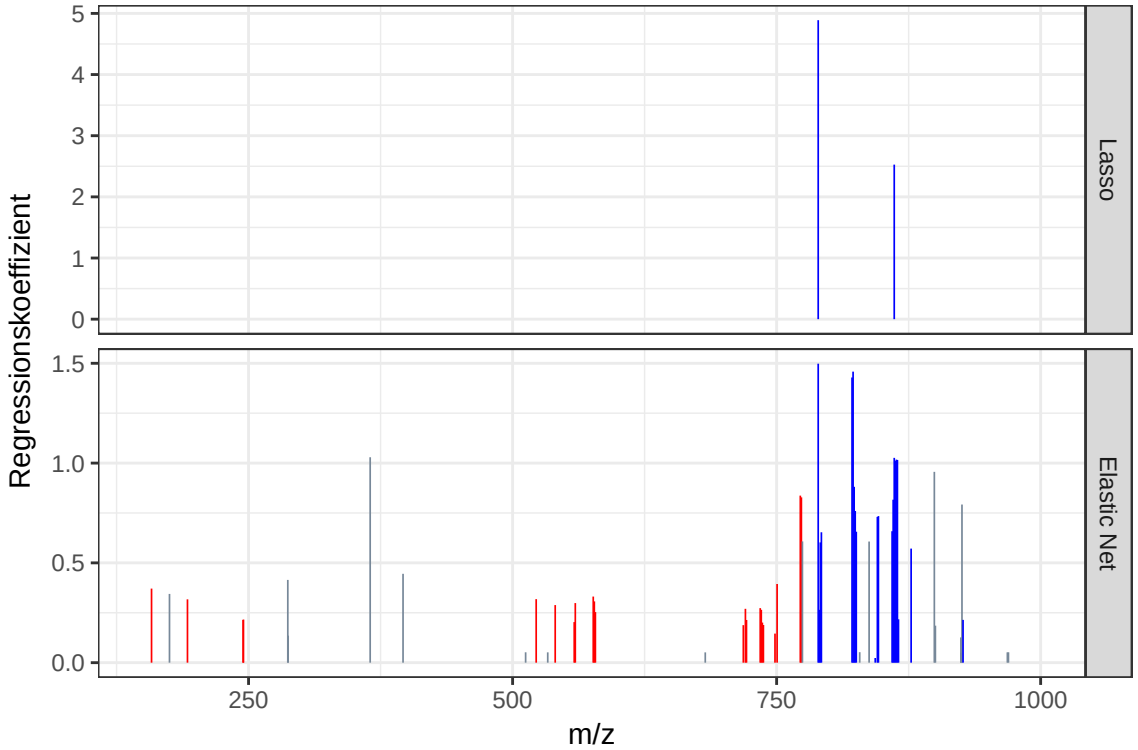


Abbildung 3.25.: Regressionskoeffizienten der Peaks im Bereich zwischen *m/z* 150 und 1000 für Lasso (oben) und Elastic Net (unten).

Zusammenfassung Alle drei Regularisierungsmethoden haben im Hinblick auf ihre Eigenschaft zur Identifizierung bioaktiver Komponenten exzellente Ergebnisse geliefert. Von den in dieser Arbeit getesteten Methoden hat die Ridge Regression mit Beschränkung auf die positiven Regressionskoeffizienten die höchsten AUC- und pAUC-Werte erzielt. Durch diese Beschränkung wird der Datensatz zwar reduziert, eine intrinsische Variablenselektion wie bei Lasso und Elastic Net findet jedoch nicht statt. Um einen kritischen Wert für die signifikant korrelierenden Peaks zu erhalten, könnte - wie bei allen Filtermethoden - zusätzlich ein Permutationstest oder eine Bootstrap Analyse durchgeführt werden.
Demgegenüber verfügen Lasso und Elastic Net über eine intrinsische Variablenselektion. Mithilfe der Elastic Net Analyse wurden insgesamt 64 Peaks selektiert, von denen 43 den 47 gesuchten Antibiotikapeaks zugeordnet werden konnten. Mit einer Trefferquote von 91 % wurden mit der Elastic Net Analyse hervorragende Ergebnisse erzielt. Durch die Selektion untereinander korrelierter Variablen wurden nahezu alle Erythromycin und Rifampicin Signale inklusive der Isotopen- und Adduktpeaks erhalten. Die Lasso Analyse hat dem hingegen ein sehr sparsames Modell mit lediglich 2 Rifampicin Peaks ergeben. Somit hat Lasso zwar den Vorteil einer maximalen Genauigkeit von 100 % bei gleichzeitig vereinfachter Interpretierbarkeit der Ergebnisse. Erythromycin Signale wurden jedoch übersehen und es fehlen die Informationen der Isotopen- und Adduktpeaks, die die Differenzierung zwischen Scheinkorrelation und kausaler Korrelation erleichtern.
Wie diese Studie gezeigt hat, kann bei schlecht konditionierten Datensätzen die Bestimmung des optimalen Penalisierungsparameters λ aufgrund der hohen Variabilität in der Kreuzvalidierung problematisch sein. Um einen Datensatz mit wenig Proben nicht weiter zu verkleineren und somit etwas stabilere Ergebnisse zu erhalten, könnte Bootstrapping eine geeignete Alternative darstellen.

3.2.7. Vergleich der Analysemethoden

Ein Vergleich der Ergebnisse der Datenanalysemethoden ist in Tabelle 3.2 dargestellt. Die in Hinblick auf Selektion der Antibiotikapeaks besten Ergebnisse wurden mit der Ridge Regression und QPAR erzielt. In beiden Fällen wurden nahezu optimale AUC- und pAUC-Werte erhalten. Unter den Top 50 Peaks befinden sich jeweils 39 respektive 40 Antibiotikapeaks, im Bereich bis 100 Peaks werden alle 47 gesuchten Antibiotikapeaks annotiert. Ähnlich gute Ergebnisse wurden mit der AcorA-Methode erhalten. Der pAUC-Wert ist mit 0,91 nur etwas geringer als bei Ridge Regression (0,95) und QPAR (0,94). Allerdings konnten innerhalb der Top 100 Peaks nur 44 Antibiotikapeaks detektiert werden.
Mit 45 und 43 gefundenen Antibiotikapeaks unter den Top 100 Peaks, haben die PLSR Beta und die Random Forest Methode ebenfalls relativ gute Resultate erzielt. Allerdings sind die pAUC-Werte im Bereich niedriger Falsch-Positiv-Raten zwischen 0 und 5 % signifikant schlechter, als bei AcorA, QPAR und Ridge Regression. Die Trefferquote unter den Top 50 Peaks ist mit 64 % (PLSR Beta) und 43 % (Random Forest) daher deutlich geringer als bei den drei zuvor genannten Methoden. Eine höhere Falsch-Positiv-Rate erschwert jedoch die Identifizierung der Peaks mit kausalen Korrelationen zur Bioaktivität.
Interessant sind die im Vergleich zur PLSR Beta Methode schlechten Resultate der PLSR-VIP. Durch die Amalgamierung der positiv und negativ korrelierenden Peaks in der Ergebnisliste wird der Anteil der falsch positiven Peaks artifiziell erhöht. Auf diese Weise sinken AUC, pAUC und Trefferquote im Vergleich zur PLSR deutlich ab.
Aufgrund der ungerichteten Natur erzielten die beiden auf Berechnung der Hauptkomponenten beruhenden Methoden PCA und PCR erwartungsgemäß die schlechtesten Ergebnisse. Unter den Top 100 Peaks wurden lediglich 41 (PCR) und 34 Peaks (PCA) annotiert. Die AUC-Werte sind mit 0,94 und 0,90 daher deutlich schlechter, als bei den anderen getesteten Methoden.
Den oben aufgeführten Datenanalysemethoden ist gemeinsam, dass sie über keine intrinsische Variablenselektion verfügen. Um die eigenschaftsrelevanten Variablen selektieren zu können, müssen in einem zweiten Schritt entsprechende Filtermethoden angewendet werden. Dies kann beispielweise ein Permutations- oder ein F-Test sein. In Tabelle 3.3 sind die Ergebnisse für die getesteten Methoden nach einem Variablenselektionschritt aufgeführt.
Zur Abschätzung der Güte der Selektionsmethoden wurden Genauigkeit (Anteil der Antibiotikapeaks in der Hitliste) und Trefferquote (Anzahl der gefundenen Antibiotikapeaks im Verhältnis zu den gesuchten Antibiotikapeaks) berechnet. Der F_1-Wert entspricht dem harmonischen Mittel aus den beiden Gütemaßen.

Tabelle 3.2.: Zusammenfassung der Analysemethoden. RMSEP, AUC, pAUC und Anzahl der Antibiotikapeaks unter den Top 50 bzw. Top 100 Peaks.

Methode	RMSEP	AUC (CI)	pAUC (CI)	Top 50 Peaks (Trefferquote)	Top 100 Peaks (Trefferquote)
AcorA	-	0,99 (0,98-1,00)	0,91 (0,86-0,95)	38 (81 %)	44 (94 %)
PCA	-	0,90 (0,84-0,96)	0,82 (0,76-0,88)	28 (60 %)	34 (72 %)
PCR	29,0	0,94 (0,89-1,00)	0,81 (0,74-0,87)	26 (55 %)	41 (87 %)
PLSR	28,4	0,99 (0,98-0,99)	0,88 (0,82-0,93)	30 (64 %)	45 (96 %)
PLSR VIP	24,9	0,96 (0,95-0,98)	0,78 (0,71-0,85)	25 (53 %)	36 (77 %)
QPAR (TP)	21,6	0,99 (0,99-1,00)	0,94 (0,91-0,97)	40 (85 %)	47 (100 %)
Random Forest	23,6	0,98 (0,97-0,99)	0,80 (0,74-0,87)	20 (43 %)	43 (91 %)
Ridge Regression	47,6	0,99 (0,99-1,00)	0,95 (0,91-0,97)	39 (83 %)	47 (100 %)
Elastic Net	18,9	-	-	39 (83 %)	43 (91 %)
Lasso	28,4	-	-	2 (4 %)	2 (4 %)

Mit einem F_1-Wert von 77 % hat die Elastic Net Methode die besten Ergebnisse erzielt. Die Methode lieferte ein sehr ausgewogenes Verhältnis aus Genauigkeit (67 %) und Trefferquote (91 %). Eine identische Genauigkeit aber geringfügig niedrigere Trefferquote (89 %) lieferte die bei AcorA verwendete Spearman Rangkorrelation. Der F_1-Wert ist mit 76 % nur minimal niedriger als bei Elastic Net.

Sehr ambivalente Ergebnisse wurden für Lasso, QPAR und den 2. Selektionsschritt in der Random Forest Analyse erhalten. In allen drei Fällen lag die Genauigkeit bei über 80 %. Aufgrund der sehr sparsamen Modelle wurden jedoch nur wenige Antibiotikapeaks in den Hitlisten annotiert. Wie der hohe pAUC-Wert zeigt, hätte insbesondere QPAR mit einer weniger restriktiven Selektionsmethode ein viel höheres Potential zur Identifizierung der eigenschaftsrelevanten Signale.

Ergebnisse von mittlerer Güte wurden mit dem 1. Selektionsschritt der Random Forest Analyse (F_1 = 45 %) und der PLSR-VIP Methode (F_1 = 46 %) erhalten. Aufgrund der hohen Anzahl der selektierten Variablen ist die Trefferquote bei der Random Forest Analyse zwar sehr hoch (98 %), allerdings enthält die Hitliste auch viele falsch positive Signale, sodass die Genauigkeit sehr niedrig ausfiel (29 %). Die PLSR-VIP Methode hingegen erzielte

Tabelle 3.3.: Zusammenfassung der Analysemethoden nach Variablenselektion.

Methode	Selektions-verfahren	Peaks nach Selektion	Antibiotika-peaks	Genauigkeit [%]	Trefferquote [%]	F_1-Maß [%]
AcorA	Permutationstest	63	42	67	89	76
Lasso	intrinsisch	2	2	100	4	8
Elastic Net	intrinsisch	64	43	67	91	77
PLSR VIP	VIP	23	16	70	34	46
QPAR (SR)	F-Test	13	12	92	26	40
Random Forest	1. Selektion	157	46	29	98	45
Random Forest	2. Selektion	8	7	88	15	25

eine vergleichsweise niedrige Trefferquote (34 %). Parallel ist die Genauigkeit mit 70 % nur von mittlerer Güte, sodass diese Methode im Vergleich zu Elastic Net und AcorA als deutlich schlechter zu bewerten ist.

3.2.8. Diskussion

In dieser Vergleichsstudie wurden eine Reihe von uni- und multivariaten Methoden im Hinblick auf ihre Eigenschaften zur Variablenselektion in massenspektrometrischen Datensätzen hin untersucht. Neben den klassischen parametrischen Methoden wie PCA und PLSR wurden auch Random Forest als nicht-parametrische Methode sowie Regularisierungsmethoden verwendet. Grundsätzlich ist die Leistungsfähigkeit einer Analysemethode abhängig vom jeweiligen Datensatz, sodass es keine universale Methode gibt, die in jedem Fall optimale Ergebnisse liefern würde [132]. In der vorliegenden Studie wurden jedoch einige Tendenzen erkennbar, die im Zusammenhang mit der Literatur im Folgenden diskutiert werden sollen:

1. Eine Hauptkomponentenanalyse ist für die Identifizierung bioaktiver Komponenten vergleichsweise ungeeignet.
2. Die Leistungsfähigkeit zur Identifizierung aktivitätsrelevanter Metaboliten nimmt in der Reihenfolge Varianz-basierte Methode (PCR) < Kovarianz-basierte Methode (PLSR) < orthogonale Kovarianz-basierte Methode (OPLS, Target Projection) zu.
3. Die Variable Importance in Projection (VIP) ist bei unilateralen Fragestellungen ungeeignet.

4. Von den Regularisierungsmethoden eignet sich die Elastic Net Analyse aufgrund der kombinierten Eigenschaften aus ℓ_1 und ℓ_2 Regularisierung besonders gut zur Selektion untereinander korrelierter Variablen.

5. Die univariate AcorA-Methode bietet ein ausgewogenes Verhältnis aus Praktikabilität und Leistungsfähigkeit in Bezug auf die Identifizierung eigenschaftsrelevanter Metaboliten.

Die PCA hat sich zur Identifzierung der aktiven Komponenten als vergleichsweise unbrauchbar erwiesen. Da die Berechnung der Hauptkomponenten lediglich die größte Varianz in **X** ausschöpft, ist *a priori* kein zur Bezug Bioaktivität gegeben. Durch die klassische Standardisierung der Datenmatrix wurde lediglich eine Auftrennung der inaktiven *Hygrophorus* Basisextrakte erzielt. Dieser Ansatz ist somit zwar für chemotaxonomische Studien interessant [340, 341], eine gezielte Selektion der bioaktiven Komponenten scheint damit jedoch nicht möglich. Durch Datenzentrierung konnten die Ergebnisse leicht verbessert werden. Allerdings wird die Information über die bioaktiven Komponenten auf mehrere Hauptkomponenten verteilt. Zudem geben die Vorzeichen der Loadings keinen Aufschluss über die Richtung der Korrelation. Die PCA wird in vergleichbaren Studien lediglich zur rein explorativen Daten- bzw. Ausreißeranalyse eingesetzt [342, 138, 142, 164, 343].
Deutlich bessere Ergebnisse lieferten die parametrischen, multivariaten Regressionsmethoden, bei denen eine Regression des Bioaktivititätsvektors **y** auf die Datenmatrix **X** erfolgt. Wie aus Tabelle 3.2 hervorgeht, nimmt die Fähigkeit zur Selektion der aktivitätsrelevanten Massensignale der getesteten Methoden in der Reihenfolge Varianz-basierte Methode (PCR), Kovarianz-basierte Methode (PLSR) und orthogonale Kovarianz-basierte Methode (Target Projection, QPAR) zu. Zu einem ähnlichem Ergebnis kommen Dumarey *et al.* [138]. Sie testeten u. a. PCR, PLSR und dessen orthogonale Variante OPLS auf ihrer Fähigkeit zur Vorhersage der oxidativen Kapazität von Grüntee Extrakten anhand von HPLC-Spektren. Analog zur vorliegenden Studie nahm der RMSEP in der Reihenfolge PCR, PLSR, OPLS ab. Simultan stieg die Interpretierbarkeit der Modelle sowie die der Regressionskoeffizienten. Der gegenüber der PCR verringerte RMSEP und insbesondere die verbesserte Selektion der aktivitätsrelevanten Signale durch die PLSR erkärt sich dadurch, dass in der PLSR die Regressanden zur Berechnung der latenten Variablen einbezogen werden. Unter der Annahme der Orthogonalität der x-Variablen geben die resultierenden Regressionskoeffizienten Aufschluss über Richtung und - nach Standardisierung der Daten - auch Stärke der Assoziation der Massensignale zur Bioaktivität [333]. Wie die Ergebnisse der PCA gezeigt haben, wird durch die Standardisierung jedoch die Bedeutung der Signale reziprok zu ihrer Intensität ver-

ändert. Im Ergebnis wurden die Antibiotikapeaks in der Fülle der Signale der nicht-aktiven Verbindungen „verschleiert“. Dieser Effekt wurde bereits in den 1980er Jahren u. a. von Wold [344] und Kvalheim [345] diskutiert. In der Konsequenz empfiehlt sich für die Analyse chemometrischer Datensätze eine Datenzentrierung oder beispielsweise eine Paretoskalierung [346]. Durch die Pareto-Skalierung wird die Bedeutung der großen Signale reduziert; die generelle Datenstruktur bleibt jedoch intakt [347].
Die Voraussetzung orthogonaler x-Variablen ist in massenspektrometrischen Rohdatensätzen aufgrund der großen Anzahl an Massensignalen sowie der vorhandenen Isotopen- und Adduktpeaks praktisch immer verletzt. Viele der Massensignale erzeugen somit eine Variation in **X**, die nicht mit der Bioaktivität assoziiert ist. Diese orthogonale Variation kann die Interpretation der Regressionskoeffizienten beeinträchtigen [348, 141].
Einen Ausweg bieten u. a. die OPLS und die in dieser Studie getestete Target Projection. Während die OPLS die zur Bioaktivität orthogonale Variation in **X** entfernt, verwendet die Target Projection die normalisierten Regressionskoeffizienten aus der PLSR, um die für die Bioaktivität relevanten Informationen zu extrahieren. Wie Kvalheim [156, 157] gezeigt hat, liefern beide Methoden mit derselben Anzahl an PLS-Komponenten identische Ergebnisse zur Vorhersage der Bioaktivität und repräsentieren somit lediglich zwei unterschiedliche Algorithmen mit dem Ziel die für y relevanten Informationen auf eine einzige prädiktive latente Variable (Komponente) zu projizieren. Auf diese Weise ergibt sich gegenüber PCR und PLSR eine verbesserte Modellinterpretation [140].
Innerhalb der vorliegenden Studie hat die Target Projection die höchsten Trefferquoten erzielt. Durch Projektion der normalisierten Regressionskoeffizienten auf **X** und anschließende Kalkulation der Target Projected Loadings, wiesen die meisten der Antibiotikapeaks höhere Werte auf als die Peaks der inaktiven Verbindungen. Durch anschließende Berechnung der Selectivity Ratios wurde das Antibiotikasignal-Rausch-Verhältnis nochmals verbessert, sodass auch kleinere Antibiotikapeaks, die in den anderen Varianz/Kovarianz-basierten Methoden übersehen wurden, innerhalb der Top 50 Peaks annotiert werden konnten.
Die QPAR Methode, bestehend aus Target Projection und Selectivity Ratios, wurde in verschiedenen Studien zur Identifizierung von aktivitätsrelevanten Metaboliten aus HPLC-Spektren erfolgreich angewandt [160, 349, 350, 158]. In einer Studie mit artifiziellen Extrakten konnten Kvalheim *et al.* sogar die gespickten bioaktiven Komponenten in der Reihenfolge ihrer Aktivität richtig vorhersagen [161].
Bedenklich ist jedoch die Bestimmung des kritischen Wertes für die Selectivity Ratios. Auf dem vorliegenden Datensatz wurden sehr viele durch die Selectivity Ratios klar von den Rauschpeaks abgegrenzten Antibiotikapeaks verworfen. Mit einer weniger konservativen

Schwelle wären deutlich mehr Peaks als Richtig-Positiv klassifiziert worden.
Ähnliches wurde auch in Studien von Farrés [351] und Tran [352] beobachtet. Farrés *et al.* verglichen die Leistungsfähigkeit der Selectivity Ratio und VIP-Methode zur Variablenselektion bei GC-MS- und DNA-Transkriptions-Datensätzen. Sie kamen zu dem Schluss, dass die SR-Methode vergleichsweise wenig Peaks selektiert und eine Tendenz zu falsch negativen Ergebnissen aufweist. In einer weiteren Vergleichsstudie untersuchten Tran *et al.* die Leistungsfähigkeit der SR- und VIP-Methode anhand eines simulierten Datensatzes mit bekannter Datenstruktur sowie eines Datensatzes bestehend aus Nahinfrarotspektren zur Qualitätskontrolle von Tabletten. Genau wie [351] attestieren sie der SR-Methode eine Tendenz zu falsch negativen Ergebnissen. Sie begründen die zu konservative Variablenselektion mit der falschen Verwendung der Freiheitsgrade in dem von Rajalahti [159] verwendeten F-Test und schlagen n-1 und n-2 (statt n-2 und n-3) Freiheitsgrade für die erklärte bzw. Restvarianz vor. Ein anderes Verfahren zur Bestimmung der kritischen Grenze verwendeten Campos *et al.* [353]. Im Anschluss an eine PLS-Regression nutzten sie den Mittelwert der Selectivity Ratios als kritischen Grenzwert und konnten auf diese Weise eine Reihe von phänotyprelevanten Genen auf Basis eines Genexpressionsdatensatzes identifizieren.

Die oben angesprochene VIP-Methode ist in der PLSR-Analyse von chemometrischen und metabolomischen Datensätzen eine weit verbreitete Variablenselektionsmethode und wird beispielsweise in Metaboanalyst [143] und SIMCA (Umetrics) standardmäßig angeboten [334, 118]. Die VIP spiegelt die Kovarianz zwischen $\mathbf{y}$ und $\mathbf{X}$ sowie den Anteil der erklärten Varianz einer Variable innerhalb einer latenten Variable a wider.
Wie [351] und [352] kongruent berichten, tendiert die VIP- im Vergleich zur SR-Methode zur Selektion einer größeren Anzahl von Signalen mit tendenziell zu vielen falsch positiven Resultaten. Die Ergebnisse der vorliegenden Dissertation bestätigen diese Tendenz. Zwar wurde der Vorhersagefehler nach der Variablenselektion mit der VIP gegenüber den PLSR Regressionskoeffizienten reduziert. Die für die Selektion der bioaktivitätsrelevanten Peaks wichtigere Trefferquote lag jedoch deutlich niedriger. Der Grund für die höhere Falsch-Positiv-Rate ist darin zu sehen, dass in der VIP Variablen mit hoher prädiktiver Bedeutung unabhängig ihres Vorzeichens gewichtet werden. Für viele Anwendungen bei denen bidirektionale Ergebnisse relevant sind (beispielsweise unter Berechnung von Fold Changes), führt dies zu dem gewünschten Effekt, dass jeweils die wichtigsten Variablen aus beiden Richtungen selektiert werden. Bei unidirektionalen Fragestellungen, bei denen jedoch nur eine Kovarianzsrichtung relevant ist, steigt durch die Vermengung von Variablen mit positiver und negativer Kovarianz die Falsch-Positiv-Rate.
Summa summarum scheint die SR-Methode der VIP-Methode für die Selektion aktivitätsre-

levanter Variablen überlegen zu sein. Durch eine geeignetere Wahl des Schwellenparameters, wie beispielsweise in [353] oder [354] beschrieben, stellt die SR-Methode in Anschluss an eine Target Projection oder (O)PLSR eine attraktive Variablenselektionsmethode dar.

Die große Anzahl von Signalen in Massenspektren führt zu dem Problem der Multikollinearität mit der Folge, dass die Kovarianzmatrix und damit auch die Regressionskoeffizienten in parametrischen Methoden nur ungenau geschätzt werden können [169, 170].
Regularisierungsmethoden wie die Ridge Regression, Lasso oder Elastic Net bieten die Möglichkeit die Varianz der Regressionskoeffizienten bei gleichzeitig nur geringer Abnahme der Erwartungstreue zu reduzieren. Der in der vorliegenden Arbeit verwendete Ansatz, den Anteil von ℓ_1 und ℓ_2 Regularisierung über den Faktor α mit Werten zwischen 0 und 1 zu variieren, bietet den Vorteil, dass innerhalb eines Skriptdurchlaufs alle drei Methoden optimiert und getestet werden können.
Von den drei getesteten Regularisierungsmethoden hat sich insbesondere das Elastic Net als exzellente Variablenselektionsmethode herausgestellt. Aufgrund der Kombination aus ℓ_1 und ℓ_2 Regularisierung scheint die Elastic Net Analyse zur Selektion von bioaktivitätsrelevanten Signalen aus Massenspektren geradezu prädestiniert zu sein. Während die ℓ_1 Regularisierung die Regressionskoeffizienten der mit der Bioaktivität nur schwach assoziierten Signale auf Null setzt, führt die ℓ_2 Regularisierung dazu, dass Isotopen- und Adduktpeaks aufgrund ihrer hohen Kollinearität untereinander gruppenweise selektiert werden. Auf diese Weise wird die Annotation der aktivitätsrelevanten Peaks und infolgedessen auch die Identifizierung der bioaktiven Komponente(n) erleichtert.
In der Tat hat die Elastic Net Analyse in den bislang wenigen Studien, in denen sie zur Prädiktion und Variablenselektion an Massenspektrometrie-basierten Datensätzen verwendet wurde, sehr gute Ergebnisse geliefert [355, 356, 357, 175, 358]. Beispielsweise verglichen Acharjee *et al.* Elastic Net, Lasso, Ridge Regression, PCR, PLSR, Random Forest und SVM Regression in Bezug auf ihre Leistungsfähigkeit zur Vorhersage der Kartoffelfarbe auf Basis eines LC-MS Datensatzes [355]. Dabei lieferten die Methoden mit intrinsischer Variablenselektion - allen voran die Elastic Net Analyse - die Modelle mit den geringsten Prädiktionsfehlern. In einer weiteren Studie zur Variablenselektion wurden verschiedene uni- und multivariate Methoden an einem simulierten Datensatz aus 400 Proben mit 10 Richtig-Positiven Prädiktoren sowie 50 Rauschvariablen mit variablem Grad an Multikollinearität getestet [359]. Auch in dieser Studie haben insbesondere die Elastic Net und die Lasso Analyse in vielen Fällen die besten Ergebnisse geliefert.
Die Elastic Net Eigenschaften könnten auch für die Auswertung von NMR-Spektren in metabolomischen/naturstoffchemischen Untersuchungen interessant sein. Ähnlich der Massen-

spektrometrie erzeugt eine Verbindung in der NMR-Spektroskopie strukturabhängig mehrere Resonanzsignale, die untereinander hoch korreliert sind. Alves *et al.* [360] haben gezeigt, dass zwei Peaksignale mit einem Korrelationskoeffizienten > 0,89 mit hoher Wahrscheinlichkeit demselben Molekül zuzuordnen sind. Besitzt diese Verbindung zudem eine hohe Assoziation zu einer gemessenen Bioaktivität, sollte man die aktivitätsrelevanten Signale der Verbindung nach Messung eines Extraktgemisches mithilfe der Eigenschaften der Elastic Net Analyse selektieren können.

Tatsächlich wird die hohe Multikollinearität von NMR-Resonanzsignalen bereits in dem s. g. Statistical Total Correlation Spectroscopy (STOCSY) Verfahren [361] genutzt. In der STOCSY wird die Korrelationsmatrix der Peakintegrale der δ-ppm Intervalle berechnet und als s. g. STOCSY-Chromatogramm (ähnlich eines TOCSY-Spektrums) dargestellt. Mithilfe des STOCSY-Spektrums können somit einzelne Resonanzsignale einer Verbindung innerhalb eines Substanzgemisches zugeordnet werden. Eine anschließende OPLS-Analyse desselben Datensatzes stellt dann den Bezug zu einer zuvor gemessenen Bioaktivität her [361]. Die Überlagerung der OPLS-Regressionskoeffizienten mit den STOCSY Daten erlaubt schließlich die Identifizierung der aktivitätsrelevanten Signale sowie deren Zuordnung als Strukturelemente eines Moleküls. Im Gegensatz zum Elastic Net sind in dem STOCSY/OPLSR-Verfahren dazu allerdings zwei Schritte notwendig. Zudem besteht durch die große Anzahl der Resonanzsignale die Gefahr der Überanpassung, für die die (O)PLS besonders anfällig ist [362, 363, 114].

Wie oben erwähnt, hat sich auch die Lasso Analyse in verschiedenen Studien insbesondere für Prädiktion und Variablenselektion bewährt. Da Lasso aus einer Reihe untereinander hoch korrelierter Signale lediglich eines zufällig auswählt, gehen bei der Auswertung massenspektrometrischer Datensätze jedoch wichtige Informationen verloren, die für die Identifizierung der aktiven Komponenten von Nutzen sein könnten. Zudem wurde mit Lasso in der vorliegenden Studie lediglich eines von zwei Antibiotika gefunden. Offensichtlich wurde der Penalisierungsparameter λ zu groß gewählt, sodass die Regressionskoeffizienten einiger relevanter Massensignale auf Null gesetzt wurden. Ursache hierfür ist wahrscheinlich die schlecht konditionierte Datenmatrix. Die mit 21 Proben ohnehin nur sehr kleine Probenanzahl wurde innerhalb der doppelten Kreuzvalidierung weiter reduziert. Damit standen zur Erstellung der Modelle jeweils nur vergleichsweise wenige Proben zur Verfügung. Als Konsequenz entstanden für verschiedene α Werte z. T. bimodale Verteilungen des Penalisierungsparameters λ, sodass die Ermittlung von λ_{opt} erschwert wurde. Wünschenswert wäre daher eine größere Anzahl von Proben. Da dies in vielen Fällen nicht möglich ist, könnte Boostrapping eine Alternative zur Verbesserung der Parameterschätzung darstellen.

Schlecht konditionierte Datenmatrizen sind ein generelles Problem der parametrischen, multivariaten Machine Learning Methoden [363, 364]. Zum einen können - wie bei den Regularisierungsmethoden beobachtet - Probleme bei der Wahl der zu optimierenden Parameter entstehen. Zum anderen können die Modelle zu stark an den Trainingsdatensatz angepasst werden, sodass weder eine sinnvolle Generalisierung noch eine brauchbare Variablenselektion möglich ist. Insbesondere die Partial-Least-Squares Methoden sind bei geringer Probenanzahl empfindlich für Überanpassung [365, 179, 178]. Im Extremfall können bei unsachgemäßer Anwendung vermeintlich gute Modelle erstellt werden, die jedoch auch durch eine rein zufällige Verteilung der Daten hätten entstehen können [179].
Aus diesen Gründen benötigen die Anwender der oben beschriebenen Methoden vertiefte Kenntnisse in der Modellierung von hochdimensionalen Datensätzen [181]. Da diese Spezialkenntnisse bei Naturstoffchemikern i. d. R. weniger vorhanden sind, stellen Methoden wie AcorA und die Random Forest Analyse eine interessante Alternative dar.
Beide Methoden benötigen keine aufwendige Parameteroptimierung und können „out-of-the-box“ verwendet werden. Im Gegensatz zu den varianzbasierten Methoden werden kleinere Peaks mit geringer Varianz durch die nicht-parametrische Modellierung bei AcorA und Random Forest nicht diskriminiert. Auf diese Weise sollten z. B. auch schlecht ionisierbare oder gering konzentrierte Wirkstoffe besser in einer Hitliste erfasst werden können.
Die Random Forest Analyse hat sich in verschiedenen Metabolomics- [117, 366, 367], QSAR- [126, 124] und Microarray-Studien [335, 336] als sehr gute Methode zur Prädiktion und Variablenselektion herausgestellt. Umso überraschender war die vergleichsweise geringere Genauigkeit bei dem hier untersuchten Datensatz. Vor allem Signale des Erythromycins wurden erst ab Position 50 in der Hitliste annotiert. Ursache hierfür war möglicherweise das Ungleichgewicht zwischen den Proben, die Rifampicin (5 Proben) und Erythromycin (2 Proben) in aktivitätsrelevanten Konzentrationen enthielten. In der Literatur ist dieses Problem vor allem bei binären Klassifikatoren auch als Klassenungleichgewicht (class imbalance) bekannt [298]. D. h. die Prädiktion der häufiger auftretenden Komponente gelingt besser als die der Komponente in der Minderheit. Inzwischen gibt es weiterentwickelte Random Forest [368] und Kreuzvalidierungs Algorithmen [298, 369, 319], die die prädiktive Güte verbessern können. Allerdings beziehen sich diese auf Klassifizierungsprobleme. Es könnte daher sinnvoll sein, die gemessene Bioaktivität in Aktivitätsklassen einzustufen und diese als Regressanden für die Random Forest Analyse zu verwenden.
Trotz des eher simplen, univariaten Verfahrens hat die vorgestellte AcorA-Methode auf dem untersuchten Datensatz z. T. bessere Ergebnisse geliefert als die deutlich komplexeren multivariaten Methoden. Tatsächlich können univariate Methoden in vielen Fällen ähnliche

Resultate liefern wie ihre multivariaten Pendants [111, 174, 300, 370, 160]. Insbesondere bei schlecht konditionierten Datensätzen ist die Schätzung der Populationskovarianzmatrix über die empirische Probenkovarianzmatrix ungenau [170, 169]. Die Folgen sind u. a. Überanpassung und instabile multivariate Modelle.
Univariate Methoden können somit eine attraktive Variante zu den komplexeren multivariaten Methoden darstellen. Tatsächlich konnten mithilfe von AcorA, sowohl in dieser Dissertation als auch in verschiedenen anderen Studien, erfolgreich bioaktive Substanzen *in silico* identifziert und durch anschließende Isolierung und Testung in ihrer Aktivität bestätigt werden [277, 293, 371, 372]. Die frühzeitige Identifizierung der aktivitätsrelevanten Massensignale erlaubt eine *m/z*-geleitete Isolierung. Die AcorA-Methode scheint somit für die Identifizierung eigenschaftsrelevanter Metabolitencluster geeignet zu sein.
Dass die Korrelation zwischen bioaktiven Extrakten/Fraktionen mit massenspektrometrischen Daten zur Identifizierung von aktivitäsrelevanten Metaboliten geeignet ist, wurde 2012 auch in einer Publikation von Inui *et al.* bestätigt [373]. Durch die Berechnung der Pearson-Korrelationen zwischen den Signalen aus GC-MS Spektren von *Oplopanax horridus* (Igelkraftwurz) Extraktfraktionen und deren Wachstumsinhibtion von *Mycobacteria tuberculosis*, konnten einige synergistisch wirkende Komponenten (Sesquiterpene und Polyketide) identifiziert werden. Da der Pearson-Korrelationskoeffizient sehr sensitiv in Bezug auf Ausreißer reagiert [374], scheint dessen Verwendung jedoch bedenklich.

Ausblick Von den untersuchten Methoden haben sich QPAR, Elastic Net und AcorA auf dem untersuchten Datensatz als hervorragende Methoden zur Identifizierung der bioaktiven Komponenten in komplexen Mischungen herausgestellt. Während für QPAR und Elastic Net vertiefte Kenntnisse in Datenmodellierung benötigt werden, besticht die entwickelte AcorA Methode durch ihren intuitiven Ansatz bei gleichzeitig hoher Leistungsfähigkeit.
Neben den untersuchten Methoden gibt es jedoch noch unzählige weitere Analysemethoden, die für die Identifizierung von eigenschaftsrelevanten Metabolitenclustern von Interesse sein könnten. So bietet beispielsweise die Anwendung der „Multivariate Adaptive Regression Splines“ (MARS) die Möglichkeit einer multivariaten, nicht-linearen Regression mit intrinsischer Variablenselektion [375, 319]. Da hier kein linearer Zusammenhang zwischen Regressoren und Regressand angenommen wird, könnte diese Methode beispielsweise zur Identifizierung von synergistisch wirkenden Substanzen interessant sein. Weitere Möglichkeiten zur nichtlinearen Regression, allerdings ohne intrinsische Variablenselektion, bieten u. a. Artificial Neural Networks (ANN) und Support Vector Machines (SVM) [105, 127, 118].
Alternativ zur Regression auf kontinuierliche Messwerte, kann die Modellierung der Bio-

aktivität auch als Klassifikationsproblem aufgefasst werden. Dies gilt umso mehr für die Fälle, bei denen die Bioaktivität nur mit hoher Standardabweichung bestimmt werden kann. Anstatt die Originalwerte der Bioaktivitäsmessung als Zielvariablen für die Regression zu verwenden, werden die Messwerte in Aktivitätsklassen eingeteilt. Beispielsweise könnte man die Bioaktivitätswerte in Kategorien wie „inaktiv“ (0 -25 %), „schwach aktiv“ (26-50 %), „mittel aktiv“ (51-80 %) und „hoch aktiv“ (81-100 %) einordnen. Anschließend können dann verschiedene Verfahren der Diskriminanzanalyse dazu verwendet werden, die Bioaktivitätsklasse aus **X** (Massenspektren, NMR-Spektren etc.) vorherzusagen. Die Variablen mit den höchsten prädiktiven Eigenschaften sollten dann am stärksten mit der Bioaktivität assoziiert sein.
Die Methode der „Nearest Shrunken Centroids“ von Tibshirani [376] stellt in diesem Zusammenhang eine besonders attraktive Klassifikationsmethode dar, da sie, ähnlich dem Lasso, über eine intrinsische Variablenselektion verfügt. Als weitere, klassische Methoden der Diskriminanzanalyse, die sowohl in Metabolomics als auch QSAR [130] Analysen Verwendung finden, seien Support Vector Machines (SVM), Artificial Neural Networks (ANN) und Lineare (LDA) bzw. quadratische Diskriminanzanalyse (QDA) genannt [3, 319].
Es gibt eine Vielzahl von weiteren Analysemethoden, die zur Prädiktion und Variablenselektion verwendet werden können. Es sollte jedoch immer im Bewusstsein bleiben, dass es keine universale Analysemethode („*Eierlegende Wollmilchsau*“) gibt, die auf jedem Datensatz die qualitativ besten Ergebnisse liefern würde [132]. Da *a priori* nicht bekannt ist, welche Methode auf dem zu analysierenden Datensatz die besten Resultate liefern wird, scheint die Aggregation von Ergebnissen verschiedener Methoden ein vielversprechender Weg zu sein [377], der auch zur Identifizierung aktivitäsrelevanter Metabolitencluster interessant sein könnte.
Ferner sollte niemals außer Acht gelassen werden, dass selbst die besten Datenanalysemethoden lediglich Modelle aufstellen, die auf vielfältige Art und Weise störanfällig sind. Letzlich können alle Modelle immer nur zur Generierung von Hypothesen dienen, die experimentell bestätigt werden müssen. Um die tatsächliche Isolierung, Strukturaufklärung und Testung im Bioassay kommt ein Experimentator nicht herum!
Mithilfe der diskutierten chemoinformatischen Methoden lassen sich die Dereplikation und der allgemeine Arbeitsaufwand jedoch erheblich reduzieren.

3.3. AcorA mit S. ampullosporum

Das Proof of Concept Experiment hat gezeigt, dass AcorA unter artifiziellen Bedingungen sehr gut geeignet ist, um biologisch aktive Verbindungen in komplexen Mischungen identifizieren zu können. Hieran schließt sich die Frage an, ob AcorA auch unter den Bedingungen, mit denen sich (Naturstoff-)Chemiker in der alltäglichen Laborarbeit konfrontiert sehen, ähnlich gute Ergebnisse liefert. D. h. können bislang unbekannte (oder auch bekannte), biologisch aktive Substanzen in komplexen Mischungen durch die Aktivitäts-Korrelations-Analyse identifiziert werden?
Grundlage für eine Identifizierung mittels AcorA ist eine möglichst hinreichende Varianz insbesondere der aktiven Verbindung(en) in den zu untersuchenden Extrakten. Michels konnte durch verschiedene physikalische und chemische Modifikationen an Aliquots eines biologisch aktiven Extraktes von *Hygrophorus latitabundus* zeigen, dass die durch die Modifikationen generierte Varianz ausreicht, um anhand der Aktivitäts-Korrelations-Analyse Hinweise auf das verursachende Prinzip zu erhalten. Sie konnte mit Hilfe von AcorA zwei Substanzen identifizieren und isolieren, die an der Wachstumsinhibition von *Bacillus subtilis* beteiligt sind [277].
Im Rahmen der vorliegenden Dissertation sollte untersucht werden, ob die natürliche Varianz, die bei der Biosynthese von Metaboliten bei unterschiedlichen Varietäten einer Spezies auftritt, für AcorA genutzt werden kann. Dazu wurden 21 *Sepedonium ampullosporum* Stämme verwendet, die von unterschiedlichen Wirten der Ordnung *Boletales* an verschiedenen Standorten in Nordamerika und Europa isoliert wurden. Die Extrakte wurden auf ihre zytotoxische Aktivität in Hinblick auf humane Darmkrebszellen untersucht. Um Hinweise auf die aktiven Verbindungen zu erhalten, wurden die durch Messung der Extrakte mit dem FT-ICR-Massenspektrometer generierten Metabolitenprofile mit Hilfe von AcorA mit dem Bioaktivitätsprofil korreliert und ausgewertet.

3.3.1. Zytotoxische Wirkung von S. ampullosporum Extrakten auf HT29-Zellen

Die methanolischen Extrakte der parasitischen Pilze wurden in 3 Konzentrationen (5 µg/mL, 0,5 µg/mL, 0,05 µg/mL) auf ihre zytotoxische Aktivität gegenüber humanen Darmkrebszellen (HT-29) getestet (Abbildung 3.26). Die Extrakte in der höchsten Verdünnungsstufe (0,05 µg/mL) beeinflussten das Zellwachstum nur marginal und zeigten keine Aktivität, die über die Aktivität der Kontrolle (BW) hinausging. Im Gegensatz dazu waren, mit Ausnah-

me der Extrakte der Stämme KSH 498, 522 und 523, die methanolischen Extrakte der Konzentration 5 µg/mL für die HT-29 Zellen zu nahezu 100 % letal. Von den drei getesteten Konzentrationen zeigten lediglich die Extrakte mit einer Massenkonzentration von 0,5 µg/mL eine für AcorA hinreichende Variation der zytotoxischen Aktivität. Das entsprechende Bioaktivitätsprofil wurde anschließend mit Hilfe des AcorA-Paketes mit den FT-ICR-MS Messungen im Positiv-Ionen-Modus korreliert.

3.3.2. Aktivitäts-Korrelation-Analyse der S. ampullosporum Extrakte

Nach Peak-Picking und Alignment wurde eine Peakliste mit insgesamt 2306 Massensignale erhalten. Aus der anschließenden Aktivitäts-Korrelations-Analyse resultierte eine Hitliste

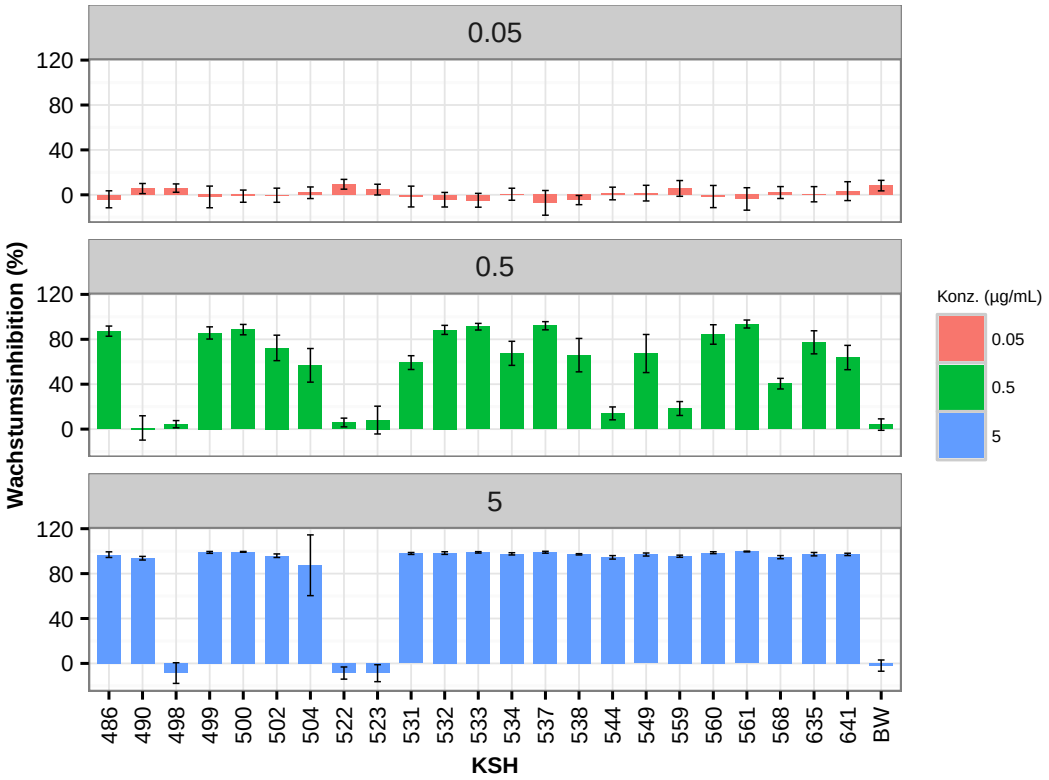


Abbildung 3.26.: Wachstumsinibition von HT-29 Zellen mit methanolischen Extrakten aus verschiedenen *S. ampullosporum* Stämmen. BW entspricht der Kontrolle, d. h. der Extrakt aus drei unbeimpften Agarplatten. Die Fehlerbalken geben die Standardabweichungen der Messungen wider.

mit 115 (Positiv-Modus) signifikanten Peaks (= 5,0 %, siehe Tabelle C.1, Anhang). Wie das Proof of Concept Experiment gezeigt hat, deuten Kombinationen aus Addukt- und Isotopenpeaks auf eine kausale Korrelation zwischen den Peaks und der Bioaktivität hin. Diese Peakcluster besitzen eine vergleichsweise hohe Peakdichte und geben somit einen Hinweis auf *m/z*-Regionen, die für die Analyse besonders interessant erscheinen. Die Auswertung der Hitliste sollte also erleichtert werden, indem man die Peakdichte in Abhängigkeit des *m/z*-Wertes graphisch darstellt.

In Abbildung 3.27 sind sowohl die Spearman Korrelationskoeffizienten als auch die Peakdichte in Abhängigkeit der *m/z*-Werte dargestellt. Im Bereich zwischen *m/z* 250 und 400 gibt es einige Peaks mit relativ hohen Korrelationskoeffizienten. Allerdings ist die Peakdichte in diesem Bereich vergleichsweise niedrig. Tatsächlich sind die betreffenden Peaks (*m/z* 274,35241, 327,20266, 367,26508 und 398,25115) in der Hitliste isoliert, d. h. es gibt keine weiteren Peaks, die mit ihnen im Sinne von Isotopen- oder Adduktpeaks assoziiert sind. Besonders auffällig sind hingegen zwei Bereiche, die in Abbildung 3.27 mit Cluster 1 und 2 gekennzeichnet sind. Beide Peakcluster verfügen über eine hohe Peakdichte und bestehen aus Massensignalen mit hoher Korrelation zur biologischen Aktivität.

3.3.2.1. Peakcluster 1 und 2

Cluster 1 liegt im Bereich zwischen *m/z* 811 und *m/z* 845. Isotopenmuster und Massenabstände von Δ10,99 zwischen den monoisotopischen Peaks *m/z* 811,98375, 822,97428 und 833,96485 deuten auf die $[M+2H]^{2+}$-, $[M+H+Na]^{2+}$- und $[M+2Na]^{2+}$-Ionen der Verbindung **61** hin (Abb. 3.28 a). Der Peak bei *m/z* 841,95210 entspricht einem Addukt aus Na^+ und K^+ von Verbindung **61** ($[M+Na+K]^{2+}$). Weiterhin können in der Hitliste signifikante Korrelationen zu einigen Isotopenpeaks der entsprechenden einfach geladenen Spezies beobachtet werden ($[M+H]^+$ = 1622,95777 und $[M+Na]^+$ = 1644,93157). Anschließende UPLC-QqTOF-MS Messungen zeigen, dass die Isotopenmuster in Peakcluster 1 faktisch auf zwei Verbindungen **61** (*m/z* 811,9848, Abb. 3.29 e) und **62** (*m/z* 812,4875, Abb. 3.29 f) zurückzuführen sind, die nach 7,1 und 7,2 Minuten von der Säule eluierten. Da sich die Verbindungen zusätzlich zur geringen Retentionszeitdifferenz nur um *m/z* 0,5 Da unterscheiden, ist davon auszugehen, dass es sich um strukturell sehr ähnliche Verbindungen handelt.

Cluster 2 (Abb. 3.27 b) besteht im Wesentlichen aus Peaks im Massenbereich zwischen *m/z* 697 und *m/z* 729. Auffällig sind wieder Massenabstände von Δ10,99 zwischen *m/z*

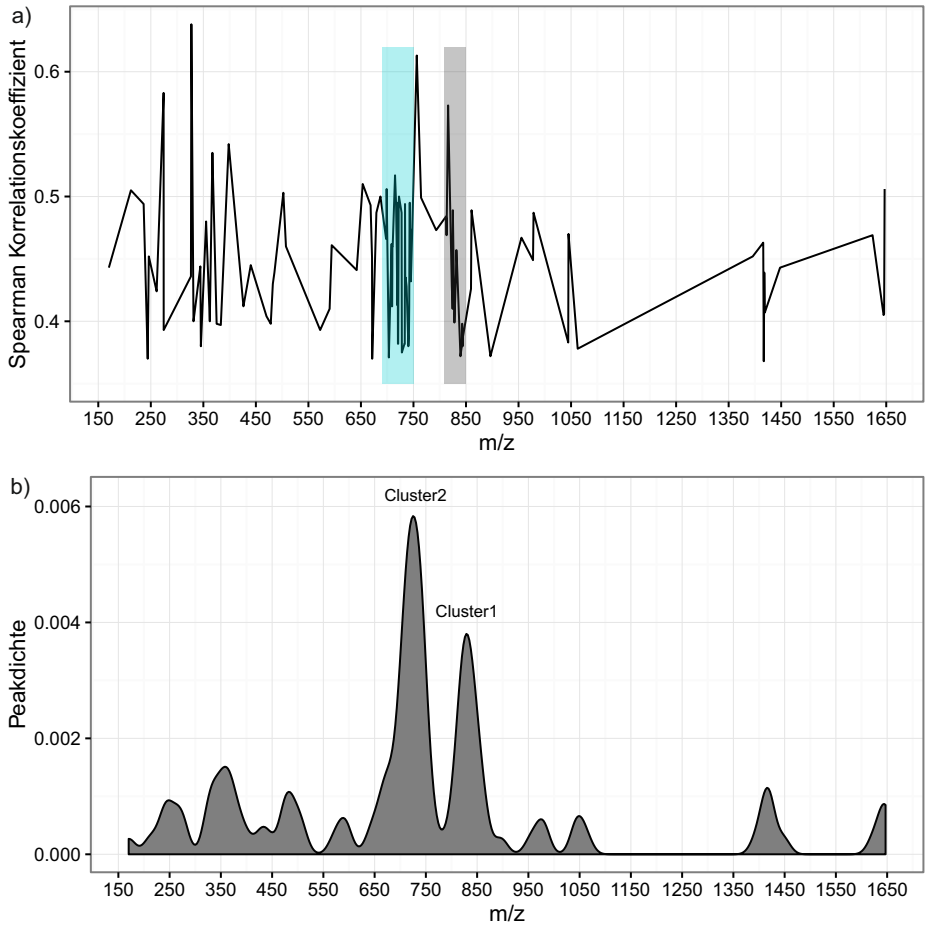


Abbildung 3.27.: a) Spearmann-Rangkorrelationskoeffizient und b) Peakdichte in Abhängigkeit von *m/z*. Bereiche mit hohem Korrelationskoeffzienten und hoher Peakdichte weisen auf signifikant korrelierende Signale, bestehend aus Isotopen- und Adduktpeaks, hin.

697,93946, 708,93053 und 719,2226, d. h. es handelt sich hier ebenfalls um eine $[M+2H]^{2+}$-, $[M+H+Na]^{2+}$- und $[M+2Na]^{2+}$-Ionenserie (Abb. 3.28 b). Die Signale bei *m/z* 708,4166 und *m/z* 719,40520 ($\triangle$10,99) weisen darauf hin, dass noch mindestens eine weitere Verbindung diesem Peakcluster zuzuordnen ist. In den vorliegenden FITCR-MS Spektren sind die $[M+H+Na]^{2+}$- und $[M+2Na]^{2+}$-Peaks in der Regel deutlich höher als die $[M+2H]^{2+}$-Peaks. Da ein hypothetischer $[M+2Na]^{2+}$-Peak bei *m/z* 730,39 nicht beobachtet werden kann, handelt es sich bei den zuvor erwähnten Peaks also vermutlich um die $[M+H+Na]^{2+}$- und $[M+2Na]^{2+}$-Ionen einer Verbindung, dessen $[M+2H]^{2+}$-Peak theoretisch bei *m/z* 697,43 liegen müsste. Die Ionenserien zwischen *m/z* 697 und *m/z* 729 bestehen somit aus einer

Überlagerung von Peaks zweier Verbindungen, die sich um eine Nominalmasse von 1 Da unterscheiden.

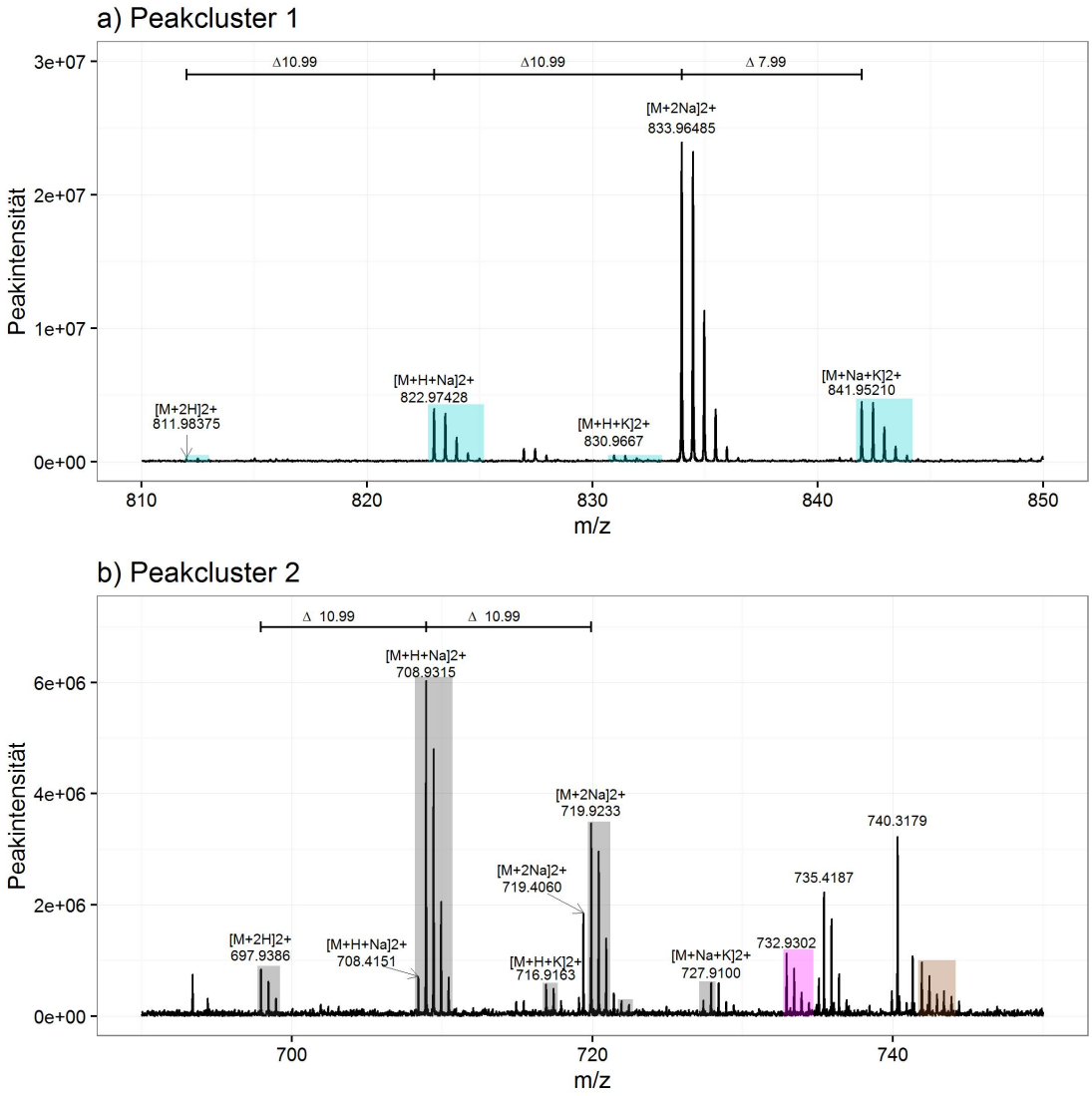


Abbildung 3.28.: Ausschnitte der Peakcluster 1 und 2 aus dem FT-ICR-MS Spektrum von KSH 537. Die signifikant korrelierenden Peaks sind farbig hervorgehoben. Peaks, die im Sinne von Isotopen- und Adduktpeaks miteinander assoziert sind, sind jeweils in der gleichen Farbe dargestellt.

Tatsächlich zeigen Messungen mit dem UPLC-QqTOF Massenspektrometer, dass es sich bei Peakcluster 2 um mindestens vier Verbindungen handeln muss, die in zwei Gruppen mit deutlich unterschiedlichen Retentionszeiten von der Säule eluierten. Verbindungen des Peakcluster 2.1 eluierten zwischen Minute 5,1 und 5,2 (Abb. 3.29 b und c). Es besteht aus den $[M+2H]^{2+}$-Peaks 697,9505 (**63**) und 698,4438 (**64**). Die Verbindungen in Peakcluster 2.2 (*m/z* 697,4345 ; **65** und *m/z* 697,9208 ; **66**) sind deutlich unpolarer und eluierten zwischen Minute 7,0 und 7,1 (Abb. 3.29 d und e) von der Säule. Wie Abbildung 3.28b zeigt, liegen innerhalb von Peakcluster 2 noch zwei weitere Peakgruppen mit signifikant korrelierenden Peaks. Die Peakgruppen scheinen jedoch weder mit den zuvor beschriebenen Peakgruppen noch untereinander im Sinne von Adduktpeaks assoziiert zu sein.

Korrelationsanalysen Peakcluster 1 und 2 Um die Beziehungen zwischen den Peaks in der Hitliste besser erfassen zu können, wurden die Pearson-Korrelationen sowie die entsprechenden Korrelationsnetzwerke berechnet. Abbildung C.1 im Anhang zeigt, dass die Peaks in Peakcluster 1 und 2 untereinander hoch korreliert sind und in der Korrelationsmatrix ein einziges Großcluster bilden. Die Annotation der Peaks (Tabelle C.1, Anhang) verdeutlicht, dass sich nahezu alle Peaks in diesem Großcluster auf Peptaibole zurückzuführen lassen. Neben den Isotopen- und Adduktpeaks der Pseudomolekülionen der Verbindungen **61** - **66**, konnten auch einige b- und y-Ionen (sowie deren Addukte) dieser Verbindungen annotiert werden. Diese sind wahrscheinlich durch Stoßaktivierung innerhalb der Ionenquelle (In-Source-Decay) entstanden. Einige wenige Peaks in diesem Peptaibolcluster (z. B. *m/z* 236,3303) konnten nicht direkt den untersuchten Peptaibolen zugeordnet werden. Die hohe Korrelation zu den Peptaibolsignalen legt jedoch nahe, dass die Peaks ebenfalls auf Peptaibolen beruhen. Möglicherweise handelt es sich dabei um Fragmentpeaks einiger in dieser Arbeit nicht sequenzierter Peptaibole.

Neben dem Großcluster ist in Abbildung C.1 ein zweites Cluster erkennbar, das aus Peaks besteht, die nicht annotiert werden konnten („NA-Cluster“). Eine Korrelation zu den Peptaibol-Peaks ist nicht vorhanden, die Peaks sind untereinander allerdings zum Teil stark korreliert (z. B. *m/z* 367,26508 und *m/z* 344,27957). Da viele dieser Peaks in nahezu allen Proben vorkommen, handelt es sich dabei jedoch vermutlich um falsch positive Korrelationen mit der Bioaktivität.

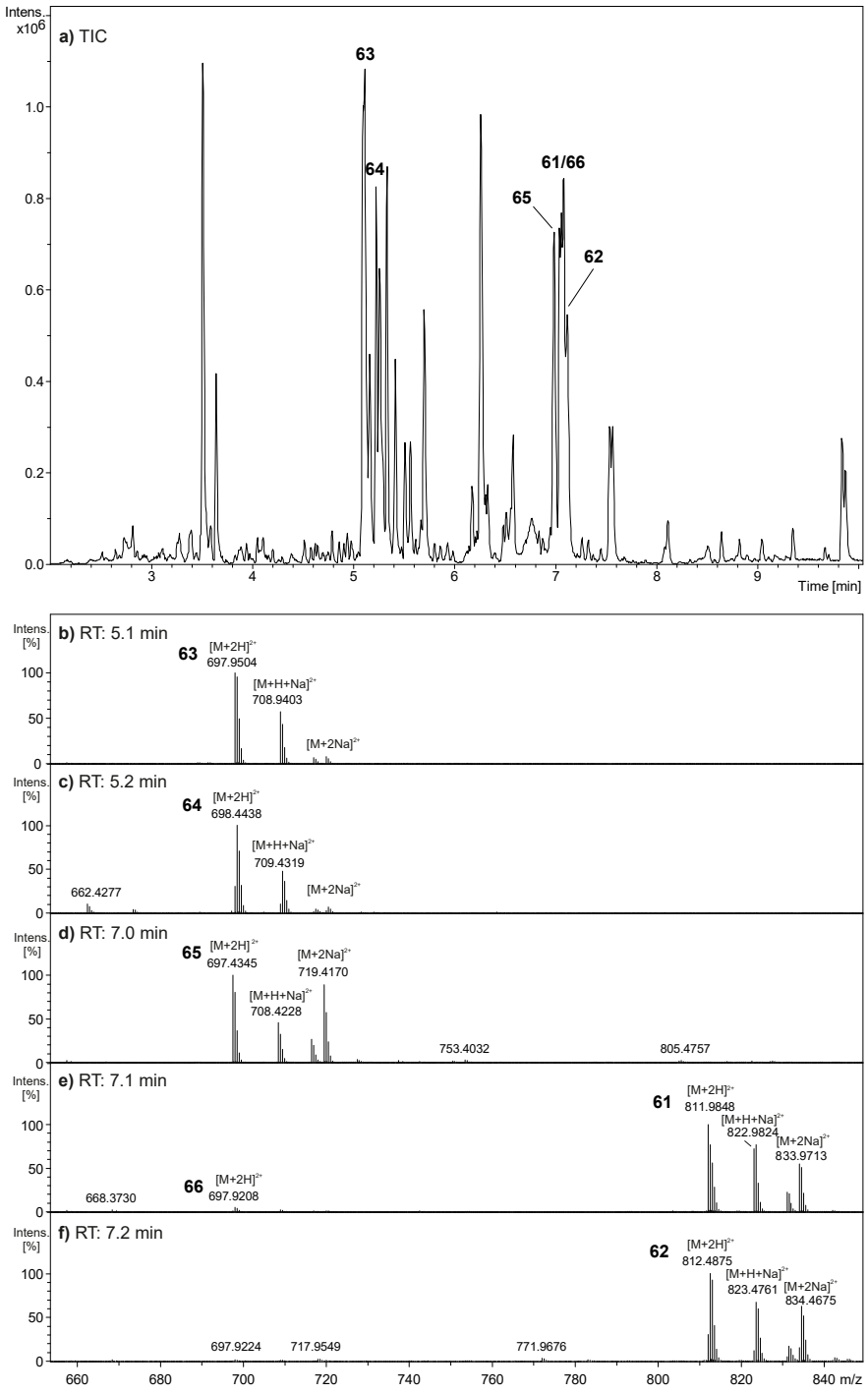


Abbildung 3.29.: Überblick über die Peakcluster 1 und 2 in der UPLC-QqTOF-MS. a) Totalionenstrom b) Verbindung **63** c) Verbindung **64** d) Verbindung **65** e) Verbindungen **66** und **61** f) Verbindung **62**.

3.3.2.2. Peakcluster 3 - 6

Interessanterweise enthält die Hitliste nahezu ausschließlich Massensignale aus den Extrakten der Stämme KSH 531, 533, 534, 537, 559, 561, 635 und 641. Die Messung der Bioaktivität gegenüber HT-29 Zellen (Abbildung 3.26) hatte jedoch gezeigt, dass eine Reihe weiterer Extrakte deutliche wachstumsinhibitorische Effekte von über 60 % besitzen (z. B. KSH 499, 500, 502, 504, 560). Aus einem nicht bekannten Grund wurden Peaks aus diesen Extrakten während der Aktivitäts-Korrelations-Analyse „maskiert", sodass sie nicht in der Hitliste auftreten.
Um nun auch die aktive(n) Verbindung(en) dieser Extrakte identifizieren zu können, wurde eine zweite Aktivitäts-Korrelations-Analyse mit einem reduzierten Datensatz durchgeführt. Dazu wurden die Peaks der Stämme KSH 531, 533, 534, 537, 559, 561, 635, und 641 aus der ursprünglichen Peakliste entfernt und sowohl Alignment als auch AcorA mit den verbliebenen Peaks (1975 Massensignale) durchgeführt.

Als Ergebnis wurde eine zweite Hitliste mit 94 Peaks erhalten (= 4,8 %, Tab. C.2). Der Peakdichte Plot zeigt vier Peakcluster, die in Abbildung 3.30 b farblich hervorgehoben sind. Auffällig ist eine generell erhöhte Peakdichte im Massenbereich zwischen *m/z* 300 und *m/z* 650. Wie die Hitliste 2 zeigt, sind in diesem Bereich überwiegend isolierte Peaks enthalten, die nicht miteinander assoziiert sind. Eine Ausnahme bildet *m/z* 510,26517 und dessen zugehörige Isotopenpeaks *m/z* 511,36895 und *m/z* 512,37068, die zu einer Verbindung **24** gehören und Peakcluster 6 bilden (Abb. 3.31 d).
In Peakcluster 3 und 4 zeigen sieben respektive zehn Peaks aus jeweils drei Peakgruppen eine signifikante Korrelation zur Bioaktivität. Sowohl die Isotopenmuster als auch die Massenabstände von Δ10,99 zwischen *m/z* 999,06118, 1010,05116 und 1021,04223 sowie *m/z* 999.56134, *m/z* 1010,54763 und 1021,54011 in Peakcluster 3 deuten auf die $[M+2H]^{2+}$-, $[M+H+Na]^{2+}$-, $[M+2Na]^{2+}$-Ionenserien zweier Verbindungen hin, deren Signale sich im FT-ICR-MS-Spektrum überlagern (Abb. 3.31 a).
Ein sehr ähnliches Muster wird auch in Peakcluster 4 beobachtet. Ausgehend von den monoisotopischen Peaks *m/z* 885,0291 und *m/z* 885,51424 sind Massensignale von mindestens zwei Verbindungen erkennbar, deren Isotopenmuster sich in drei Peakgruppen aus $[M+2H]^{2+}$-, $[M+H+Na]^{2+}$-, $[M+2Na]^{2+}$-Addukten überlagern (Abb. 3.31 b).

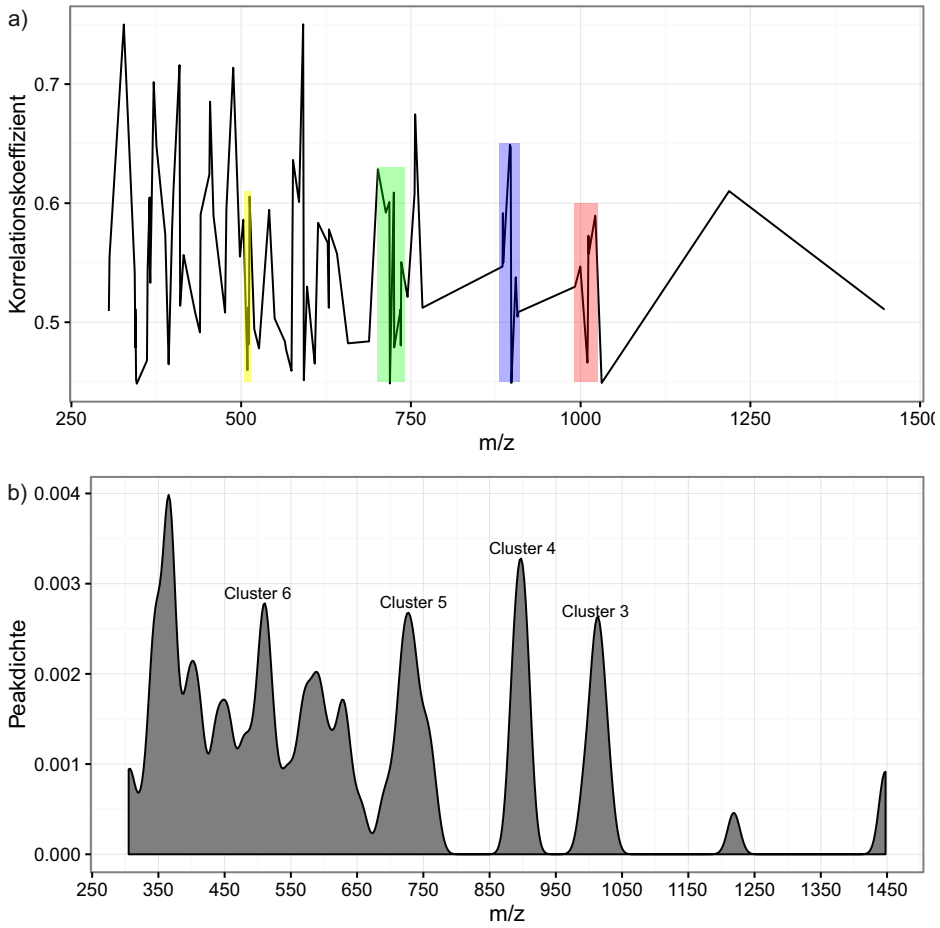


Abbildung 3.30.: a) Spearman-Rangkorrelationskoeffizient und b) Peakdichte in Abhängigkeit von *m/z* der zweiten Hitliste mit den Peakclustern 3-6.

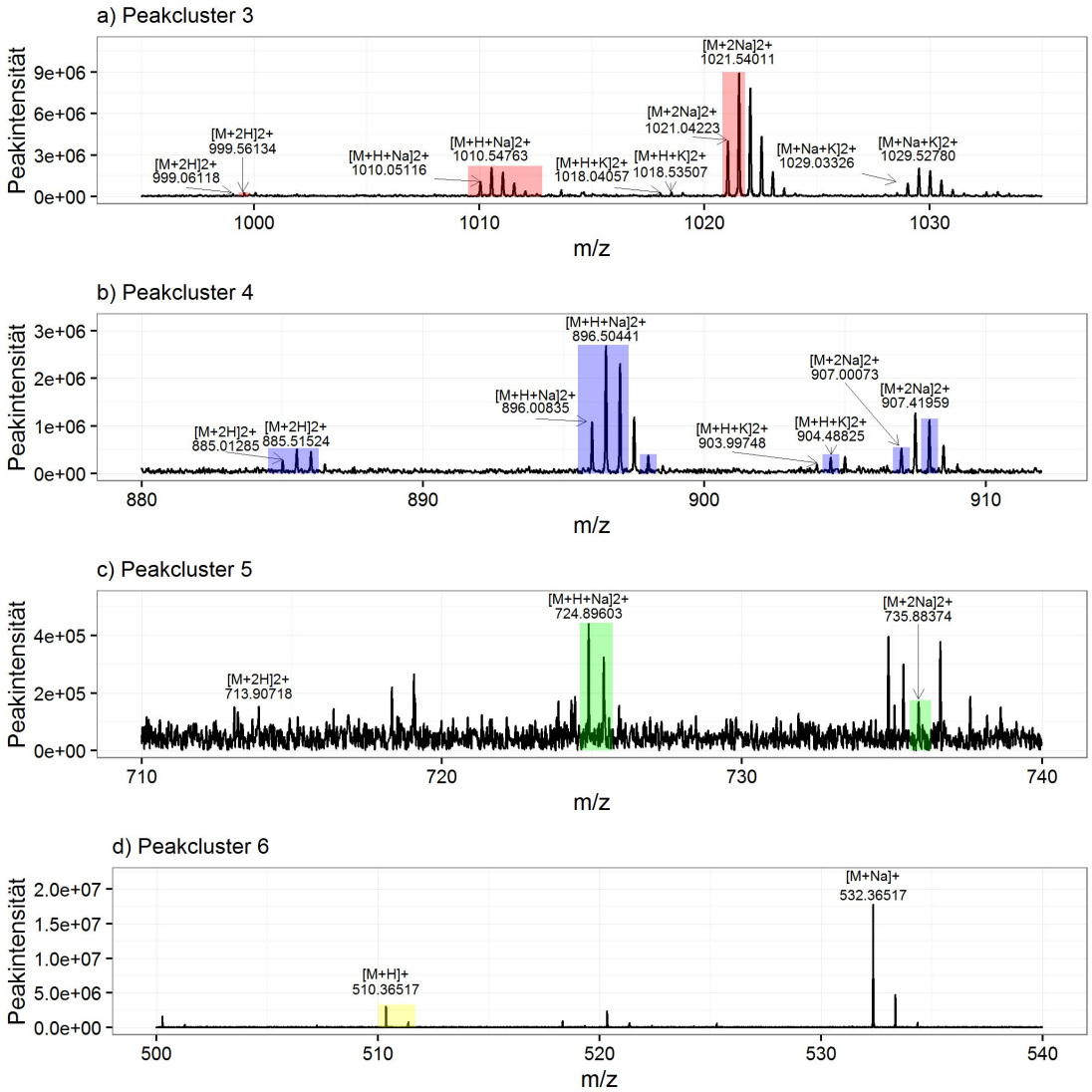


Abbildung 3.31.: Ausschnitte der Peakcluster 3 bis 6 aus dem FT-ICR-MS-Spektrum von KSH 499. Die signifikant korrelierenden Peaks sind farbig hervorgehoben. Peaks, die im Sinne von Isotopen- und Adduktpeaks miteinander assoziert sind, sind jeweils in der gleichen Farbe dargestellt.

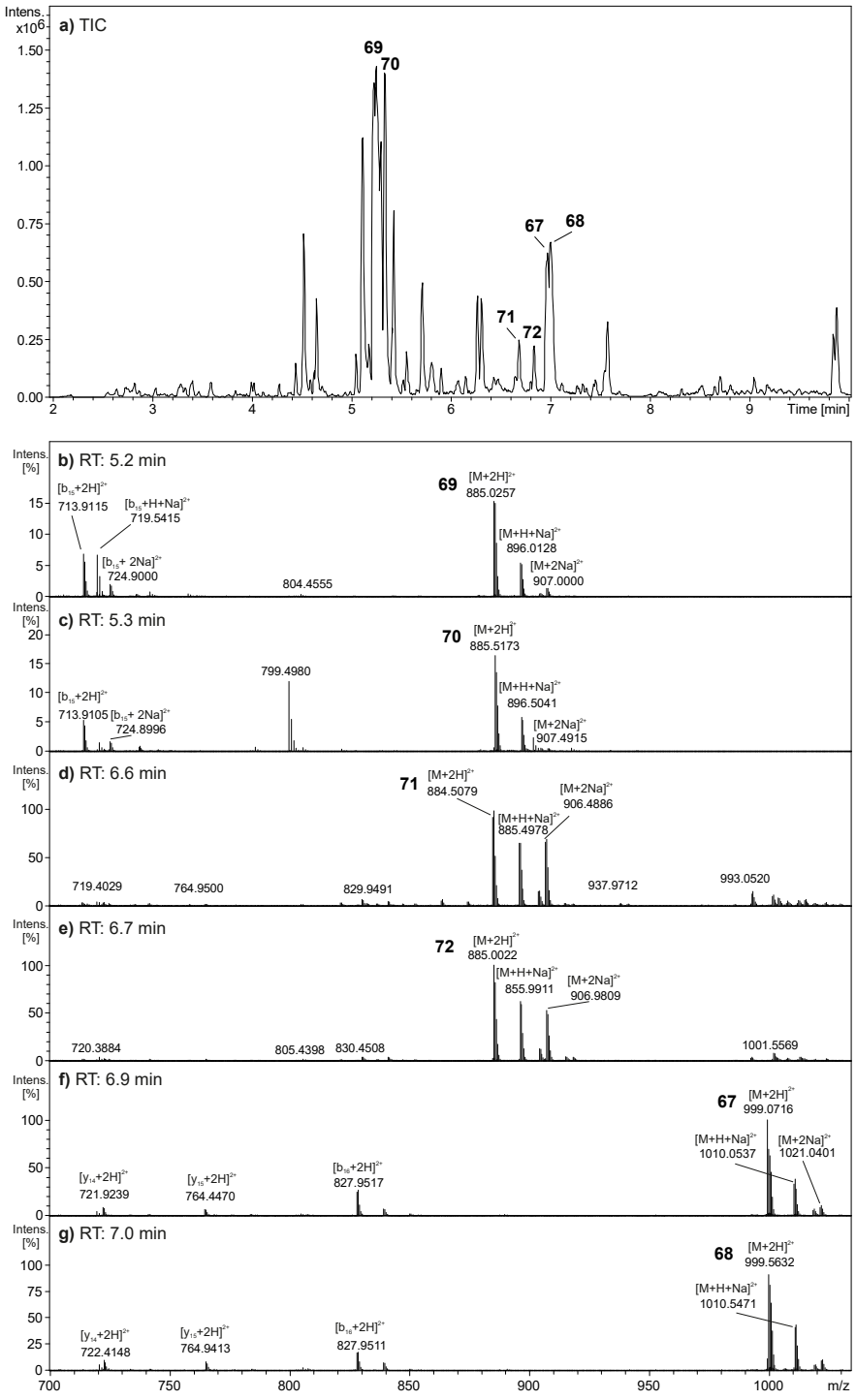


Abbildung 3.32.: Überblick Peakcluster 3, 4 und 5 in der UPLC-QqTOF-MS. a) Totalionenstrom b) Verbindung **69** c) Verbindung **70** d) Verbindung **71** e) Verbindung **72** f) Verbindung **67** g) Verbindung **68**.

Die Überlagerung von Peaksignalen mehrerer Verbindungen wurden durch Messungen mit dem UPLC-QqTOF-MS bestätigt (Abb. 3.32). Die chromatographische Trennung führte zur Separierung von zwei Verbindungen in Peakcluster 3 und vier Verbindungen in Peakcluster 4. Die Verbindungen in Peakcluster 3 eluierten nach 6.9 (**67**) und 7,0 Minuten (**68**) und unterscheiden sich lediglich um 1 Da (Abb. 3.32 f und g).

Die Verbindungen in Peakcluster 4.1 eluierten zwischen Minute 5,2 und 5,3 und bestehen aus den Verbindungen **69** und **70** mit den monoisotopischen Massen $[M+2H]^{2+}$ = 885,0257 und $[M+2H]^{2+}$ = 885,5173 (Abb. 3.32 b und c). Die Verbindungen in Peakcluster 4.2 eluierten zwischen Minute 6,6 und 6,7 und bestehen aus den Verbindungen **71** (*m/z* 884,5083) und **72** (*m/z* 885,0033) (Abb. Abb. 3.32 d und e).
Bei den positiv korrelierenden Peaks in Peakcluster 5 (*m/z* 724,89603, 725,39727, 735,88374) handelt es sich um Adduktpeaks einer Verbindung **73**, dessen $[M+2H]^{2+}$-Peak bei 713,90718 detektiert wurde. Die UPLC-QqTOF-MS Messungen zeigen, dass diese Verbindung die gleiche Retentionszeit besitzt, wie die Peaks in Peakcluster 4.1 (Abb. 3.32 b und c). Tatsächlich zeigen Fragmentierungsstudien (siehe Abschnitt 3.3.3.5, Seite 181), dass es sich bei *m/z* 713,90718 mit hoher Wahrscheinlichkeit um das doppelt-protonierte b_{15}-Ion der Verbindungen **69** und **70** handelt, welches durch Bruch der labilen Aib-Pro Bindung in der Ionenquelle (in-source-decay) entstanden ist.

Korrelationen innerhalb der Peakcluster 3-6 In der Pearson-Korrelationsmatrix der Hitliste 2 (Abb. C.3, Anhang) sind fünf Cluster erkennbar, in denen jeweils hohe Korrelationen zwischen den Peaks auftreten. Nach deren Annotation (Tab. C.2) wird deutlich, dass drei dieser Cluster in hohem Maße Peaks enthalten, die Peptaibolen zugeordnet werden können (Pep1, Pep2, Pep3). Jedes der drei Peptaibolcluster enthält Addukt- oder Fragment Peaks verschiedener Peptaibole, d. h. die Bildung der Cluster ist nicht direkt substanzabhängig. Die drei Cluster enthalten jedoch auch Peaks, die nicht annotiert werden konnten. Die hohe Korrelation zu den Peptaibolpeaks weist jedoch auf einen gemeinsamen biochemischen Ursprung zu Peptaibolen hin. Den drei Peptaibolclustern stehen zwei Cluster (NA1, NA2) gegenüber, bei denen der Ursprung der Peaks nicht näher bekannt ist. Durch Darstellung im Korrelationsnetzwerk wird die Trennung zwischen NA- und Peptaibol Clustern noch besser deutlich. Die Cluster I und II enthalten die Peptaibolcluster Pep2, Pep1 und Pep3. Die NA-Cluster NA1 und NA2 verteilen sich auf Cluster III und IV. Die Berechnung des GGM-Netzwerkes (Abb. C.4, Anhang) ergab keine bessere Feinauflösung der beschriebenen Cluster.

3.3.2.3. Zusammenfassung der Aktivitäts-Korrelations-Analyse

22 methanolische *Sepedonium ampullosporum* Rohextrakte wurden mithilfe von AcorA in Hinblick auf ihre zytotoxische Wirkung auf HT-29 Kolonkarzinomzellen analysiert. In zwei Iterationen wurden zwei Hitlisten generiert, in denen insgesamt 6 Peakcluster besonders auffällig waren, die sowohl einen hohen Rangkorrelationskoeffizienten als auch eine hohe Peakdichte aufweisen. Letzteres ist ein Hinweis auf Isotopen- und Adduktpeaks, die - nach der Erfahrung aus dem Proof of Concept Experiment - auf eine mögliche kausale Korrelation mit der Bioaktivität hinweisen. Die Übersicht in Tabelle 3.4 macht deutlich, dass es aufgrund der fehlenden chromatographischen Trennung und des geringen Massenunterschiedes in den hochauflösenden FT-ICR-MS Messungen zu einer Superposition der Isotopenmuster mehrerer Verbindungen in verschiedenen Peakclustern gekommen ist. Theoretisch wäre es möglich, dass alle aufgeführten Verbindungen eine zytotoxische Wirkung besitzen.
In einem nächsten Schritt wurden die Verbindungen durch Vergleich der massenspektrometrischen Daten mit Naturstoffdatenbanken und gegebenenfalls *de novo* Sequenzierung identifiziert.

Tabelle 3.4.: Überblick über die monoisotopischen Peaks in Peakcluster 1-5. Die angegebenen *mz*-Werte beziehen sich auf die Messungen mit dem UPLC-QqTOF Massenspektrometer.

Peakcluster	Verbindung	rt	$[M+2H]^{2+}$	$[M+H+Na]^{2+}$	$[M+H+K]^{2+}$	$[M+2Na]^{2+}$
1	**61**	7,1	811,9848	823,4833	830,9600	833,9713
1	**62**	7,2	812,4875	823,4761	831,4579	834,4675
2.1	**63**	5,1	697,9515	708,9399	716,9169	719,9241
2.1	**64**	5,2	698,4434	709,4316	717,4110	720,4189
2.2	**65**	7,0	697,4346	708,4224	716,4042	719,4170
2.2	**66**	7,1	697,9200	708,9103	716,8983	719,9021
3	**67**	6,9	999,0712	1010,0558	1018,0326	1021,0446
3	**68**	7,0	999,5646	1010,5485	1018,5245	1021,5329
4.1	**69**	5,2	885,0291	896,0169	903,9972	907,0049
4.1	**70**	5,3	885,5219	896,5090	904,4900	907,4979
4.2	**71**	6,6	884,5083	895,4981	903,4775	906,4900
4.2	**72**	6,7	885,0033	895,9919	903,9707	906,9824
5	**73**	5,1/5,2	713,9121	724,9018	n.d.	735,8855

3.3.3. Strukturaufklärung der signifikant korrelierenden Verbindungen

3.3.3.1. Peakcluster 1

Unter Berücksichtigung der massenspektrometrischen Daten aus den FT-ICR-MS und UPLC-QqTOF-MS Messungen, muss die monoisotopische Masse der gesuchten Verbindung **61** im Bereich zwischen 1621,95 Da und 1621,98 Da liegen. Durch eine entsprechende Datenbanksuche im „Dictionary of Natural Products“ [287] wurde als einziger Treffer das bereits 1997 von Ritzau aus *Sepedonium ampullosporum* isolierte Ampullosporin A erhalten [227]. Um diesen Verdacht zu bestätigen wurde der methanolische Extrakt von KSH 537 chromatographisch getrennt und das Signal m/z 811,9, das der theoretischen Masse des doppelt protonierten Ampullosporin A entspricht, mit zwei verschiedenen Energien (15, 60 eV) im QqTOF-Massenspektrometer fragmentiert (Abbildung 3.33). Die b-Ionen Serie kann durch Überlagerung der Informationen aus den beiden Spektren vom b_1- bis zum b_{14}-Fragment vollständig identifiziert werden. Das C-terminale Fragment ergibt sich aus dem komplementären y_1-Fragment mit der Masse *m/z* 118,1232. Die aus dieser Masse abgeleitete Summenformel von $C_6H_{16}N_1O_1^+$ (berechnet 118,1226, 4,7 ppm) entspricht einem Leucinol oder Isoleucinol. Der für Ampullosporine typische *N*-Terminus aus Acetyl-Tryptophan ist mit dem Peak bei *m/z* 229,0993 ($C_{13}H_{13}N_2O_2^+$, berechnet 229,0972, 9,2 ppm) deutlich erkennbar. Das entsprechende a_1-Fragment (*m/z* 201,1029, $C_{12}H_{13}N_2O_1^+$, berechnet 201,1022, 3,5 ppm), das Immoniumion bei *m/z* 159,0901 ($C_{10}H_{11}N_2^+$, berechnet 159,0917, 10,0 ppm) sowie ein weiteres für Tryptophan typisches Fragment bei *m/z* 130,0656 ($C_9H_8N^+$, berechnet 130,0651, 3,8 ppm) weisen ebenfalls auf die Anwesenheit von Tryptophan hin. Die aus dem Fragmentierungsmuster abgeleitete Primärstruktur entspricht somit der des Ampullosporin A. Da mit dieser Methode nicht ohne Weiteres zwischen L- und D-Aminosäuren sowie zwischen Leucin(ol)und Isoleucin(ol) differenziert werden kann, bleibt an dieser Stelle eine Restunsicherheit bestehen.
Für Verbindung **62** konnte kein Datenbankeintrag gefunden werden. Der *m/z*-Wert von 812,4875 $[M+2H]^{2+}$ ergibt einen monoisotopischen *m/z*-Wert von 1622,96 Da und unterscheidet sich lediglich um 1 Da von **61**. Diese geringe Massendifferenz, sowie die geringe Retentionszeitdifferenz von 0,1 Minute, deuten eine hohe strukturelle Ähnlichkeit der beiden Verbindungen an. In der Tat zeigt die massenspektrometrische Fragmentierung von **62**, dass die b-Ionenserien beider Verbindungen bis zum b_{13}^+-Ion identisch sind (Tab. 3.5 und Tab. 3.6). An Position 14 enhält **62** ein Glutamat (b_{14}^+ = 1506,7258) anstelle eines Glutamins in **61** (b_{14}^+ = 1505,7480). Dieser Unterschied erklärt die Massendifferenz von 1 Da sowie die geringe Retentionszeitdifferenz der beiden Substanzen.

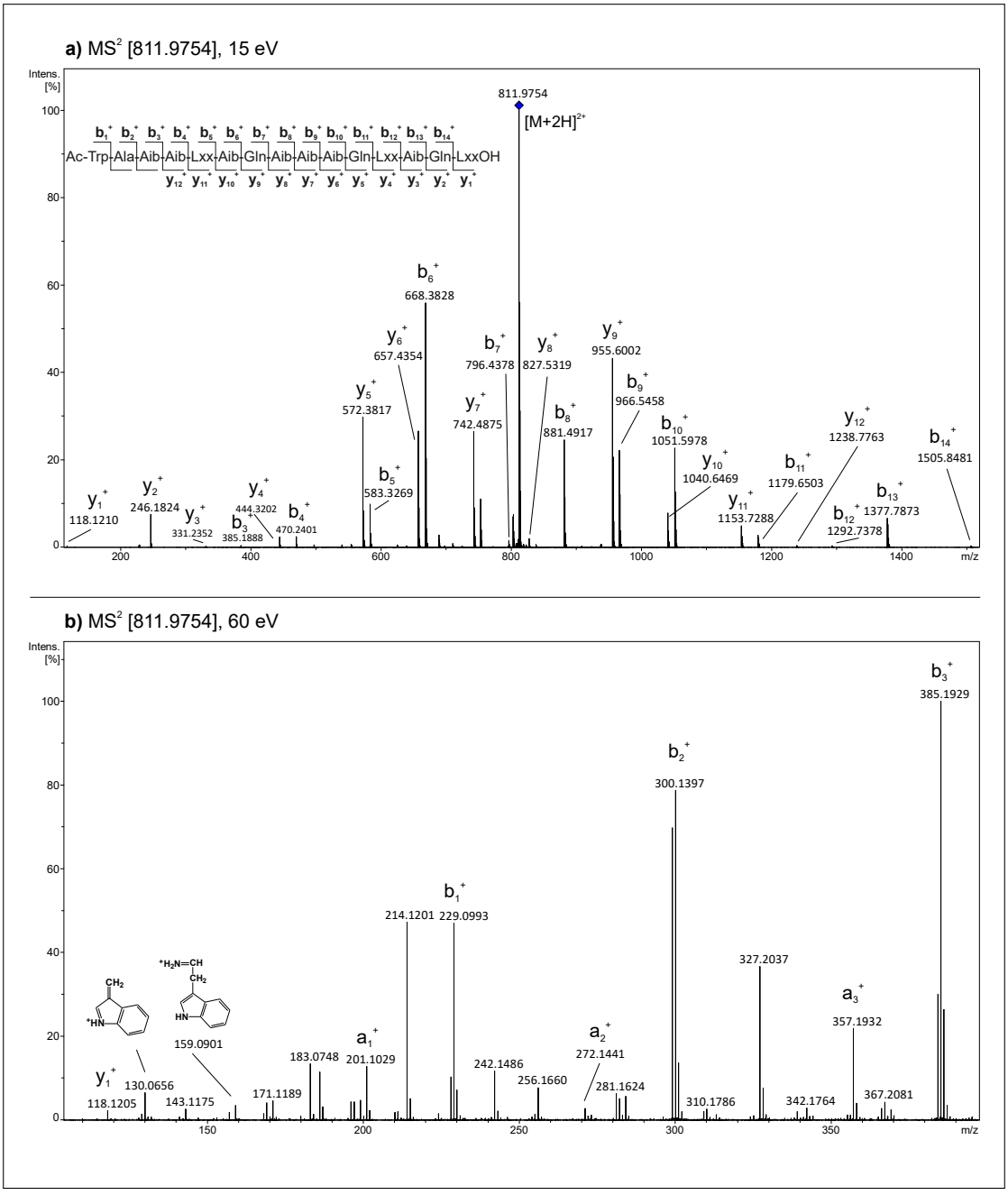


Abbildung 3.33.: Massenspektrometrische Fragmentierung der Verbindung **61** in Peakcluster 1 mit UPLC-QqTOF-MS/MS. a) (+) MS_2 [811,9754], 15 eV b) MS_2 [811,9754], 60 eV.

3.3.3.2. Peakcluster 2

Neben den Peaks aus Peakcluster 1 wird die Hitliste 1 von signifikanten Peaks einiger weiterer Verbindungen dominiert. Die UPLC-QqTOF-MS Messungen haben gezeigt, dass es sich dabei um die Verbindungen **63** - **66** handelt, die in zwei Retentionszeitbereichen (Peakcluster 2.1 und Peakcluster 2.2) von der Säule eluierten. Durch Hochrechnung der *m/z*-Werte ergibt sich, dass diese Verbindungen eine Molekülmasse zwischen 1392,8 Da und 1394,8 Da besitzen müssen. Eine Datenbanksuche blieb erfolglos, aber die relativ großen Massen und die Tatsache, dass die Verbindungen in exakt den selben Extrakten detektiert wurden, wie die Verbindungen **61** und **62**, lassen vermuten, dass es sich ebenfalls um Peptaibole handeln könnte.

Peakcluster 2.1 Die massenspektrometrische Charakterisierung zeigt, dass es sich bei **63** und **64** mit hoher Wahrscheinlichkeit um *N*-terminal trunkierte Varianten von **61** und **62** handelt. Durch Fragmentierung im QqTOF Massenspektrometer konnten die Primärstrukturen von **63** und **64** anhand der b- und y-Ionenserien von Position 3 bis zum C-Terminus eindeutig bestimmt werden (Abb. 3.34, Tabellen 3.5 und 3.6). Das b_2-Ion entspricht mit einer Masse von *m/z* 157,0972 ($C_7H_{13}N_2O_2^+$, 0,3 ppm) einem Ala-Aib oder Aib-Ala Dipeptid. Es wurden weder das b_1- noch das a_1-Ion oder entsprechende Fragmente der oberen y-Ionenserie detektiert, so dass eine Differenzierung zwischen den beiden Varianten nicht ohne Weiteres möglich ist. Jaworski *et al.* haben in ihrer Arbeit gezeigt, dass die Fragmentierung von Peptaibolen im Negativ-Ionen-Modus insbesondere die komplementären y-Fragmente zu den *N*-terminalen b-Ionen liefert [378]. Um zwischen den Dipeptid Varianten differenzieren zu können, wurden daher die $[M-H]^-$ Spezies (*m/z* 1393, bzw. 1394) der beiden Peptaibole im Negativ-Ionen-Modus fragmentiert. Ausgehend von dem jeweiligen $[M-H]^-$-Pseudomolekülionenpeak weisen die konsekutiven Neutralionenverluste von 71 (y_3^- = 1321,82) und 85 (y_{12}^- = 1236,76) eindeutig ein Ala-Aib Dipeptid als *N*-Terminus nach (Abb. 3.34 c).
Die geringe Retentionszeitdifferenz (0,1 min) der beiden Peptaibole beruht auf dem Unterschied von Glutamin (**63**) und Glutamat (**64**) an Position 13. Die übrigen Sequenzpositionen sind bei beiden Peptaibolen identisch und stimmen mit Ausnahme des *N*-terminalen Acetyltryptophans mit **61** und **62** überein. Die freie Aminogruppe am *N*-Terminus von **63** und **64** erklärt die gegenüber **61**/**62** deutlich verringerte Retentionszeit.

Tabelle 3.5.: Putative Aminosäuresequenzen der Peptaibole in Peakcluster 1 und 2.

Verbindung	rt [min]	$[M+2H]^{2+}$	Aminosäure			
			1	2	3	4
61	7,1	811,9946	AcTrp	Ala	Aib	Aib
62	7,2	812,4969	AcTrp	Ala	Aib	Aib
63	5,1	697,9439	Ala	Aib	Aib	Lxx
64	5,2	698,4395	Ala	Aib	Aib	Lxx
65	7,0	697,4312	$C_7H_{10}NO_3^+$	Aib	Lxx	Aib
66	7,1	697,9223	$C_7H_{10}NO_3^+$	Aib	Lxx	Aib

Peakcluster 2.2 Ähnlich den Verbindungen in Peakcluster 2.1, besitzen auch die Verbindungen in Peakcluster 2.2 eine hohe Sequenzhomologie zu **61**/**62**. So sind die Aminosäuresequenzen der Verbindungen **65** und **66** von Position 2 bis 13 identisch mit den Aminosäuresequenzen von **61** und **62** an Position 4 bis 15 (Tab. 3.5). Die Sequenzen lassen sich zweifelsfrei aus den komplementären Fragmenten der b- und y-Ionenserie der UPLC-QqTOF-MS Messungen ableiten (Abb. 3.35 a, Tab. 3.6).

Beide Verbindungen enthalten jedoch einen ungewöhnlichen *N*-Terminus, für den lediglich eine Summenformel von $C_7H_{10}NO_3^+$ mit 3,5 Doppelbindungsäquivalenten entsprechend dem b_1-Ion bei *m/z* 156,0640 (berechnet für $C_7H_{10}NO_3^+$ 156,06552, 9,7 ppm) angegeben werden kann. Um diesen eindeutig bestimmen zu können, wurden zunächst die Mutterionen durch Anlegen einer hohen Quellenspannung in der Ionenquelle fragmentiert. Durch Selektion des relevanten Fragmentions (hier: *m/z* 354) im Quadrupol Q1 und anschließender CID-Fragmentierung in der Kollisionszelle, konnten - ähnlich einer 3D-Ionenfalle - mit dem UPLC-QqTOF-MS weitere Strukturinformationen gewonnen werden. Abbildung 3.35 b zeigt das resultierende s. g. Pseudo-MS^3 Spektrum von *m/z* 354,2 der Verbindung **65**. Im unteren Massenbereich sind die Immonium-Ionen von Aib (*m/z* 58,0651, berechnet für $C_3H_8N^+$ 58,06513, 0,5 ppm) und Lxx (*m/z* 86,0950, berechnet für $C_5H_{12}N^+$ 86,09643, 16,8 ppm) deutlich erkennbar. Das *m/z* 354,2 Fragment sollte folglich mindestens ein Aib und ein Lxx enthalten. Die Abspaltung von Lxx^3 ist mit dem Ion bei *m/z* 241,1177 erkennbar, das entsprechende a_2-Ion ist bei *m/z* 213,1233 zu sehen. Die weitere Abspaltung von Aib kann mit dem b_1-Ion bei *m/z* 156,0640 sowie dem a_1-Ion bei *m/z* 128,0693 beobachtet werden. Insgesamt erhält man somit für das b_3-Ion *m/z* 354,2 eine Sequenzabfolge von $C_7H_{10}NO_3{}^{+1}$-Aib^2-Lxx^3.

Eine *N*-terminale Sequenz $AcAib^1$-Lxx^2-Lxx^3 entsprechend einer b-Ionenserie 128,0693 -

Tabelle 3.5.: (Fortsetzung)

5	6	7	8	9	10	11	12	13	14	15
Lxx	Aib	Gln	Aib	Aib	Aib	Gln	Lxx	Aib	Gln	LxxOH
Lxx	Aib	Gln	Aib	Aib	Aib	Gln	Lxx	Aib	Glu	LxxOH
Aib	Gln	Aib	Aib	Aib	Gln	Lxx	Aib	Gln	LxxOH	
Aib	Gln	Aib	Aib	Aib	Gln	Lxx	Aib	Glu	LxxOH	
Gln	Aib	Aib	Aib	Gln	Lxx	Aib	Gln	LxxOH		
Gln	Aib	Aib	Aib	Gln	Lxx	Aib	Glu	LxxOH		

241,1177 - 354,2 wäre zwar ebenfalls denkbar. Allerdings liegt die Masse eines theoretischen deacetylierten Aib-Ions bei 86,06004 Da (berechnet für $C_4H_8NO^+$) und läge mit einer Massenabweichung von 312 ppm unter der beobachteten Masse von *m/z* 86,0950. Zudem wäre das Signal bei *m/z* 156,0640 dadurch nicht erklärt.
Messungen mit der Ionenfalle im Negativ-Ionen-Modus zeigen eindeutig einen Neutralionenverlust von 155 Da für das *N*-terminale Peptid (Abb. 3.35 c). Von dem y_{12}^--Ion bei *m/z* 1236,81 kann die Sequenz bis zum y_4^--Ion identifiziert werden und bestätigt nochmals die mit dem UPLC-QqTOF Massenspektrometer erhaltenen b-Ionenserie (vom b_1^+ bis zum b_9^+-Fragment).
Auffällig ist im Negativ-Ionen-Modus eine Abspaltung von 44 Da, die zu einem stabilen Fragment bei *m/z* 1347,86 führt (Abb. 3.35 c). Aufgrund der hohen Messabweichung des Ionenfallen Massenspektrometers kann nicht zwischen der Abspaltung einer Carboxylgruppe (CO_2, *m/z* 43,9898) oder eines Acetaldehyds (C_2H_4O, *m/z* 44,0262) differenziert werden. Der Neutralionenverlust von 44 Da ist im MS/MS Spektrum des Negativ-Ionen-Modus der Verbindung **63**, die, mit der Ausnahme des *N*-Terminus, eine identische Primärstruktur besitzt, nicht zu beobachten (Abb. 3.34 c). Dieses Fragment scheint somit ein Strukturelement des *N*-Terminus der Verbindungen **65** und **66** darstellen.

Tabelle 3.6.: Diagnostische Fragmentionen [*m/z*] der Peptaibole in Peakcluster 1 und 2.

Verbindung	**61**	**62**	**63**	**64**	**65**	**66**
rt [min]	7.1	7.2	5.1	5.2	7.0	7.1
$[M+2H]^{2+}$	811.9946	812.4969	697.9439	698.4395	697.4312	697.9223
Immoniumion Trp	159.0892	159.0892	-	-	-	-
a1	201.1009	201.0996	n.d.	n.d.	128.0702	128.0690
a2	272.1416	272.1420	129.1024	129.1017	213.1241	213.1225
b1	229.0955	229.0963	n.d.	n.d.	156.0659	156.0660
b2	300.1400	300.1361	157.0982	157.0969	241.1163	241.1159
b3	385.1915	385.1884	242.1490	242.1482	354.2041	354.2028
b4	470.2426	470.2384	355.2326	355.2319	439.2565	439.2567
b5	583.3310	583.3236	440.2845	440.2850	567.3163	567.3149
b6	668.3825	668.3779	568.3438	568.3443	652.3688	652.2665
b7	796.4286	796.4293	653.3983	653.3973	737.4190	737.4130
b8	881.4812	881.4815	738.4470	738.4480	822.4673	822.4644
b9	966.5241	966.5238	823.4968	823.4982	950.5122	950.5212
b10	1051.5683	1051.5673	951.5425	951.5440	1063.5778	1063.5927
b11	1179.5993	1179.6037	1064.6118	1064.6192	1148.6252	1148.6261
b12	1292.6664	1292.6723	1149.6591	1149.6572	1276.6543	1277.6485
b13	1377.7075	1377.7097	1277.6925	1278.6871	n.d.	n.d.
b14	1505.7480	1506.7258	n.d.	n.d.	-	-
y1	118.1232	118.1229	118.1228	118.1235	118.1234	118.1225
y2	246.1803	247.1623	246.1808	247.1614	246.1802	247.1621
y3	331.2319	332.2162	331.2340	332.2173	331.2344	332.2209
y4	444.3168	445.3020	444.3172	445.3016	444.3168	445.2987
y5	572.3803	573.3643	572.3777	573.3619	572.3799	573.3610
y6	657.4335	658.4137	657.4303	658.4147	657.4321	658.4132
y7	742.4828	743.4640	742.4775	743.4641	742.4800	743.4643
y8	827.5277	828.5146	827.5224	828.5085	827.5253	828.5110
y9	955.5812	956.5652	955.5732	956.5883	955.5770	956.5970
y10	1040.6153	1041.6048	1040.6172	1041.5996	1040.6177	1041.6015
y11	1153.6835	1154.6764	1153.6785	1154.6710	1153.6820	1154.6465
y12	1238.7206	1239.7102				
$[M-H]^-$	-	-	1393	1394	1392	1393
$[M-H-H_2O]^-$	-	-	-	-	1374	1375
$[M-H-44]^-$	-	-	-	-	1348	1349
$y13^-$	-	-	1322	1323	-	-
$y12^-$	-	-	1237	1238	1237	1238
$y11^-$	-	-	1152	1153	1152	1153
$y10^-$	-	-	1039	1040	1039	1040
$y9^-$	-	-	954	955	954	955

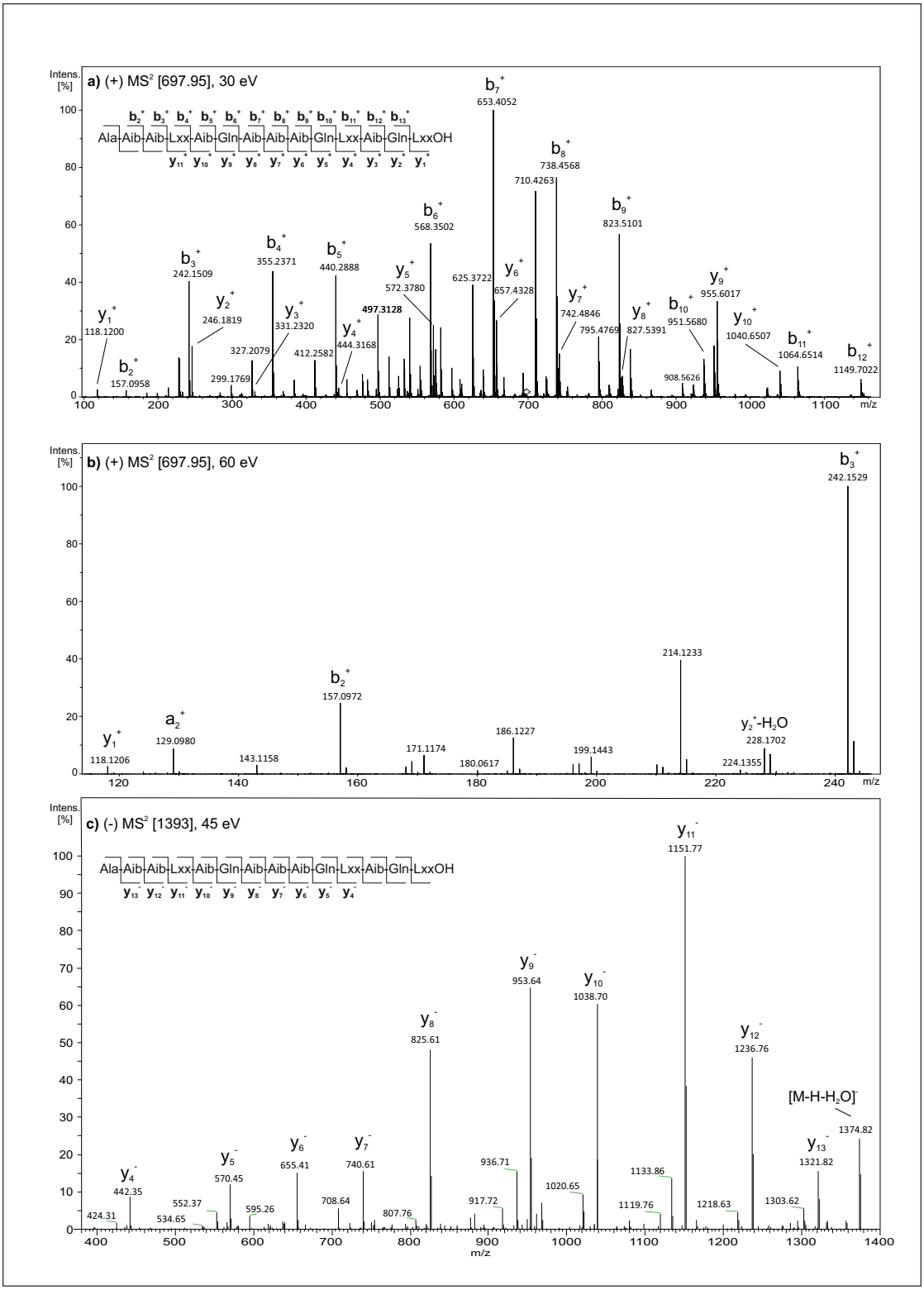


Abbildung 3.34.: Massenspektrometrische Fragmentierung von Verbindung **63** in Peakcluster 2.1 mit UPLC-QqTOF-MS/MS (a-b) und UPLC-IT-MS^n (c). a) (+) MS^2 [697,95], 30 eV b) (+) MS^2 [697,95], 60 eV c) (-) MS^2 [1393], 45 eV.

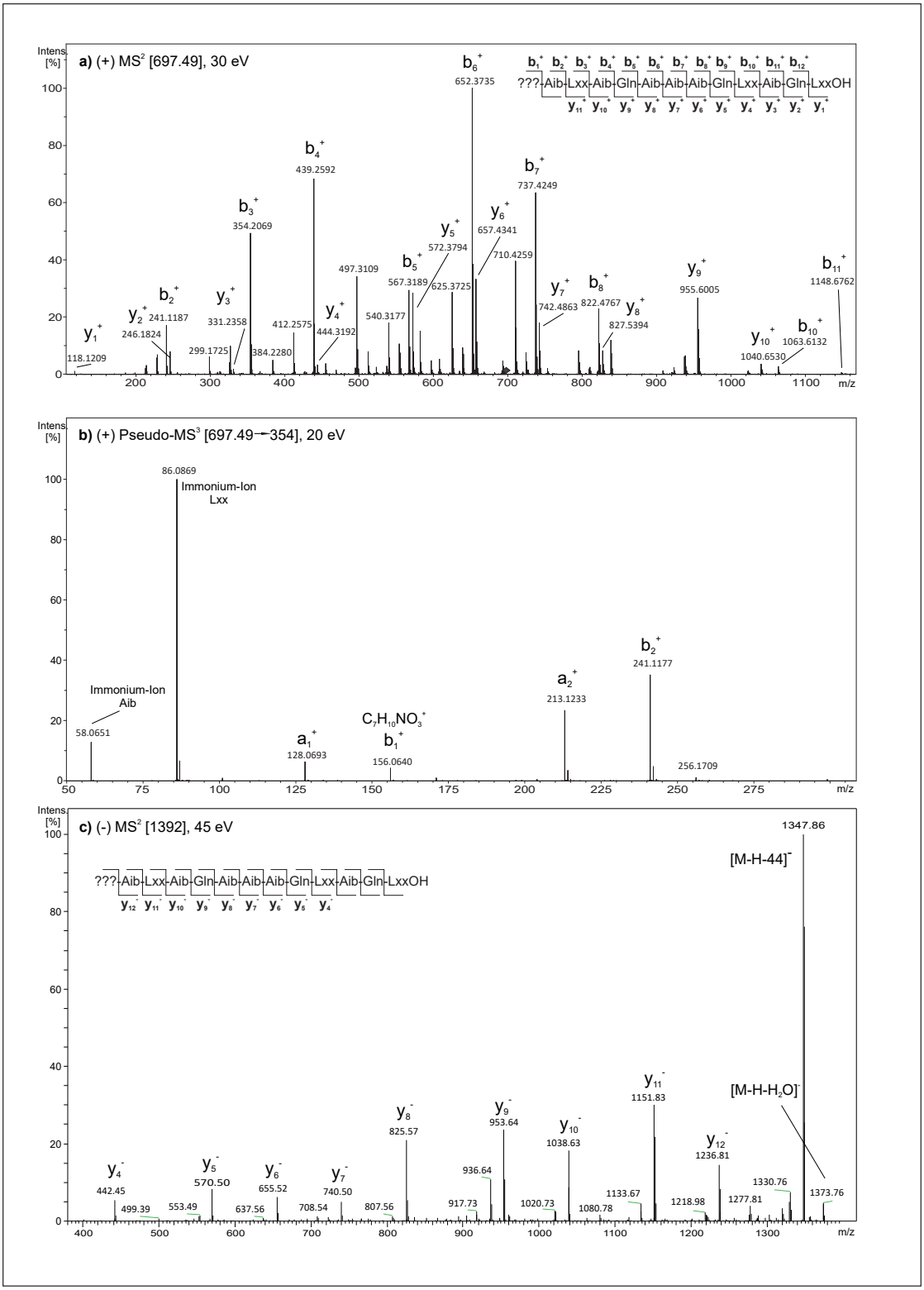


Abbildung 3.35.: Massenspektrometrische Fragmentierung von Verbindung **65** in Peakcluster 2.2 mit UPLC-QqTOF-MS/MS (a-b) und UPLC-IT-MS^n (c). a) (+) MS^2 [697,49], 30 eV b) (+) MS^2 [697,49 $\mapsto$ 354], 20 eV c) (-) MS^2 [1392], 45 eV.

3.3.3.3. Peakcluster 3

Peakcluster 3 besteht aus den Verbindungen **67** und **68**. Eine Literatursuche blieb ergebnislos. Die Verbindungen wurden daher im methanolischen Extrakt von KSH 549 mit dem UPLC-QqTOF Massenspektrometer mit drei verschiedenen Kollisionsenergien fragmentiert. Die Fragmentierung von **67** zeigt bereits bei geringer Fragmentierungsenergie (15 eV) ein abundantes Signal bei *m/z* 343,2345 (52 % relative Signalintensität, Abb. 3.36 a). Das komplementäre Ion bei *m/z* 1654,7362 besitzt mit 15 % relativer Signalintensität eine deutlich geringere Intensität. Dieses Fragmentierungsverhalten ist typisch für Prolin-haltige Peptaibole und deutet darauf hin, dass es sich bei *m/z* 343,2345 um ein Fragment der y-Serie mit einem terminalen Prolin handelt, da nach Bruch der labilen Aib-Pro Bindung die Ladung überwiegend auf dem Prolin enthaltenden y-Fragment verbleibt [379]. Ausgehend von dem b_{16}-Ion bei *m/z* 1654,7362 kann die b-Ionenserie bis zum b_2-Ion (*m/z* 300,1354) zurückverfolgt werden (Abb. 3.36 b und c). Das *N*-terminale Acetyltryptophan ist jedoch nur indirekt über das a_1-Ion (*m/z* 201,1021; berechnet für $C_{12}H_{13}N_2O^+$ 201,10224, 0,7 ppm), sowie über das deacetylierte b_1-Fragment (*m/z* 187,0860; berechnet für $C_{11}H_{11}N_2O^+$ 187,08659, 3,2 ppm) und das Immonium-Ion von Tryptophan (*m/z* 159,0923; berechnet für $C_{10}H_{11}N_2{}^+$ 159,09167, 3,9 ppm) nachweisbar. Zudem wurde mit *m/z* 130,0646 (berechnet für $C_9H_8N^+$ 130,06513 4,1 ppm) ein weiteres für Tryptophan typisches Fragment detektiert. Die Präsenz von Tyrosin im Molekül wird durch das entsprechende Immoniumion bei *m/z* 136,0736 (berechnet für $C_8H_{10}NO^+$ 136,07569, 15 ppm) bestätigt.
Messungen mit der Ionenfalle sind aufgrund der fehlenden y-Ionen übersichtlicher und bestätigen die Sequenz vom b_2- bis zum b_{16}-Ion (Abb. 3.36 d und e).
Da die b-Ionenserie oberhalb des b_{16}-Fragments abbricht, wurden die verbliebenen C-terminalen Aminosäuren über die y-Ionenserie bestimmt. Das y_1-Fragment *m/z* 118,1238 (berechnet für $C_6H_{16}N_1O_1{}^+$ 118,12264, 9,8 ppm) konnte als Leucinol/Isoleucinol identifiziert werden. Das C-terminale LxxOH wird durch die Signale bei *m/z* 325,2215 (AS19 -H_2O) und *m/z* 226,1173 (AS19-18), d. h. durch Abspaltung von H_2O und LxxOH vom y_3-Fragment bestätigt. Die Masse von *m/z* 226,1173 entspricht einem Dipeptid aus Pro-Gln (berechnet für $C_{10}H_{16}N_3O_3{}^+$, 226,11862, 5,8 ppm), sodass für das C-terminale Tripeptid ein Pro-Gln-LxxOH postuliert werden kann. Tatsächlich sind auch das y_2-Ion bei *m/z* 246,1785 (berechnet für $C_{11}H_{24}N_3O_3{}^+$ 246,18122, 11 ppm) und das y_3 bei *m/z* 343,2343 (berechnet für $C_{16}H_{31}N_4O_4{}^+$ 343,23398, 0,9 ppm) erkennbar und bestätigen somit die C-terminale Sequenz von Pro^{17}-Gln^{18}-$LxxOH^{19}$ in **67**.

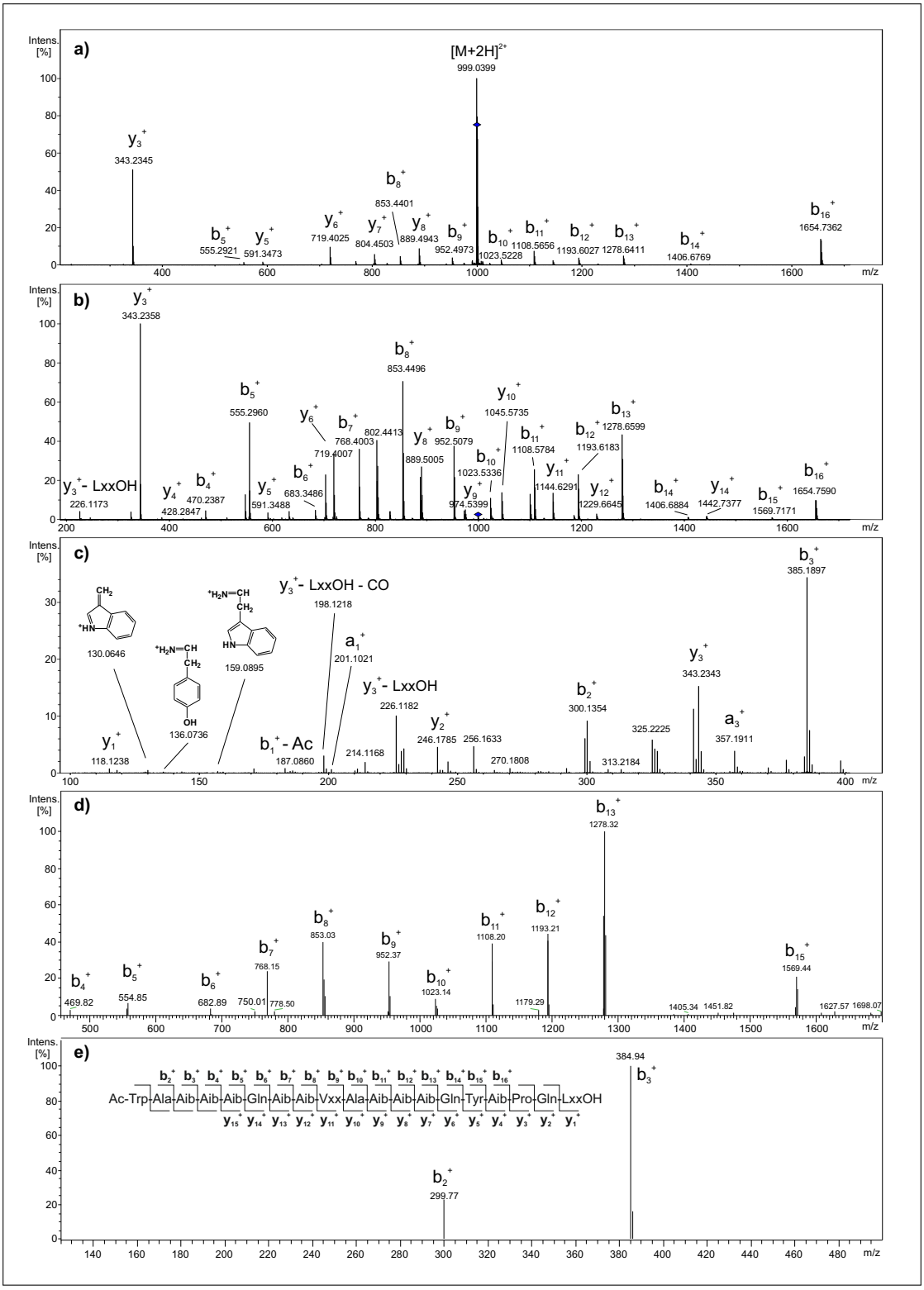


Abbildung 3.36.: Massenspektrometrische Fragmentierung der Verbindung **67** in Peakcluster 3 mit UPLC-QqTOF-MS/MS (a-c) und UPLC-IT-MS^n (d, e). a) (+) MS^2 [999.05], 15 eV b) (+) MS^2 [999.05], 30 eV c) (+) [999.05], 60 eV d) (+) MS^3 [999 $\mapsto$ 1654] e) (+) MS^5 [999 $\mapsto$ 1654 $\mapsto$ 768 $\mapsto$ 470] .

Hinweise für ein isobares Ala-Gly Dipeptid anstelle von Gln^{18} wurden nicht gefunden.
Sowohl die geringe Retentionszeitdifferenz (0,1 min), als auch die Massendifferenz von 1 Da zwischen Verbindung **67** und **68** lassen vermuten, dass es sich um strukturell verwandte Verbindungen handeln muss.
Die b-Ionenserie ist bei beiden Verbindungen bis zum b_{16}-Ion (*m/z* 1654,7350) identisch. Unterschiede zeigen sich erst bei dem C-terminalen Tripeptid, dessen abundantes y_3-Ion von Verbindung **68** bei *m/z* 344,2204 (37 % rel. Intensität) detektiert wurde (Tab. 3.8). Eine Massendifferenz von 0,98 Da zum y_3-Ion von Verbindung **67** entspräche einer Substitution von Glutamin durch Glutamat an Sequenzposition 18. Tatsächlich entspricht die Massendifferenz zwischen dem C-terminalen y_1-Ion (118,1230; $C_6H_{16}N_1O_1^+$, LxxOH) und dem y_2-Ion (247,1613) der Inkrementmasse des Glutamats. Die y-Ionenserie ist vom y_1-Fragment bis zum y_{12} durchgängig erkennbar und durch die Substitution an Position 18 ab dem y_2-Ion um jeweils eine Nominalmasse gegenüber der y-Ionenserie von **67** versetzt (Tab. 3.8).

3.3.3.4. Peakcluster 4

Peakcluster 4 besteht aus vier Verbindungen, die sich aufgrund der Retentionszeitunterschiede in zwei Gruppen mit je zwei Verbindungen einteilen lassen. Die Substanzen konnten weder durch Literaturrecherche noch durch Suche in der „Comprehensive Peptaibol Database“ [380] identifiziert werden. Sie wurden daher nach Trennung mittels UPLC mit dem QqTOF Massenspektrometer fragmentiert und *de novo* sequenziert.

Peakcluster 4.1 Die Verbindungen in Peakcluster 4.1 sind in Bezug auf ihre Primärstruktur mit der Ausnahme der *N*-Termini identisch mit den Verbindungen in Peakcluster 3. Die 15 eV Fragmentspektren von **69** und **70** weisen Ähnlichkeiten zu den 15 eV Fragmentspektren von **67** und **68** auf. So erkennt man prominente Signale bei 343,2336 und 1426,7066 (**69**, Abb. 3.37 a und b) respektive 344,2191 und 1426,7189 (**70**, Tab. 3.8), die wieder auf Prolin in der Nähe der C-Termini schließen lassen. Die beiden Fragmente sind im Rahmen der Messgenauigkeit identisch mit den y_3-Fragmenten von **67** und **68**. Die detektierten Massen der y_1- (*m/z* 118,1224) und y_2-Ionen (*m/z* 246,1798 **69**, *m/z* 247,1638 **70**) weisen, wie bereits bei der Verbindungen **67** und **68** in Peakcluster 3, C-Termini mit den Sequenzen Pro-Gln-LxxOH (**69**) und Pro-Glu-LxxOH (**70**) nach. Die Peaks der y-Ionenserien weisen in beiden Verbindungen zumeist nur geringe Intensitäten auf, sind jedoch vom y_1 bis zum y_{12} Fragment erkennbar (Tab. 3.8). Demgegenüber sind die Peaks der b-Ionenserien deutlich

Tabelle 3.7.: Putative Aminosäuresequenzen der Peptaibole in Peakcluster 3 und 4

Cluster	Verbindung	rt [min]	$[M+2H]^{2+}$	Aminosäure					
				1	2	3	4	5	6
3	**67**	6,9	999,0712	AcTrp	Ala	Aib	Aib	Aib	Gln
3	**68**	7,0	999,5646	AcTrp	Ala	Aib	Aib	Aib	Gln
4.1	**69**	5,2	885,0291	Ala	Aib	Aib	Aib	Gln	Aib
4.1	**70**	5,3	885,5219	Ala	Aib	Aib	Aib	Gln	Aib
4.2	**71**	6,6	884,5083	$C_7H_{10}NO_3^+$	Aib	Aib	Gln	Aib	Aib
4.2	**72**	6,7	885,0033	$C_7H_{10}NO_3^+$	Aib	Aib	Gln	Aib	Aib

stärker ausgeprägt und vom b_2- bis zum b_{15}-Fragment in den 30 und 45 eV Fragmentspektren verfolgbar (Abb.3.37 a und b, Tab. 3.8). Das jeweilige b_2-Ion entspricht mit einer Masse von 157,0969 (Verbindung **69**) und 157,0959 (Verbindung **70**) einem Dipeptid aus Ala-Aib oder Aib-Ala. Das Pseudomolekülion weist im Negativ-Ionen-Modus sukzessive Neutralionenverluste von 71 (y_{17}^-, *m/z* 1695,76) und 85 Da (y_{16}^-, *m/z* 1610,96) auf, was eindeutig auf einen Ala-Aib als *N*-Terminus zurückzuführen ist. Von diesem Ion aus lässt sich die y-Ionenserie im Negativ-Ionen-Modus bis zum y_5^- Fragment weiterverfolgen (Abb. 3.37 d)

Insgesamt betrachtet, scheint es sich bei den Verbindungen **69** und **70** um *N*-terminal trunkierte Varianten der Verbindungen **67** und **68** zu handeln. Die Aminosäuresequenzen sind bis auf den terminalen Acetyltryptophan Rest identisch. Die freie Aminogruppe am *N*-Terminus erklärt die gegenüber **67** und **68** deutlich verringerte Retentionszeit.

Peakcluster 4.2 Die Verbindungen in Peakcluster 4.2 sind deutlich unpolarer als die Verbindungen in Peakcluster 4.1. Die Ähnlichkeit der detektierten Massen in den Übersichts- und MS/MS-Spektren weist jedoch auf strukturell verwandte Verbindungen hin.

Analog zu den Verbindungen in Peakcluster 3 und 4.1, sind in den 15 eV Spektren von **71** und **72** Massensignale bei *m/z* 343,2324 (**71**) und *m/z* 344,2196 (**72**) erkennbar (Abb. 3.38, Tab. 3.8). Wie bereits zuvor weisen die y_1-, y_2- und y_3-Ionen eine C-terminale Sequenz aus Pro-Gln-LxxOH (**71**) bzw. Pro-Glu-LxxOH (**72**) nach. Die y-Ionenserien wurden durchgängig von y_1 bis y_{12} detektiert und sind im Rahmen der Messgenauigkeit identisch mit den y-Ionenserien der Verbindungen in Peakcluster 3 und 4.1. Die Massenspektren werden jedoch von der b-Ionenserie dominiert, die von b_1 bis b_{14} identifiziert werden konn-

Tabelle 3.7.: (Fortsetzung)

7	8	9	10	11	12	13	14	15	16	17	18	19
Aib	Aib	Vxx	Ala	Aib	Aib	Aib	Gln	Tyr	Aib	Pro	Gln	LxxOH
Aib	Aib	Vxx	Ala	Aib	Aib	Aib	Gln	Tyr	Aib	Pro	Glu	LxxOH
Aib	Vxx	Ala	Aib	Aib	Aib	Gln	Tyr	Aib	Pro	Gln	LxxOH	
Aib	Vxx	Ala	Aib	Aib	Aib	Gln	Tyr	Aib	Pro	Glu	LxxOH	
Vxx	Ala	Aib	Aib	Aib	Gln	Tyr	Aib	Pro	Gln	LxxOH		
Vxx	Ala	Aib	Aib	Aib	Gln	Tyr	Aib	Pro	Glu	LxxOH		

te. Anhand der ermittelten Ionenserien wird deutlich, dass die Aminosäuresequenz von **71** ab Position 2 identisch ist mit den Aminosäuresequenzen von **67** und **69** ab Position 3 bzw. 4. Äquivalentes gilt für die Sequenzabfolge in den Verbindungen **72**, **68** und **70** (vgl. Tab. 3.7). D. h. die strukturellen Unterschiede zwischen den Verbindungen der beschriebenen Peakcluster müssen am *N*-Terminus lokalisiert sein. Während die Verbindungen in Peakcluster 3 ein für *S. ampullosporum* typisches Acetyltryptophan aufweisen, beginnen die Verbindungen in Peakcluster 4.1 mit einem Ala-Aib Motiv. Bei den Verbindungen in Peakcluster 4.2 wiederum entspricht das b_1-Ion mit einer Masse von *m/z* 156,0618 einer Summenformel von $C_7H_{10}N_1O_3^+$ (berechnet 156,06552, 23 ppm) und lässt somit eine strukturelle Identität zum *N*-Terminus der Verbindungen in Peakcluster 2.2 vermuten. Die Messungen im Negativ-Ionen-Modus bestätigen diese Annahme (Abb. 3.38 d). Ausgehend vom Pseudomolekülion $[M\text{-}H]^-$ = 1766 wurde mit dem Peak bei *m/z* 1611,0 (y_{16}^-) ein Neutralionenverlust von 155 Da detektiert, der, wie bereits bei **65** und **66**, *m/z* 156,0618 als b_1^+-Ion bestätigt. Analog zu **65** und **66** ist eine Abspaltung von 44 Da, d. h. einer Carboxylgruppe oder eines Acetaldehyds, zu beobachten, die zu dem großen Signal bei *m/z* 1722 führt. Dieser Neutralionenverlust ist in den Spektren der Verbindungen in Peakcluster 4.1 nicht zu beobachten. Da sich die Primärstrukturen der Verbindungen in Peakcluster 4.1 und 4.2 ausschließlich in der ersten Aminosäure unterscheiden, entspricht der Neutralionenverlust von 44 Da vermutlich einem Strukturelement der *N*-Termini von **71** und **72** in Peakcluster 4.2. Der Retentionszeitunterschied (0,1 min) dieser beiden Verbindungen erklärt sich durch die leicht verändete Polarität aufgrund der Präsenz von Gln respektive Glu an Position 16.

3.3.3.5. Peakcluster 5

Die in Peakcluster 5 signifikant korrelierenden Peaks besitzen Massen von *m/z* 724,89603 ($[M+H+Na]^{2+}$), *m/z* 725,39727 ($[M+H+Na]^{2+}$, 1.Isotopenpeak) und *m/z* 735,88374 ($[M+2Na]^{2+}$). Sie gehören zu einem Peakcluster aus doppelt-geladenen Ionen, dessen $[M+2H]^{2+}$-Ion bei *m/z* 713,90718 detektiert wurde (**73**, Abb. 3.31). Für das entsprechende $[M+H]^{+}$-Ion ergibt sich daraus eine Masse von *m/z* 1426,8139. Diese Masse entspricht mit einer Massenabweichung von 6 ppm der theoretischen Masse (1426,80528) des b_{15}-Fragmentes der Verbindungen **69** und **70**. Das b_{15}-Fragment entstand - wie zuvor diskutiert - durch Bruch der labilen Aib-Pro Bindung aufgrund von Stoßaktivierung in der Ionenquelle (ISCID). Die UPLC-QqTOF-MS Messungen zeigen, dass **73** die gleiche Retentionszeit besitzt wie **69** und **70** (Abb. 3.32). Weiterhin wird in den MS/MS Messungen von **69** und **70** neben dem $[M+H]^{+}$-Ion auch das $[M+2H]^{2+}$-Ion bei *m/z* 713,9037 des b_{15}-Fragments beobachtet. Man kann also davon ausgehen, dass es sich bei Verbindung **73** mit hoher Wahrscheinlichkeit um das doppelt geladene b_{15}-Fragment der Verbindungen **69** und **70** handelt und aufgrund der labilen Aib-Pro Bindung durch Stoßaktivierung in der Ionenquelle entstanden ist.

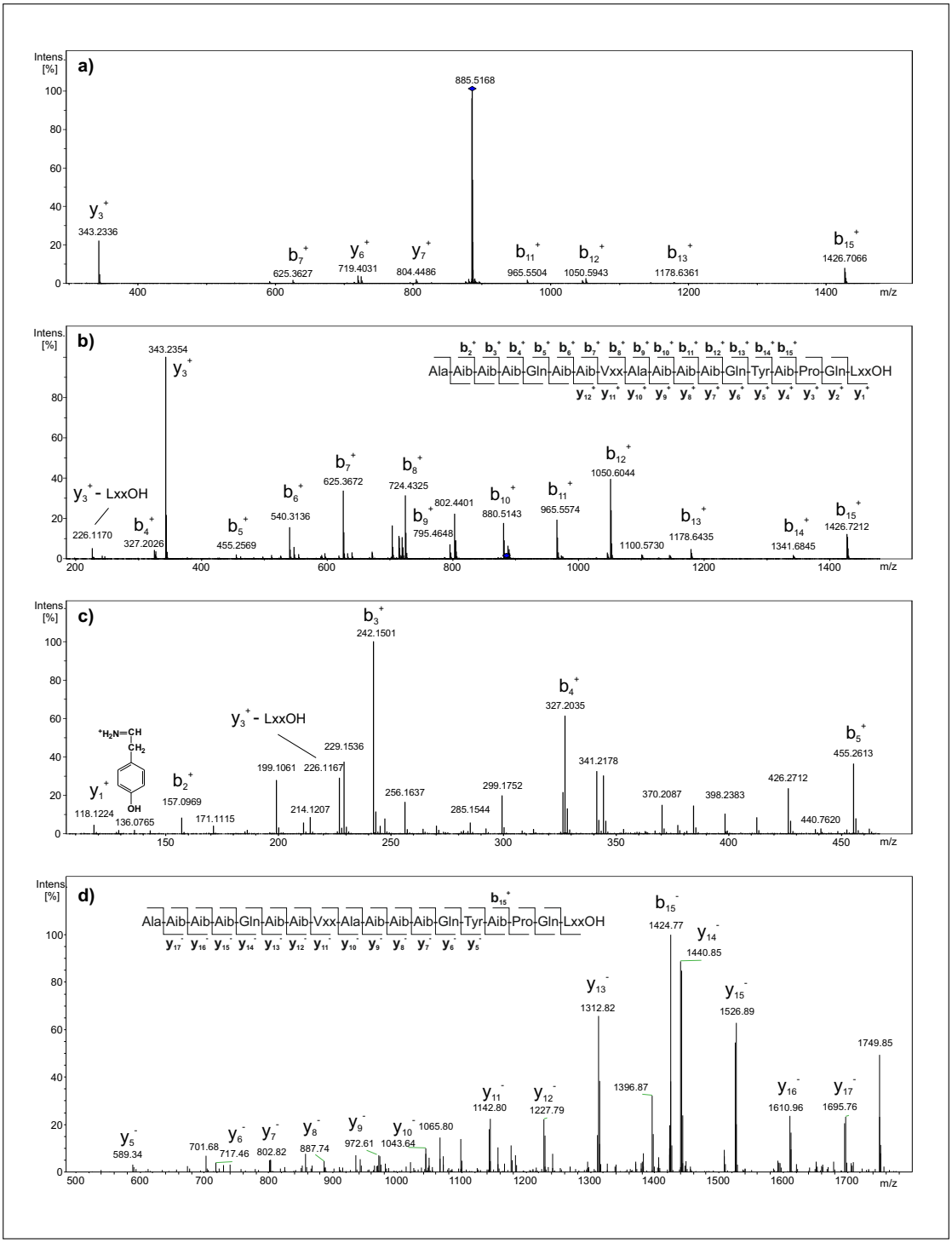


Abbildung 3.37.: Massenspektrometrische Fragmentierung der Verbindung **69** in Peakcluster 4.1 mit UPLC-QqTOF-MS/MS (a-c) und UPLC-IT-MS^n (d). a) (+) MS^2 [885.01], 15 eV b) (+) MS^2 [885.01], 30 eV c) (+) [885.01], 60 eV d) (-) MS^2 [1767].

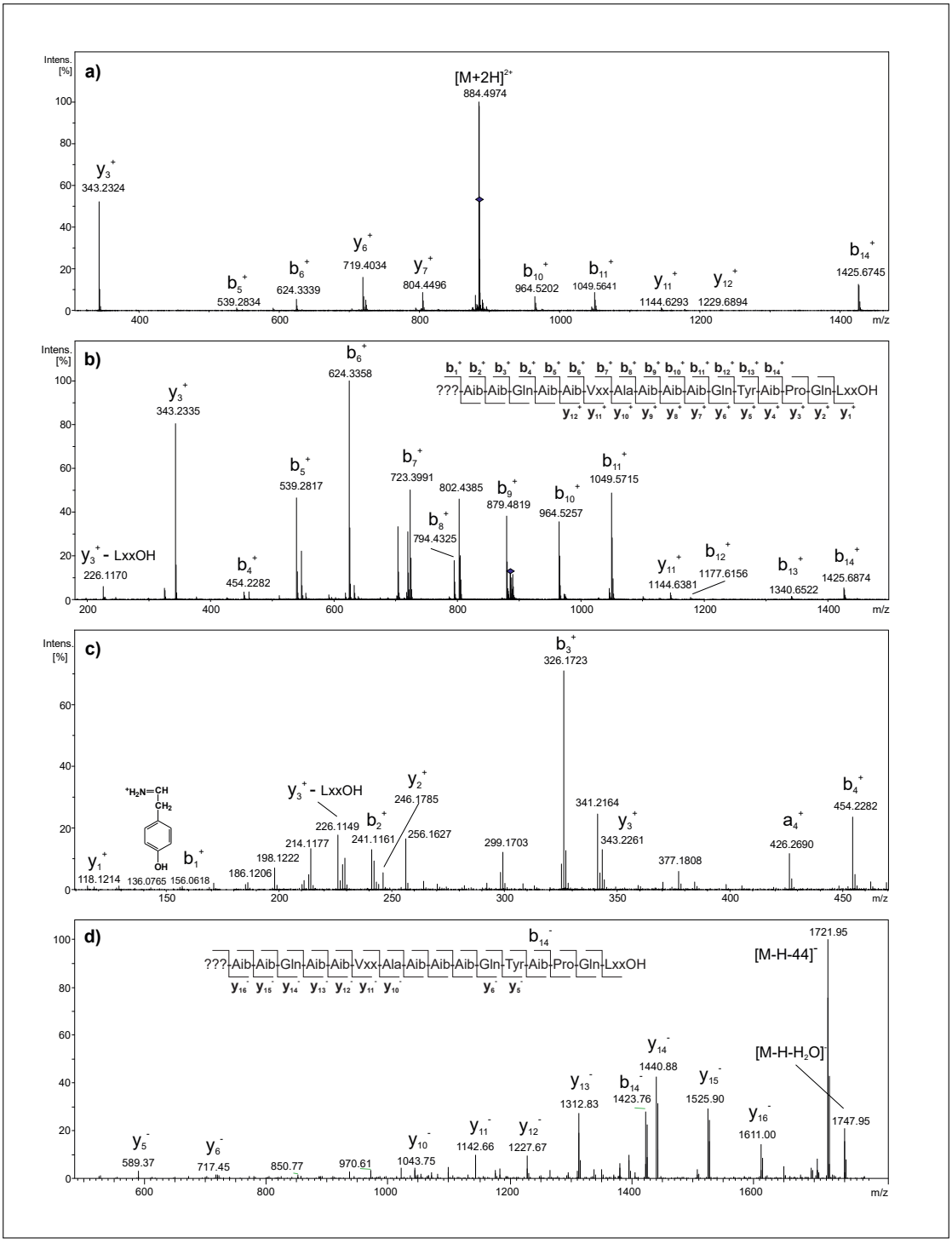
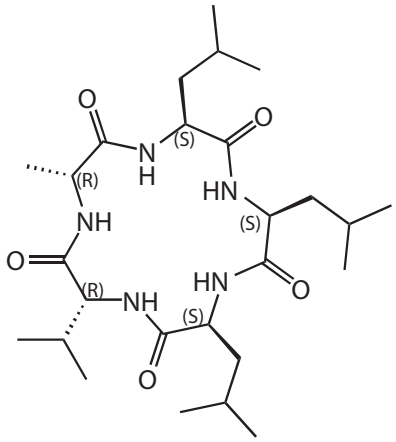


Abbildung 3.38.: Massenspektrometrische Fragmentierung der Verbindung **71** in Peakcluster 4.2 mit UPLC-QqTOF-MS/MS (a-c) und UPLC-IT-MS^n (d). a) (+) MS^2 [884.01], 15 eV b) (+) MS^2 [884.01], 30 eV c) (+) [884.01], 60 eV d) (-) MS^2 [1766].

3.3.3.6. Peakcluster 6

Die signifikant korrelierenden Peaks in Peakcluster 6 können einer Verbindung **24** zugeordnet werden, die mit Ausnahme von KSH490 in allen untersuchten Extrakten beobachtet wurde. In den UPLC-QqTOF-MS Messungen wurde **24** mit *m/z* 510,3688 ($[M+H]^+$) und *m/z* 532.3516 ($[M+Na]^+$) bei einer Retentionszeit von 6,25 Minuten detektiert. Die genaueren FT-ICR-MS Messungen ergaben ein *m/z* von 510,36517. Dies entpricht mit einer Massendifferenz von 0,3 ppm einer Verbindung, die von Mitova aus *Sepedonium chrysospermum* isoliert wurde und als Chrysosporid bezeichnet wird [265]. Das Chrysosporid ist ein zyklisches Pentapeptid mit einer Sequenz cyclo-(L-Val-D-Ala-L-Leu-L-Leu-D-Leu). Es zeigte schwache zytotoxische Eigenschaften (IC_{50} = 33,4 µM) gegen die murine P388 Leukämie Zelllinie.



Summenformel: $C_{26}H_{47}N_5O_5$
Exakte Masse: 509.35772

Abbildung 3.39.: Struktur von Chrysosporid (**24**) [265].

Tabelle 3.8.: Diagnostische Fragmentionen [*m/z*] der Peptaibole in Peakcluster 3 und 4.

Verbindung	**67**	**68**	**69**	**70**	**71**	**72**
rt [min]	6.9	7.0	5.2	5.3	6.6	6.7
$[M+2H]^{2+}$	999.0494	999.5415	885.0140	885.5067	884.4965	884.9920
Immoniumion Tyr	136.0735	136.0737	136.0759	136.0753	136.0748	136.0739
Immoniumion Trp	159.0923	159.0899	-	-	-	-
b1-42	187.0860	187.0866	-	-	-	-
a1	201.1021	201.1009	n.d.	n.d.	n.d.	n.d.
b1	n.d.	n.d.	n.d.	n.d.	156.0612	156.0655
b2	300.1358	300.1357	157.0944	157.0959	241.1161	241.1158
b3	385.1907	385.1896	242.1490	242.1494	326.1720	326.1744
b4	470.2386	470.2384	327.2016	327.2034	454.2291	454.2296
b5	555.2958	555.2946	455.2603	455.2597	539.2830	539.2828
b6	683.3486	683.3493	540.3139	540.3148	624.3357	624.3368
b7	768.4001	768.4000	625.3674	625.3579	723.3999	723.4014
b8	853.4491	853.4492	724.4328	724.4331	794.4318	794.4337
b9	952.4978	952.5067	795.4645	795.4654	879.4817	879.4811
b10	1023.5240	1023.5365	880.5133	880.5131	964.5252	964.5270
b11	1108.5711	1108.5787	965.5562	965.5589	1049.5702	1049.5709
b12	1193.6067	1193.6194	1050.6030	1050.6037	1177.6189	1177.6078
b13	1278.6447	1278.6614	1178.6430	1178.6418	1340.6515	1340.6493
b14	1406.6690	1406.6747	1341.6805	1341.6762	1425.6856	1425.6862
b15	1569.7176	1570.7169	1426.7182	1426.7189	n.d.	n.d.
b16	1654.7449	1654.7350	n.d.	n.d.	n.d.	n.d.
b17	n.d.	n.d.	n.d.	n.d.	-	-
b18	n.d.	n.d.	n.d.	n.d.	-	-
y1	118.1235	118.1230	118.1247	118.1233	118.1240	118.1232
y2	246.1788	247.1613	246.1798	247.1638	246.1797	247.1635
y3	343.2358	344.2204	343.2354	344.2191	343.2335	344.2196
y4	428.2827	429.2647	428.2763	429.2608	428.2897	429.2736
y5	591.3482	592.3326	591.3508	592.3335	591.3511	592.3326
y6	719.4080	720.3906	719.4042	720.3867	719.4047	720.3895
y7	804.4534	805.4391	804.4520	805.4343	804.4530	805.4380
y8	889.5009	890.4857	889.4968	890.4856	889.4969	890.4830
y9	974.5391	975.5254	974.5423	975.5276	974.5421	975.5253
y10	1045.5727	1046.5575	1045.5743	1046.5500	1045.5705	1046.5510
y11	1144.6282	1145.6118	1144.6290	1145.6000	1144.6372	1145.6027
y12	1229.6614	1230.6525	1229.6684	n.d.	1229.6846	n.d
y13	1314.6926	n.d.	n.d.	n.d.	n.d.	n.d.
y14	1442.7436	1443.7230	n.d.	n.d.	n.d.	n.d.
y15	1527.7640	n.d.	n.d.	n.d.	n.d.	n.d.
y3	343.2358	344.2204	343.2354	344.2191	343.2335	344.2196
y3 - H_2O	325.2215	326.2072	325.2210	326.2066	325.2196	326.2040
y3 - AS (n)	226.1179	227.1026	226.1159	227.1010	226.1167	227.1006
$[M-H]^-$	-	-	1767	1768	1766	1767
$[M-H-H_2O]^-$	-	-	1749	1750	1748	1749
$[M-H-44]^-$	-	-	-	-	1722	1723
$y18^-$	-	-	1696	1697	-	-
$y17^-$	-	-	1611	1612	1611	1612
$y16^-$	-	-	1526	1527	1526	1527
$y15^-$	-	-	1441	1442	1441	1442
$y14^-$	-	-	1313	1314	1313	1314

3.3.4. Verteilung der identifizierten Peptaibole in S. ampullosporum

Abbildung 3.40 stellt die Verteilung der identifizierten Peptaibole in *S. ampullosporum* anhand einer Heatmap dar. Man erkennt deutlich, dass die Peptaibole aus Peakcluster 1 (**61**, **62**) und 2 (**63**, **64**, **65**, **66**) ausschließlich in den Stämmen KSH 531, 533, 534, 537, 561, 635 und 641 auftreten. Die Peptaibole aus den Peakclustern 3 und 4 bilden eine weitere klar abgegrenzte Gruppe, die aus den Stämmen KSH 499, 500, 502, 544, 549 und 560 besteht. Die methanolischen Extrakte der Pilzstämme der beiden separierten Gruppen wiesen, mit der Ausnahme von KSH544, eine mittlere bis hohe zytotoxische Aktivität auf.

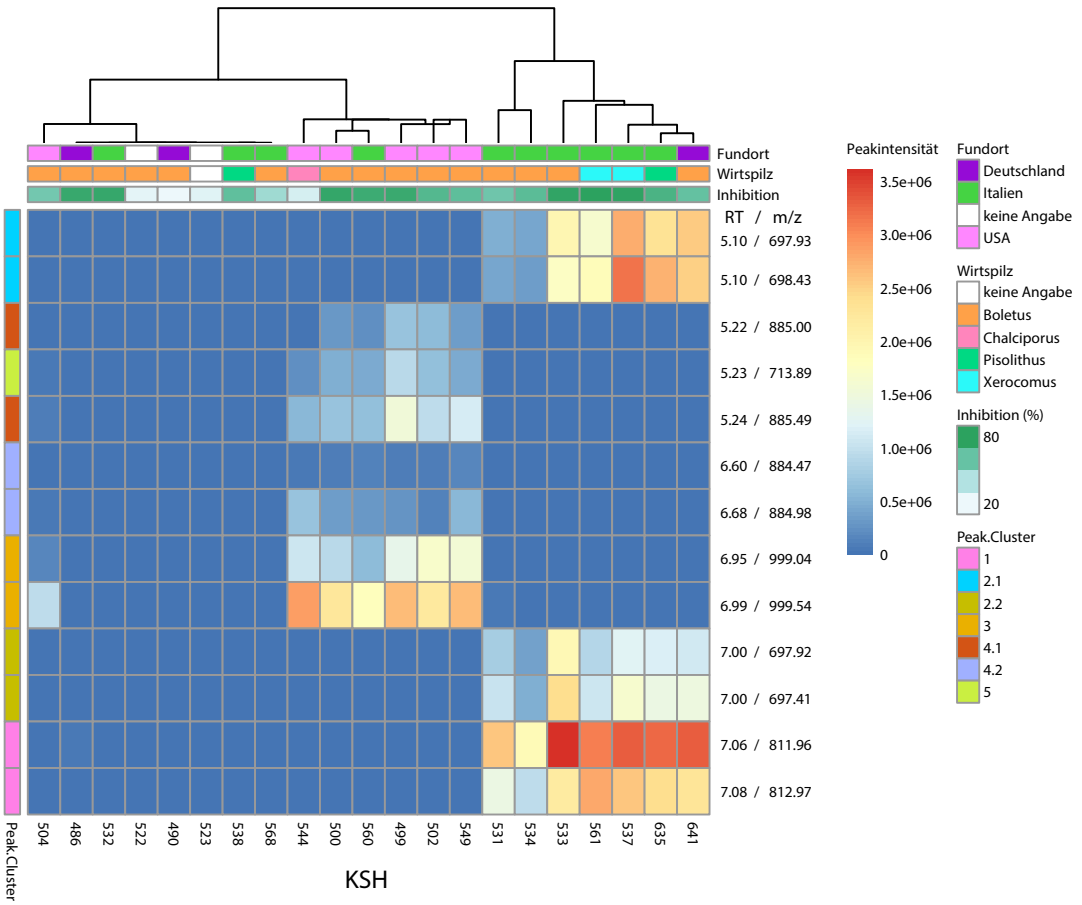


Abbildung 3.40.: Heatmap: Intensität der Peaksignale der identifizierten Peptaibole in der UPLC-QqTOF-MS Analyse der untersuchten *S. ampullosporum* Stämme. Das hierarchische Clustering erfolgte mit der Centroid Methode anhand der Manhattan Distanz.

Demgegenüber fehlen die Peptaibole der Peakcluster 1-4 in den zytotoxisch inaktiven Extrakten der Stämme KSH 490, 522 und 523. Diese Beobachtungen weisen nochmals darauf hin, dass zumindest einige der identifizierten Peptaibole ursächlich für die Wachstumsinhibition verantwortlich sind.
In den Stämmen KSH 486, 532 und 538 konnten keine der identifizierten Peptaibole nachgewiesen werden. Dennoch wurde für die Extrakte der Stämme eine mittlere bis hohe Wachstumsinhibition gemessen. Dies weist darauf hin, dass in diesen Extrakten noch weitere, bislang nicht identifizierte zytotoxisch aktive Verbindungen (möglicherweise weitere Peptaibole?) vorhanden sein müssen.
Weder der Fundort noch die Art des Wirtspilzes scheinen in einem direkten Bezug zur Ausstattung der Peptaibole in *S. ampullosporum* zu stehen.

3.3.5. Isolierung, Strukturaufklärung und zytotoxische Aktivität der Verbindung 61

Isolierung AcorA bietet für die Isolierung einen entscheidenden Vorteil. Da die Masse der zu isolierenden Substanz durch die *in silico* Identifizierung bereits bekannt ist, kann die Substanz in den Fraktionen, die während des Aufreinigunsprozesses generiert werden, durch Massenspektrometrie sehr leicht verfolgt werden. Eine zeitraubende aktivitäts-geleitete Fraktionierung, bei der die einzelnen Fraktionen mit einem entsprechenden Assay getestet werden müssten, ist also nicht erforderlich.
Durch die *m/z*-geleitete Isolierung konnten innerhalb nur weniger Arbeitstage 24 mg von Verbindung (**61**) isoliert werden. Die verwendete Aufreinigungsstrategie orientierte sich dabei an der von Ritzau beschriebenen Methode [227]. Dazu wurden 50 Erlenmeyerkolben à 150 mL Malz-Pepton-Medium mit KSH 533 inokuliert und 25 Tage bei Raumtemperatur inkubiert. Nach Filtration des Pilzmycels wurden Kulturfiltrat und Mycel je 3x mit Ethylacetat extrahiert. Die vereinigten Ethylacetat Extrakte (HAM115E, 989 mg) wurden anschließend mit Sephadex LH-20 säulenchromatographisch aufgetrennt. Die Fraktionen 30-34 (HAM115ES30-34, 134 mg) enthielten die Verbindung **61**. Diese Fraktionen wurden vereinigt und mittels präparativer HPLC weiter aufgereinigt. Die Fraktion mit dem Hauptpeak (HAM115_ES_30-34P2, 40 mg) wurde anschließend mit einer präparativen HPLC noch ein weiteres Mal aufgereinigt (HAM115_ES_30-34P2.2).

Strukturaufklärung Das Totalionenstromchromatogramm (Abb. 3.41 a) der Fraktion HAM115_ES_30-34P2.2 zeigt einen einzelnen großen Peak bei RT = 4,9 min. Das zugehörige Übersichtsspektrum (Abb. 3.41 b) zeigt drei Massensignale bei *m/z* 812,28, *m/z* 1622,27 und *m/z* 1644,94, die sich auf die $[M+2H]^{2+}$-, $[M+H]^{+}$- und $[M+Na]^{+}$-Ionen der Verbindung **61** zurückführen lassen. Im MS/MS-Spektrum von *m/z* 812 ist die b-Ionenserie von b_2 bis b_{13} nahezu vollständig erkennbar. Mit der Ausnahme des y_3-Ions - ist auch die komplementäre y-Ionenserie von y_2 bis y_{11} gut sichtbar (Abb. 3.41 c). Das y_2-Ion entspricht einem Dipeptid aus Gln^{13}-$LxxOH^{14}$. Das b_1-Ion entspricht mit einer Masse von *m/z* 228,89 einem Acetyltryptophan. Die Präsenz von Tryptophan im Molekül wird durch das UV-Vis Spektrum bestätigt (siehe C.3.2, Anhang). Neben dem Hauptmaximum bei 220 nm, welches der Absorption der Peptidbindung entspricht, enthält das Spektrum zwei kleinere Maxima bei 281 und 290 nm, die charakteristisch für Tryptophan sind [381]. Durch UPLC-IT-MS^n Messungen der isolierten Verbindung **61** wurde die in den methanolischen Rohextrakten mittels UPLC-QqTOF-MS/MS bestimmte Primärstruktur somit nochmals

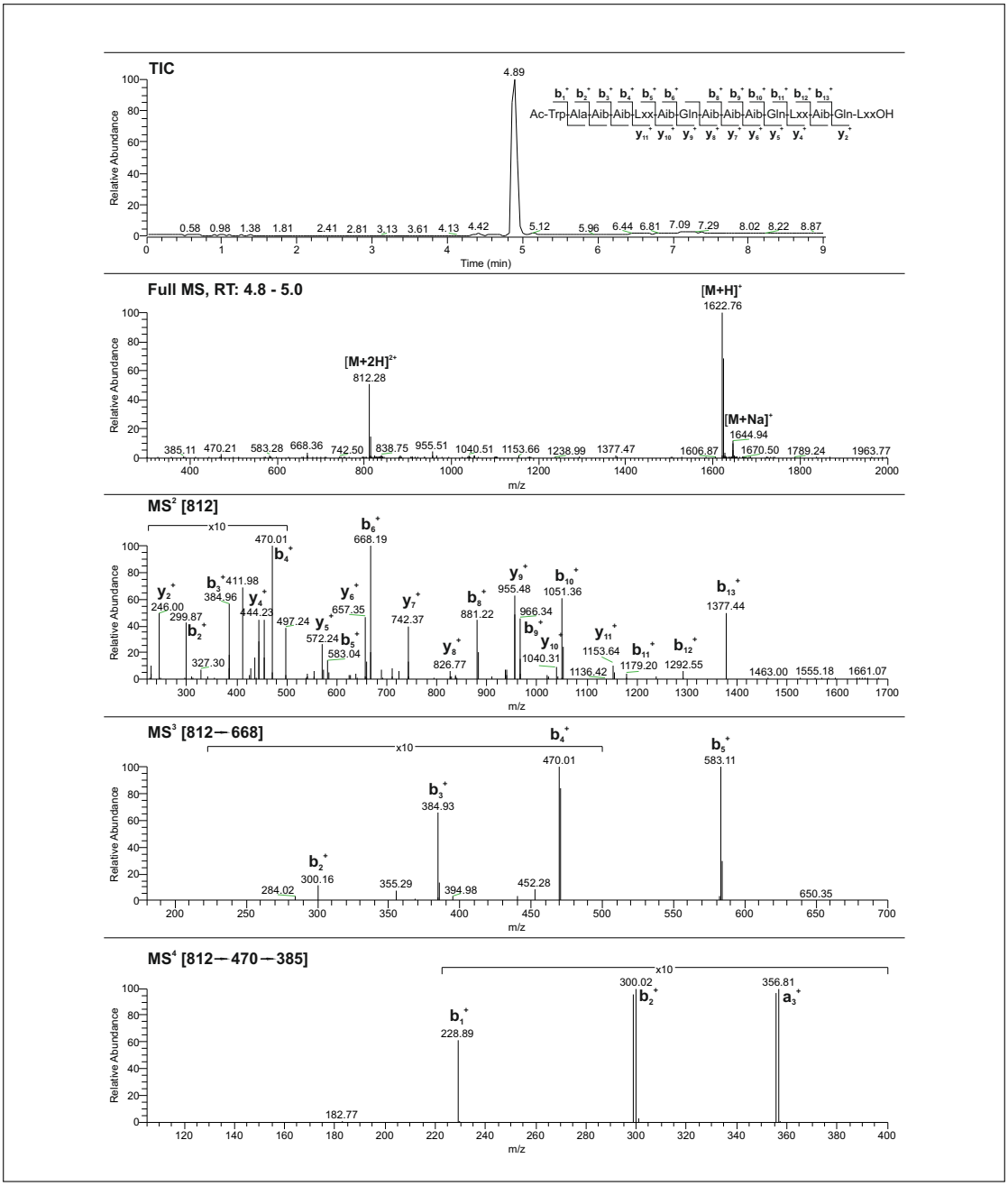


Abbildung 3.41.: UPLC-IT-MS^n-Messungen der aufgereinigten Fraktion HAM115_ES_30-34P2.2 zur Strukturaufklärung von Verbindung **61**.

bestätigt. Die Aminosäuresequenz ist - ohne Berücksichtigung möglicher enantiomerer Aminosäuren - identisch mit der von Ampullosporin A [227].

Biologische Aktivität der Verbindung 61 Die isolierte Verbindung **61** wurde in einem Konzentrationsbereich zwischen 50 µM und 16 nM auf ihre Zytotoxizität in Hinblick auf die HT-29 Krebszelllinie untersucht. Der aus der Kurve berechnete IC_{50}-Wert liegt bei 4,46 µM (95 % Konfidenzintervall: 4,02, 4,95 µM). Verbindung **61** zeigt somit eine mittelstarke Toxizität und ist vergleichbar mit der von Chrysaibol auf die murine Leukämie Zelllinie P388 (IC_{50} = 6,61 µM) [221].

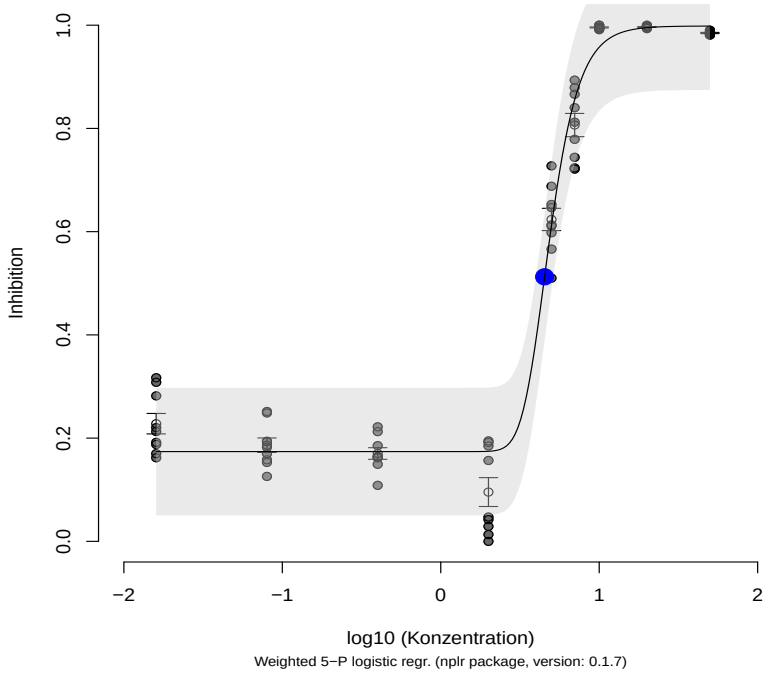


Abbildung 3.42.: Wachstumsinhibition von HT-29 Zellen nach Zugabe von **61** in einem Konzentrationsbereich zwischen 50 µM und 16 nM. Der IC_{50}-Wert wurde auf 4,46 µM (Konfidenzint.95%: 4,02, 4,95 µM) kalkuliert. Die Fehlerbalken geben den Standardfehler wieder. $R^2 = 0.9736$.

3.3.6. Diskussion

Im Rahmen dieser Dissertation wurde die Aktivitäts-Korrelations-Analyse (AcorA) als „Quick & Easy“ Methode zur Identifizierung von bioaktiven Substanzen in komplexen Mischungen vorgestellt. Grundlage für AcorA ist die Korrelation zwischen massenspektrometrischen oder anderen spektroskopischen/chromatografischen Signalen der Verbindungen einer Probe und dessen gemessener Aktivität. Um einen statistisch auswertbaren Datensatz zu erhalten, muss die aktive Komponente über eine Reihe von Extrakten in möglichst hoher Konzentrationsvarianz verteilt werden. Michels hat gezeigt, dass dies beispielsweise durch unterschiedliche Behandlung eines aliquotierten Rohextraktes geschehen kann [277]. Nach der erfolgreichen Testung in dem Proof of Concept Experiment wurde AcorA zur Identifizierung von zytotoxisch wirkenden Metaboliten in dem fungicol lebenden Pilz *Sepedonium ampullosporum* verwendet. Die Anwendung der Aktivitäts-Korrelations-Analyse auf den mit FT-ICR-MS ermittelten Metabolitenprofilen resultierte in zwei Hitlisten, bestehend aus *m/z*-Werten mit signifikanter Korrelation zur Bioaktivität. Bereiche in der Hitliste mit hohen Korrelationskoeffizienten und hoher Peakdichte weisen auf Isotopen- und Adduktpeaks von Verbindungen hin, die statistisch signifikant mit der zytotoxischen Wirkung korreliert sind. Auf diese Weise konnten insgesamt sechs Peakcluster lokalisiert werden.
Durch Datenbanksuche und *de novo* Sequenzierung konnten diesen Peakclustern insgesamt 12 Peptaibole (**61**-**72**) zugeordnet werden, die mit der zytotoxischen Aktivität korreliert sind. Im folgenden soll nun erörtert werden, ob diese Korrelationen einen kausalen Ursprung besitzen.

Biologische Aktivität der identifizierten Peptaibole Durch die *m/z*-geleitete Isolierung konnten innerhalb weniger Wochen 20 mg von Verbindung **61** isoliert werden. Die massenspektrometrische und spektroskopische Analyse hat ergeben, dass es sich dabei mit hoher Wahrscheinlichkeit um das bereits zuvor aus *S. ampullosporum* isolierte AmpA handelt [227]. Der für die Zytotoxizität ermittelte IC_{50}-Wert von 4,46 µM liegt im Bereich anderer bereits charakterisierter Peptaibole [219, 274, 221, 382, 383]. Die biologische Aktivität geht mutmaßlich auf die membranmodifizierenden Eigenschaften der Peptaibole zurück, die auf ihrem amphiphilen Charakter und der Fähigkeit zur Bildung von helikalen Strukturen basieren [201, 187].
Röntgenstrukturanalysen von AmpA zeigen, dass der hydrophile Anteil über die β-Carboxylamidgruppen von Gln^{7}, Gln^{11} und Gln^{14} sowie über die Carbonylsauerstoffe von Aib^{10} und Gln^{11} vermittelt wird [197]. Die hydrophoben Anteile, die die Integration in

Zellmembranen ermöglichen, werden über das Ac-Trp^{1}, Leu^{5}, Leu^{12} und LeuOH^{15} gebildet. Röntgenstrukturanalysen und Festphasen-NMR [198] zeigen, dass Ampullosporin A in einer gemischten $3_{10}/\alpha$-Helix vorliegt. Interessanterweise scheint sich der Anteil der $3_{10}/\alpha$-Helix eines in einer Membran lokalisierten Peptaibols zugunsten der 3_{10}-Helix zu verschieben [198, 215]. Auf diese Weise nimmt das Peptaibol eine gestrecktere Konformation an, was eine tiefere Penetration in die Zellmembran erlaubt.
Der membranmodifizierende Effekt von AmpA wurde an verschiedenen Modellmembranen untersucht. In niedrigen Konzentrationen lagert sich AmpA zunächst parallel zur Oberfläche an die Membran an bzw. in die Membran ein [198, 214, 215]. Das Molekül ist dabei so ausgerichtet, dass die hydrophoben Abschnitte, insbesondere Ac-Trp^{1}, innerhalb der Membran liegen, und die hydrophilen Gln$^{7,\ 11,\ 14}$ mit den polaren Kopfgruppen der Phosphatidylcholine in Wechselwirkung treten.
Wird eine bestimmte Schwellenkonzentration überschritten kommt es zur Schwächung der Membran und schließlich zu dessen Disruption. Eid [214], Bortolus [215] und Salnikov [198] kommen aufgrund ihrer Analysen unabhängig voneinander zu dem Schluss, dass AmpA sowohl detergenzartig, d. h. nach dem Carpet-Mechanismus, als auch in spannungsabhängiger Weise über die Bildung von Membranporen, agiert.
Diese Beobachtungen lassen Rückschlüsse auf die mögliche Bioaktivität der Verbindungen **62**, **63**, **64**, **65** und **66** zu.
Verbindung **62** enthält an Position 14 ein - möglicherweise durch Deamidierung entstandenes - Glu anstatt eines Gln in **61**. Wie oben beschrieben, treten Gln$^{7,\ 11,\ 14}$ mit den polaren Kopfgruppen der Phosphatidylcholine in Kontakt. Dieser Kontakt wäre auch bei der Substitution Gln^{14} $\rightarrow$Glu^{14} gewährleistet. Zudem sind die hydrophoben Bereiche des Moleküls nicht betroffen. Es kann daher angenommen werden, dass die Funktion von **62** durch den Austausch nicht wesentlich beeinträchtigt wird und vermutlich eine ähnliche Bioaktivität aufweist wie **61**.
Eine andere Situation ergibt sich für die Verbindungen **63** und **64**. In beiden Fällen fehlt das Ac-Trp^{1} am *N*-Terminus, sodass die Verbindungen *N*-terminal mit einer freien Aminogruppe beginnen. Wie die UPLC-QqTOF-MS Messungen gezeigt haben, ist die Polarität im Vergleich zu **61** und **62** erwartungsgemäß deutlich erhöht. Ein Eindringen in die Membran scheint aufgrund der hohen Polarität nicht mehr möglich. Tatsächlich haben Messungen von [des-Trp^{1}]Ampullosporin A und [des-Ac-Trp^{1}]Ampullosporin A gezeigt, dass die bei AmpA normalerweise beobachtete Induktion der Pigmentbildung bei *Phoma destructiva* sowie die Ausbildung eines hypothermischen Effekts bei Mäusen, bei den trunkierten Derivaten vollständig ausbleiben [267]. Anhand von Membranleitfähigkeitsmessungen haben Grigoriev

et al. demonstriert, dass die Zugabe von [des-Ac-Trp^1]Ampullosporin A zu Doppelschichtmembranen aus Sojabohnenphosphatidylcholin im Gegensatz zu den Ampullosporinen A-D nicht zu einer Änderung der Membranleitfähigkeit führt, wie sie nach Membrandisruption zu erwarten wäre [269]. Es ist daher davon auszugehen, dass die Verbindungen **63** und **64** nicht über zytotoxische Eigenschaften verfügen. Da AcorA mit den FT-ICR-MS-Spektren durchgeführt wurde, ist allerdings zu beachten, dass sich die Signale aus Peakcluster 2.1 und Peakcluster 2.2 ohne vorherige chromatographische Trennung überlagern. Daher bestand für die Korrelationsanalyse kein Unterschied zwischen den in Hinblick auf die Polarität deutlich unterschiedlichen Verbindungen. Die Tatsache, dass sie von AcorA trotzdem als zur Zytotoxizität signifikant korrelierende Verbindungen eingestuft wurden, beruht mutmaßlich darauf, dass ihre Massensignale in hoher Weise mit denen des AmpA korreliert sind.

Die Zytotoxizität der Verbindungen in Peakcluster 2.2 (**65** und **66**) ist nur schwer einzuschätzen, da die Struktur des *N*-terminalen Fragments nicht zweifelsfrei bestimmt werden konnte (s. u.). Ein mögliches *N*-Acetyl-hydroxyprolin wäre dem bei den bekannten Ampullosporinen auftretenden *N*-Acetyl-Tryptophan strukturell recht ähnlich. Die geringen Retentionszeitdifferenzen zwischen **65**/**66** und AmpA deuten zudem darauf hin, dass nur geringe Unterschiede in der Polarität bestehen. Eine dem AmpA ähnliche zytotoxische Aktivität wäre daher plausibel.
Die Aktivitäten der Verbindungen in Peakcluster 3 und 4 können ebenfalls nur anhand der über die MS/MS Analysen ermittelten Primärstrukturen abgeschätzt werden. Das Sequenzalignment zeigt eine hohe Übereinstimmung von **67** mit AmpA (siehe C.4). Mit der Ausnahme der konservativen Substitutionen an Position 5 (Leu^5 $\rightarrow$Aib^5), 10 (Aib^{10} $\rightarrow$Vxx^{10}) und 15 (Leu^{15} $\rightarrow$Tyr^{15}) enthält **67** gegenüber AmpA zwei Insertionen an Position 11-13 (Ala^{11}-Aib^{12}-Aib^{13}) und Position 17 (Pro^{17}). Strukturell handelt sich bei **67** somit lediglich um eine elongierte Variante von AmpA. In gleichem Sinne ist **68** eine Strukturvariante von **62**. Der hohe Anteil von Aib lässt weiterhin auf die für Peptaibole typischen helikalen Strukturen schließen. Da sich die Retentionszeiten von **67** und **68** nur marginal von **61** und **62** unterscheiden, kann ein ähnliches Maß an Lipophilie angenommen werden. Zusammengefasst lassen diese Beobachtungen den Schluss zu, dass die Verbindungen in Peakcluster 3 über zytotoxische Eigenschaften verfügen.
Analog zu den Betrachtungen zur Zytotoxizität der Verbindungen in Peakcluster 2.1 ist eine zytotoxische Wirkung der Verbindungen in Peakcluster 4.1 nicht zu erwarten. Wie von Grigoriev *et al.* beschrieben [269] ist das *N*-Acetyltryptophan essenziell für die biologische Aktivität in AmpA. Aufgrund der hohen Sequenzhomologie zu AmpA ist anzunehmen, dass das Fehlen des *N*-Acetyltryptophans und der somit freien *N*-terminalen Aminogruppe

in **67** und **68** ebenfalls mit einem Verlust an Bioaktivität verbunden ist. Die signifikanten Korrelationen der Massensignale der Verbindungen in Peakcluster 4.1 zu der gemessenen Bioaktivität beruhen daher auf einer hohen Korrelation zu den Massensignalen der mutmaßlich zytotoxischen Verbindungen in Peakcluster 3.
Die Verbindungen in Peakcluster 4.2 wiederum weisen eine hohe Strukturhomologie zu den Verbindungen in Peakcluster 2.2 auf. **71** und **72** unterscheiden sich im Wesentlichen durch die Insertionen an Position 11-13 und 17 von den Verbindungen **65** und **66**. Weiterhin konnte eine Summenformel von $C_7H_{10}NO_3^+$ für das *N*-terminale Fragment aller vier Verbindungen ermittelt werden. Da dem *N*-Terminus eine zentrale Rolle für die biologische Aktivität bei den Ampullosporinen zukommt, ist eine Abschätzung der Bioaktivität der Verbindungen aufgrund der nicht exakt bekannten Struktur für den *N*-Terminus nur schwer möglich. Der hohe Anteil von Aib und die allgemein hohe Sequenzhomologie zu AmpA lassen jedoch helikale Strukturen vermuten. Zudem liegen die Retentionszeiten von **71** und **72** in einem ähnlichen Bereich wie AmpA, sodass von einer vergleichbaren Polarität ausgegangen werden kann. Eine zytotoxische Aktivität der Verbindungen in Peakcluster 4.2 wäre somit plausibel.

Strukurvorschläge für $C_7H_{10}NO_3^+$ Für das *N*-terminale Fragment der Verbindungen **65** und **66** in Peakcluster 2.2 sowie **71** und **72** in Peakcluster 4.2 konnte mithilfe der Pseudo-MS^3 Messungen eine Summenformel von $C_7H_{10}NO_3^+$ bestimmt werden. Zusätzlich wurde mit der Ionenfalle im Negativ-Ionen-Modus ein deutlich sichtbarer Neutralionenverlust von 44 Da beobachtet, der ausschließlich bei der Fragmentierung der oben genannten Verbindungen auftrat. Dies und die Tatsache, dass **65** und **71** keine weiteren Aminosäuren enthalten, die zu einer Abspaltung von 44 Da führen könnten, weisen daraufhin, dass das *N*-terminale Fragment entweder eine zur Carboxylgruppe (CO_2, *m/z* 43,9898) oder zum Acetaldehyd ($C_2H_4O_1$, *m/z* 44,0262) fragmentierbare funktionelle Gruppe (etwa eine Acetylgruppe) als Strukturelement enthält.
Geht man von einer klassischen Peptidbindung aus, entspräche das einfach-geladene Fragment $C_7H_{10}NO_3^+$ einer Verbindung mit der Summenformel $C_7H_{11}NO_4$. Sucht man mit dieser Summenformel in der Naturstoffdatenbank „Supernatural II“ (ca. 355000 Einträge) [384], erhält man ohne Berücksichtigung von Stereoisomeren insgesamt 11 Treffer (siehe Anhang Tabelle C.5). Von diesen 11 Verbindungen entsprechen wiederum 6 Verbindungen der Struktur einer α-Aminosäure. Beschränkt man sich anschließend auf die Grundstrukturen bereits bekannter Peptaibole [186], verbleiben Acetyl-hydroxyprolin und Carboxypipecolinsäure als mögliche Strukturvarianten.

Das Auftreten von Hydroxyprolin in Peptaibolen ist nicht ungewöhnlich. Eine Suche in der CPDB ergibt 63 Peptaibole mit mindestens einem Hydroxyprolin in der Aminosäuresequenz. In diesen Peptaibolen (vor allem Antiamoebine [385], Bergofungine [386], Cephaibole [229], Emerimicine [387] und Zervamicine [388, 389]) wird das Hydroxyprolin fast ausschließlich an Position 10 und 13 beobachtet. In Cicadapeptinen [217] befindet sich das Hydroxyprolin am *N*-Terminus; allerdings ist es am Ringstickstoff mit Decansäure verestert. Ein *N*-terminales Acetylprolin wurde bereits im Adenopeptin (*Chrysosporium* sp.) [390] und im Acremopeptin (*Acremonium* sp.) nachgewiesen [230]. Vor diesem Hintergrund sowie der Annahme, dass der beobachtete Neutralionenverlust von 44 Da einer für Peptaibole klassischen Acetylgruppe entspricht, scheint ein *N*-terminales *N*-Acetyl-hydroxyprolin in den Verbindungen **65**, **66**, **71** und **72** eine plausible Erklärung zu sein.
Geht man andererseits davon aus, dass der Neutralionenverlust von 44 Da der Abspaltung einer Carboxylgruppe entspricht, könnte das *N*-terminale Fragment auch einer Carboxypipecolinsäure entsprechen. Die Pipecolinsäure ist ein Abbauprodukt des Lysins und wurde bislang u. a. im Adenopeptin, Efrapeptinen [391] und Neoefrapeptinen [392] nachgewiesen. In den beiden letztgenannten Peptaibolklassen tritt die Pipecolinsäure als *N*-Acetylpipecolinsäure an Sequenzposition 1 auf. Die Efrapeptine und Neoefrapeptine zeichnen sich durch eine zytotoxische Aktivität gegenüber diversen Krebszelllinien aus. Der Mechanismus beruht allerdings nicht auf einer membranmodifizierenden Eigenschaft, sondern auf der Inhibition der F_0F_1-ATPase der Atmungskette [235].
Eine Carboxypipecolinsäure wurde bislang noch nicht in Peptaibolen beobachtet und würde aufgrund des pks-Wertes der *N*-terminalen Carboxylgruppe von -0,35 dissoziiert vorliegen. Eine Membranpermeation der entsprechenden Peptaibole wäre somit eher unwahrscheinlich. Zudem wurde bei Peptaibolen bislang keine Übertragung einer Carboxylgruppe auf eine *N*-terminale Aminosäure beschrieben, sodass auch eine postsynthetisch carboxylierte Pipecolinsäure im Vergleich zum *N*-Acetyl-hydroxyprolin eher unwahrscheinlich ist.
Die exakte Natur des *N*-terminalen Fragments bleibt somit unklar und kann letztendlich nur über die Isolierung der Peptaibole und anschließende NMR-spektroskopische Charakterisierung erfolgen.

Biosynthese und NRPS Reiber *et al.* konnten zeigen, dass die Synthese der Ampullosporine A-E mit zwei Proteinen HMWP1 (1,5 MDa) und HMWP2 (350 kDa) assoziiert ist, die eine für NRPS typische Adenylierungsdomäne tragen [237]. Sie postulieren, dass das kleinere HMWP2 Protein aus einer Polyketidsynthase (PKS) zur Generierung der Ace-

tylgruppe aus Malonyl-CoA, sowie den Modulen für die ersten beiden Aminosäuren (Trp, Ala) besteht. Das größere HMWP1 Protein beherbergt die 13 weiteren Module sowie die Reduktasedomäne.
Da die Primärstruktur von **61** der des Amp A entspricht und die weiteren Verbindungen in Peakcluster 1 und 2 lediglich leicht modifizierte Derivate von AmpA darstellen, ist anzunehmen, dass die Verbindungen in Peakcluster 1/2 durch das beschriebene NRPS-System synthetisiert wurden. Die Ergebnisse der vorliegenden Dissertation weisen jedoch darauf hin, dass in den untersuchten *S. ampullosporum* Spezies mindestens zwei verschiedene NRPS Systeme für die Biosynthese der Peptaibole verantwortlich sein müssen. Hierzu können verschiedene Gründe angeführt werden:
Strukturell bestehen die Peptaibole in Peakcluster 1 und 2 aus 14 bis 15 Aminosäuren, die Peptaibole in Peakcluster 3 und 4 aus 18 bis 19 Aminosäuren. Da für die Insertion jeder Aminosäure ein Modul, bestehend aus Kondensations-, Adenylierungs- und Thiolierungsdomäne, notwendig ist, müssten für die Synthese der Peptaibole in Peakcluster 3/4 19 anstatt 15 Module in Peakcluster 1/2 zur Verfügung stehen. Theoretisch wäre es zwar denkbar, dass Peptaibole mit unterschiedlicher Kettenlänge durch s. g. Modul Skipping, ähnlich wie bei *Trichoderma virens* [245], von der selben NRPS synthetisiert werden könnten. Die Peptaibole in Peakcluster 1/2 bzw. 3/4 wurden jedoch niemals zusammen innerhalb eines KSH Stammes detektiert, sondern immer nur exklusiv. Auf einen gemeinsamen biosynthetischen Ursprung weisen zudem die hohen Korrelationen der Massensignale innerhalb der Peakcluster 1/2 bzw. 3/4 hin.
Über den evolutionsbiologischen Ursprung der hier postulierten zwei NRPS Systeme kann nur spekuliert werden, da für *Sepedonium* spp. bislang keine genomischen Daten oder Nukleotidsequenzen der für die NRPS kodierenden *tex* Gene publiziert wurden.
Ausgangspunkt für die beiden postulierten NRPS-Systeme ist womöglich eine NRPS mit 19 Modulen. Die Entstehung der NRPS für die 15-AS Peptaibole (Peakcluster 1/2) ließe sich durch zwei Deletionen in den Modulen für die Positionen 11-13 und 17 von **67** erklären. Eine weitere Möglichkeit zur Entstehung der kürzeren Peptaibole in Peakcluster 1/2 wäre alternatives Spleißen der mRNA-Transkripte.
Die Entstehung der *N*-terminal trunkierten Varianten von **61**, **62**, **69** und **70** scheint rätselhaft. Das Auftreten von *N*-terminal trunkierten Peptaibolvarianten wurde zuvor schon von Degenkolb [243, 393] und Neuhof [244] bei verschiedenen *Trichoderma* Arten beobachtet. Da Peptaibole an den NRPS vom *N*-Terminus zum C-Terminus synthetisiert werden, kann es sich bei den oben genannten Verbindungen nicht um einfache biosynthetische Abbruchprodukte handeln.

Eine artifizielle Bildung trunkierter Varianten, wie sie beispielsweise von Theis *et al.* [394] mit Trifluoressigsäure induziert wurde, scheint ebenfalls unwahrscheinlich, da die Extrakte lediglich während der 10 minütigen Chromatographie mit Säure (0,1 % Essigsäure) in Kontakt kamen.

Ausblick Mithilfe der Aktivitäts-Korrelations-Analyse konnten insgesamt 12 Peptaibole identifiziert werden, die positiv mit der Zytotoxizität der *S. ampullosporum* Extrakte korreliert sind. Während die Zytotoxizität von Verbindung **61** (Ampullosporin A) eindeutig bestimmt wurde, konnten zur Wirksamkeit der 11 weiteren Verbindungen lediglich Hypothesen aufgestellt werden. Um die Zytotoxizität der Verbindungen **62**-**72** zu testen, wäre daher die Isolierung dieser Verbindungen aus *S. ampullosporum* interessant. Mithilfe von *de novo* Sequenzierungen konnten zwar die Primärstrukturen der Verbindungen **62**-**72** größtenteils bestimmt werden. Nichtsdestotrotz müssen die Strukturen dieser Peptaibole - auch in Hinblick auf die Stereochemie - anhand von NMR-spektroskopischen Methoden validiert werden. Dies gilt insbesondere für die Peptaibole **65**, **66**, **71** und **72**, die offenbar über einen ungewöhnlichen *N*-Terminus verfügen.

Wie die vorliegende Studie gezeigt hat, sind FT-ICR-MS Messungen zur metabolomischen Untersuchung von *Sepedonium* Extrakten nur bedingt geeignet. Die Mikroheterogenität der Peptaibole führte in den FT-ICR-MS Spektren zu Signalüberlagerungen verschiedener Peptaibole mit deutlich unterschiedlicher Polarität, sodass Kokorrelationen zwischen diesen Peakclustern die Aktivitäts-Korrelations-Analyse beeinträchtigt haben. Insbesondere für solche Fälle könnte eine vorherige chromatographische Trennung der Extrakte sinnvoll sein und zu einer Verbesserung der Ergebnisse einer Aktivitäts-Korrelations-Analyse führen.

Die Untersuchungen der multivariaten Analysemethoden haben gezeigt, dass insbesondere die Elastic Net Analyse (aber auch QPAR) exzellente Ergebnisse im Hinblick auf die Variablenselektion im Proof of Concept Experiment geliefert hat. Es wäre daher interessant zu untersuchen, ob die Anwendung dieser Methode auf die generierten Datensätze der *S. ampullosporum* Extrakte die gleichen (oder zumindest ähnliche) Ergebnisse liefert, wie die AcorA-Methode.

Das Biosynthesemuster der Peptaibole in den untersuchten *Sepedonium* Stämmen deutet auf zwei unterschiedliche NRPS-Systeme hin. Die molekularbiologische Sequenzierung der entsprechenden *tex* Gene könnte daher Aufschluss über die tatsächlichen Biosynthese der Peptaibole in *S. ampullosporum* geben. Durch Sequenzierung der ITS, EF1-α und RPB2 Sequenzen könnte zusätzlich geklärt werden, ob die Biosynthese der verschiedenen Peptaibole innerhalb der untersuchten *S. ampullosporum* Stämme sogar auf phylogenetischen Differenzen beruhen.

Literaturverzeichnis

[1] K. Adachi, Matrix-Based Introduction to Multivariate Data Analysis. Singapore: Springer Singapore, 2016. [Online]. Available: http://link.springer.com/10.1007/978-981-10-2341-5

[2] L. Fahrmeir, T. Kneib, and S. Lang, Regression. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. [Online]. Available: http://link.springer.com/10.1007/978-3-642-01837-4

[3] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning, 2nd ed., ser. Springer Series in Statistics. New York, NY: Springer New York, mar 2009. [Online]. Available: http://www.springer.com/de/book/9780387848570

[4] P. Filzmoser, B. Liebmann, and K. Varmuza, "Repeated double cross validation," Journal of Chemometrics, vol. 23, no. 4, pp. 160–171, 2009. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/cem.1225/abstract

[5] S. Wetzel, R. S. Bon, K. Kumar, and H. Waldmann, "Biology-oriented synthesis," Angewandte Chemie - International Edition, vol. 50, no. 46, pp. 10 800–10 826, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22038946

[6] F. E. Koehn and G. T. Carter, "The evolving role of natural products in drug discovery," Nature Reviews Drug Discovery, vol. 4, no. 3, pp. 206–220, 2005. [Online]. Available: https://www.nature.com/articles/nrd1657

[7] D. J. Newman and G. M. Cragg, "Natural Products as Sources of New Drugs from 1981 to 2014." Journal of natural products, vol. 79, no. 3, pp. 629–61, mar 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26852623

[8] ——, "Natural products as sources of new drugs over the 30 years from 1981 to 2010." Journal of natural products, vol. 75, no. 3, pp. 311–35, mar 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3721181

[9] M. S. Butler, A. A. Robertson, and M. A. Cooper, "Natural product and natural product derived drugs in clinical trials." Natural product reports, vol. 31, no. 11, pp. 1612–61, nov 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25204227

[10] D. A. Dias, S. Urban, and U. Roessner, "A Historical Overview of Natural Products in Drug Discovery," Metabolites, vol. 2, no. 2, pp. 303–336, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24957513

[11] A. G. Atanasov, B. Waltenberger, E.-M. Pferschy-Wenzig, T. Linder, C. Wawrosch, P. Uhrin, V. Temml, L. Wang, S. Schwaiger, E. H. Heiss, J. M. Rollinger, D. Schuster, J. M. Breuss, V. Bochkov, M. D. Mihovilovic, B. Kopp, R. Bauer, V. M. Dirsch, and H. Stuppner, "Discovery and resupply of pharmacologically active plant-derived natural products: A review." Biotechnology advances, vol. 33, no. 8, pp. 1582–1614, dec 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0734975015300276

[12] S. Anderson, Making Medicines: A Brief History of Pharmacy and Pharmaceuticals, 1st ed., S. Anderson, Ed. London: Pharmaceutical Press, 2005. [Online]. Available: http://www.pharmpress.com/product/9780857110992/making-medicines

[13] W. E. Müller and G. Holoubek, "Plausibilität für eine therapeutische Anwendung als Antidepressivum?: Die Pharmakologie von Johanniskrautextrakt," Pharmazie in unserer Zeit, vol. 32, no. 3, pp. 220–226, may 2003. [Online]. Available: http://doi.wiley.com/10.1002/pauz.200390068

[14] M. Unger, "Pflanzliche Sedativa: Neue Aspekte zu altbewährten Arzneipflanzen," Pharmazie in unserer Zeit, vol. 36, no. 3, pp. 206–212, may 2007. [Online]. Available: http://doi.wiley.com/10.1002/pauz.200600219

[15] E. Luppold, "Matricaria chamomilla: eine alte und neue Arzneipflanze," Pharmazie in Unserer Zeit, vol. 13, no. 3, pp. 65–70, 1984. [Online]. Available: http://doi.wiley.com/10.1002/pauz.19840130301

[16] P. Janssen, "A review of the chemical features associated with strong morphine-like activity." British journal of anaesthesia, vol. 34, no. 4, pp. 260–8, apr 1962. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14451235

[17] R. S. Vardanyan and V. J. Hruby, "Fentanyl-related compounds and derivatives: current status and future prospects for pharmaceutical applications." Future

medicinal chemistry, vol. 6, no. 4, pp. 385–412, mar 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24635521

[18] K. Miotto, A. K. Cho, M. A. Khalil, K. Blanco, J. D. Sasaki, and R. Rawson, "Trends in Tramadol: Pharmacology, Metabolism, and Misuse." Anesthesia and analgesia, vol. 124, no. 1, pp. 44–51, jan 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27861439

[19] E. Sanchez-Rangel and S. E. Inzucchi, "Metformin: clinical use in type 2 diabetes," Diabetologia, vol. 60, no. 9, pp. 1586–1593, sep 2017. [Online]. Available: http://link.springer.com/10.1007/s00125-017-4336-x

[20] C. J. Bailey, "Metformin: historical overview," Diabetologia, vol. 60, no. 9, pp. 1566–1576, sep 2017. [Online]. Available: http://link.springer.com/10.1007/s00125-017-4318-z

[21] H. Müller and H. Reinwein, "Zur Pharmakologie des Galegins," Arch Exp Path Pharmacol, vol. 125, no. 3-4, pp. 212–228, 1927. [Online]. Available: https://link.springer.com/article/10.1007/BF01862957

[22] C. Watanabe, "Studies in the metabolism changes induced by administration of guanidien bases: Influence of injected guanidin hydrochloride upon blood sugar content." Journal of Biological Chemistry, vol. 33, pp. 253–265, 1918. [Online]. Available: http://www.jbc.org/content/33/2/253.citation

[23] J. Sterne, "Du nouveau dans les antidiabetiques. La NN dimethylamine guanyl guanide (N.N.D.G.)," Maroc Medical, vol. 36, pp. 1295–1296, 1957.

[24] Agency for Healthcare Research and Quality (AHRQ), "Metformin Hydrochloride, Drug Usage Statistics, United States, 2004 - 2014," 2014. [Online]. Available: http://clincalc.com/DrugStats/Drugs/MetforminHydrochloride

[25] World Health Organization (2015), "WHO Model List of Essential Medicines," 2015.

[26] B. M. Heckman-Stoddard, A. DeCensi, V. V. Sahasrabuddhe, and L. G. Ford, "Repurposing metformin for the prevention of cancer and cancer recurrence," Diabetologia, pp. 1–9, 2017. [Online]. Available: https://link.springer.com/article/10.1007/s00125-017-4372-6

[27] M. J. R. Desborough and D. M. Keeling, "The aspirin story - from willow to wonder drug." British journal of haematology, vol. 177, no. 5, pp. 674–683, jun 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28106908

[28] J. Oates, "The 1982 Nobel Prize in Physiology or Medicine," Science, vol. 218, no. 19, pp. 765–768, 1982. [Online]. Available: http://science.sciencemag.org/content/218/4574/765

[29] G. Taubes, "The bacteria fight back." Science (New York, N.Y.), vol. 321, no. 5887, pp. 356–61, jul 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18635788

[30] V. Kontis, J. E. Bennett, C. D. Mathers, G. Li, K. Foreman, and M. Ezzati, "Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble." Lancet (London, England), vol. 389, no. 10076, pp. 1323–1335, apr 2017. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(16)32381-9

[31] R. Suzman and J. Beard, "Global Health and Aging," NIH Publication no 117737, vol. 1, no. 4, pp. 273–277, 2011.

[32] I. D. Federation, "IDF Diabetes Atlas, 7th Edition, 2015," International Diabetes Federation (IDF), Tech. Rep., 2015. [Online]. Available: www.diabetesatlas.org

[33] M. Heinrich, "Ethnopharmacy and natural product research – Multidisciplinary opportunities for research in the metabolomic age," Phytochemistry Letters, vol. 1, no. 1, pp. 1–5, apr 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1874390007000043

[34] D. C. Swinney and J. Anthony, "How were new medicines discovered?" Nature reviews. Drug discovery, vol. 10, no. 7, pp. 507–19, jun 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21701501

[35] F. Vincent, P. Loria, M. Pregel, R. Stanton, L. Kitching, K. Nocka, R. Doyonnas, C. Steppan, A. Gilbert, T. Schroeter, and M.-C. Peakman, "Developing predictive assays: the phenotypic screening "rule of 3"." Science translational medicine, vol. 7, no. 293, p. 293ps15, jun 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26109101

[36] J. G. Moffat, F. Vincent, J. A. Lee, J. Eder, and M. Prunotto, "Opportunities and challenges in phenotypic drug discovery: an industry perspective." Nature reviews. Drug discovery, vol. 16, no. 8, pp. 531–543, aug 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28685762

[37] S. I. Miller, "Antibiotic Resistance and Regulation of the Gram-Negative Bacterial Outer Membrane Barrier by Host Innate Immune Molecules." mBio, vol. 7, no. 5, pp.

e01 541–16, sep 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27677793

[38] M. A. Seeger and H. W. van Veen, "Molecular basis of multidrug transport by ABC transporters." Biochimica et biophysica acta, vol. 1794, no. 5, pp. 725–37, may 2009. [Online]. Available: http://dx.doi.org/10.1016/j.bbapap.2008.12.004

[39] J.-M. Pagès and L. Amaral, "Mechanisms of drug efflux and strategies to combat them: challenging the efflux pump of Gram-negative bacteria." Biochimica et biophysica acta, vol. 1794, no. 5, pp. 826–33, may 2009. [Online]. Available: http://dx.doi.org/10.1016/j.bbapap.2008.12.011

[40] J. M. Munita and C. A. Arias, "Mechanisms of Antibiotic Resistance," Virulence Mechanisms of Bacterial Pathogens, Fifth Edition, vol. 4, no. 2, pp. 481–511, 2016. [Online]. Available: http://www.asmscience.org/content/book/10.1128/9781555819286.chap17

[41] R. El-Awady, E. Saleh, A. Hashim, N. Soliman, A. Dallah, A. Elrasheed, and G. Elakraa, "The Role of Eukaryotic and Prokaryotic ABC Transporter Family in Failure of Chemotherapy." Frontiers in pharmacology, vol. 7, no. JAN, p. 535, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28119610

[42] J. Eder, R. Sedrani, and C. Wiesmann, "The discovery of first-in-class drugs: Origins and evolution," Nature Reviews Drug Discovery, vol. 13, no. 8, pp. 577–587, 2014. [Online]. Available: http://dx.doi.org/10.1038/nrd4336

[43] A. L. Hopkins and C. R. Groom, "The druggable genome." Nature reviews. Drug discovery, vol. 1, no. 9, pp. 727–30, 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12209152

[44] J. Inglese, R. L. Johnson, A. Simeonov, M. Xia, W. Zheng, C. P. Austin, and D. S. Auld, "High-throughput screening assays for the identification of chemical probes." Nature chemical biology, vol. 3, no. 8, pp. 466–79, aug 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17637779

[45] J. Zhang, P. L. Yang, and N. S. Gray, "Targeting cancer with small molecule kinase inhibitors." Nature reviews. Cancer, vol. 9, no. 1, pp. 28–39, jan 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19104514

[46] A. Nieto Gutierrez and P. H. McDonald, "GPCRs: Emerging anti-cancer drug targets." Cellular signalling, vol. 41, no. September 2017, pp. 65–74, jan 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28931490

[47] F. Reimann and F. M. Gribble, "G protein-coupled receptors as new therapeutic targets for type 2 diabetes." Diabetologia, vol. 59, no. 2, pp. 229–33, feb 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26661410

[48] J. Wölcke and D. Ullmann, "Miniaturized HTS technologies - uHTS." Drug discovery today, vol. 6, no. 12, pp. 637–646, jun 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11408200

[49] L. M. Mayr and D. Bojanic, "Novel trends in high-throughput screening." Current opinion in pharmacology, vol. 9, no. 5, pp. 580–8, oct 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19775937

[50] D. Hughes and A. Karlén, "Discovery and preclinical development of new antibiotics." Upsala journal of medical sciences, vol. 119, no. 2, pp. 162–9, may 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24646082

[51] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, and G. S. Sittampalam, "Impact of high-throughput screening in biomedical research." Nature reviews. Drug discovery, vol. 10, no. 3, pp. 188–95, 2011. [Online]. Available: http://dx.doi.org/10.1038/nrd3368

[52] C. M. Dobson, "Chemical space and biology," Nature, vol. 432, no. 7019, pp. 824–828, 2004. [Online]. Available: https://www.nature.com/articles/nature03192

[53] R. Macarron, "Critical review of the role of HTS in drug discovery." Drug discovery today, vol. 11, no. 7-8, pp. 277–9, apr 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16580969

[54] L. M. Mayr and P. Fuerst, "The future of high-throughput screening." Journal of biomolecular screening, vol. 13, no. 6, pp. 443–8, jul 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18660458

[55] P. Ettmeyer, R. Schnitzer, A. Bergner, and H. Nar, "Hit And Lead Generation Strategies," in Comprehensive Medicinal Chemistry III, 3rd ed., S. Chackalamannil, D. Rotella, and S. Ward, Eds. Elsevier Ltd, 2017, ch. 2.02, pp. 33–63. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124095472123133

[56] A. Mullard, "The phenotypic screening pendulum swings," Nature Reviews Drug Discovery, vol. 14, no. 12, pp. 807–809, 2015. [Online]. Available: http://dx.doi.org/10.1038/nrd4783

[57] D. G. I. Kingston, "Modern natural products drug discovery and its relevance to biodiversity conservation." Journal of natural products, vol. 74, no. 3, pp. 496–511, mar 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/np100550t

[58] R. Macarron, "Chemical libraries: How dark is HTS dark matter?" Nature chemical biology, vol. 11, no. 12, pp. 904–5, dec 2015. [Online]. Available: http://dx.doi.org/10.1038/nchembio.1937

[59] A. M. Wassermann, E. Lounkine, D. Hoepfner, G. Le Goff, F. J. King, C. Studer, J. M. Peltier, M. L. Grippo, V. Prindle, J. Tao, A. Schuffenhauer, I. M. Wallace, S. Chen, P. Krastel, A. Cobos-Correa, C. N. Parker, J. W. Davies, and M. Glick, "Dark chemical matter as a promising starting point for drug lead discovery." Nature chemical biology, vol. 11, no. 12, pp. 958–966, 2015. [Online]. Available: http://dx.doi.org/10.1038/nchembio.1936

[60] O. Sticher, "Natural product isolation," Natural Product Reports, vol. 25, no. 3, pp. 517–554, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18497897

[61] S. D. Sarker, Z. Latif, and A. I. Gray, Natural Products Isolation, 2nd ed., S. D. Sarker, Z. Latif, and A. I. Gray, Eds. Totowa, NJ: Humana Press, 2006. [Online]. Available: http://link.springer.com/10.1385/1592599559

[62] S. D. Sarker and L. Nahar, "An Introduction to Natural Products Isolation," in Natural Products Isolation. Methods in Molecular Biology (Methods and Protocols), 864th ed., S. D. Sarker and L. Nahar, Eds. Humana Press, 2012, pp. 1–25. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22367891

[63] F. Bucar, A. Wube, and M. Schmid, "Natural product isolation–how to get from biological material to pure compounds." Natural product reports, vol. 30, no. 4, pp. 525–45, apr 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23396532

[64] M. G. Weller, "A unifying review of bioassay-guided fractionation, effect-directed analysis and related techniques." Sensors (Basel, Switzerland), vol. 12, no. 7, pp. 9181–209, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23012539

[65] J. A. Beutler, Y. Kashman, M. Tischler, J. H. Cardellina, G. N. Gray, M. J. Currens, M. E. Wall, M. C. Wani, and M. R. Boyd, "A reinvestigation of Maprounea

triterpenes." Journal of natural products, vol. 58, no. 7, pp. 1039–46, jul 1995. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7561897

[66] B. Gilbert and L. F. Alves, "Synergy in plant medicines." Current medicinal chemistry, vol. 10, no. 1, pp. 13–20, jan 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12570718

[67] Y. Yang, Z. Zhang, S. Li, X. Ye, X. Li, and K. He, "Synergy effects of herb extracts: Pharmacokinetics and pharmacodynamic basis," Fitoterapia, vol. 92, pp. 133–147, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.fitote.2013.10.010

[68] H. A. Junio, A. A. Sy-Cordero, K. A. Ettefagh, J. T. Burns, K. T. Micko, T. N. Graf, S. J. Richter, R. E. Cannon, N. H. Oberlies, and N. B. Cech, "Synergy-directed fractionation of botanical medicines: a case study with goldenseal (Hydrastis canadensis)." Journal of natural products, vol. 74, no. 7, pp. 1621–9, jul 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21661731

[69] K. Hostettmann, A. Marston, J.-L. L. Wolfender, M. A. Hostettmann K., J.-L. L. Wolfender, K. Hostettmann, and A. Marston, "Strategy in the Search for New Lead Compounds and Drugs from Plants," CHIMIA International Journal for Chemistry, vol. 59, no. 6, pp. 291–294, jun 2005. [Online]. Available: http://www.ingentaselect.com/rpsv/cgi-bin/cgi?ini=xref{&}body=linker{&}reqdoi=10.2533/000942905777676326

[70] T. Elufioye, "Bioassay-coupled Chromatographic Analysis of Medicinal Natural Products: A Review," Tropical Journal of Natural Product Research, vol. 1, no. 3, pp. 100–104, sep 2017. [Online]. Available: https://www.tjnpr.org/viewarticle.aspx?articleid=64

[71] S. Bräm and E. Wolfram, "Recent Advances in Effect-directed Enzyme Assays based on Thin-layer Chromatography." Phytochemical analysis : PCA, vol. 28, no. 2, pp. 74–86, mar 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28146298

[72] L. H. Gu, L. P. Liao, H. J. Hu, S. W. Annie Bligh, C. H. Wang, G. X. Chou, and Z. T. Wang, "A thin-layer chromatography-bioautographic method for detecting dipeptidyl peptidase IV inhibitors in plants." Journal of chromatography. A, vol. 1411, pp. 116–22, sep 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26283532

[73] K. Kazafeos, "Incretin effect: GLP-1, GIP, DPP4." Diabetes research and clinical

practice, vol. 93 Suppl 1, no. SUPPL. 1, pp. S32–6, aug 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21864749

[74] S. Dewanjee, M. Gangopadhyay, N. Bhattacharya, R. Khanra, and T. K. Dua, "Bioautography and its scope in the field of natural product chemistry," Journal of Pharmaceutical Analysis, vol. 5, no. 2, pp. 75–84, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.jpha.2014.06.002

[75] A. Marston, "Thin-layer chromatography with biological detection in phytochemistry." Journal of chromatography. A, vol. 1218, no. 19, pp. 2676–83, may 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21236438

[76] D. van Elswijk and H. Irth, "Analytical tools for the detection and characterization of biologically active compounds from nature," Phytochemistry Reviews, vol. 1, no. 3, pp. 427–439, oct 2002. [Online]. Available: http://link.springer.com/10.1023/A:1026089809218

[77] D. A. Van Elswijk, U. P. Schobel, E. P. Lansky, H. Irth, and J. Van Der Greef, "Rapid dereplication of estrogenic compounds in pomegranate (Punica granatum) using on-line biochemical detection coupled to mass spectrometry," Phytochemistry, vol. 65, no. 2, pp. 233–241, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14732284

[78] A. R. de Boer, T. Letzel, D. A. van Elswijk, H. Lingeman, W. M. A. Niessen, and H. Irth, "On-line coupling of high-performance liquid chromatography to a continuous-flow enzyme assay based on electrospray ionization mass spectrometry." Analytical chemistry, vol. 76, no. 11, pp. 3155–61, jun 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15167796

[79] N. Aggarwal and B. F. Sloane, "Cathepsin B: multiple roles in cancer." Proteomics. Clinical applications, vol. 8, no. 5-6, pp. 427–37, jun 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24677670

[80] A. Liesener and U. Karst, "Monitoring enzymatic conversions by mass spectrometry: a critical review." Analytical and bioanalytical chemistry, vol. 382, no. 7, pp. 1451–64, aug 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16007447

[81] H. Bisswanger, "Enzyme assays," Perspectives in Science, vol. 1, no. 1-6, pp. 41–55, may 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S2213020914000068

[82] J. Bitzer, B. Köpcke, M. Stadler, V. Hellwig, Y.-M. Ju, S. Seip, and T. Henkel, "Accelerated Dereplication of Natural Products, Supported by Reference Libraries," CHIMIA International Journal for Chemistry, vol. 61, no. 6, pp. 332–338, jun 2007. [Online]. Available: http://www.ingentaconnect.com/content/scs/chimia/2007/00000061/00000006/art00005

[83] J. Hubert, J.-M. Nuzillard, and J.-H. Renault, "Dereplication strategies in natural product research: How many tools and methodologies behind the same concept?" Phytochemistry Reviews, vol. 16, no. 1, pp. 55–95, feb 2017. [Online]. Available: https://doi.org/10.1007/s11101-015-9448-7

[84] D. G. Corley and R. C. Durley, "Strategies for Database Dereplication of Natural Products," Journal of Natural Products, vol. 57, no. 11, pp. 1484–1490, nov 1994. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/np50113a002

[85] J.-L. Wolfender, G. Marti, A. Thomas, and S. Bertrand, "Current approaches and challenges for the metabolite profiling of complex natural extracts." Journal of chromatography. A, vol. 1382, pp. 136–64, feb 2015. [Online]. Available: http://dx.doi.org/10.1016/j.chroma.2014.10.091

[86] S. P. Gaudêncio and F. Pereira, "Dereplication: racing to speed up the natural products discovery process." Natural product reports, vol. 32, no. 6, pp. 779–810, jun 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25850681

[87] A. Aharoni, C. H. Ric de Vos, H. A. Verhoeven, C. A. Maliepaard, G. Kruppa, R. Bino, and D. B. Goodenowe, "Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry." Omics : a journal of integrative biology, vol. 6, no. 3, pp. 217–34, jan 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12427274

[88] T. Kind and O. Fiehn, "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry." BMC bioinformatics, vol. 8, p. 105, mar 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17389044

[89] S. Bertrand, Y. Guitton, and C. Roullier, "Successes and pitfalls in automated dereplication strategy using liquid chromatography coupled to mass spectrometry data: A CASMI 2016 experience," Phytochemistry Letters, vol. 21, pp. 297–305, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.phytol.2016.12.025

[90] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, and O. Yanes, "Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects," TrAC - Trends in Analytical Chemistry, vol. 78, pp. 23–35, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.trac.2015.09.005

[91] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak, "METLIN: a metabolite mass spectral database." Therapeutic drug monitoring, vol. 27, no. 6, pp. 747–51, dec 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16404815

[92] Y. Sawada, R. Nakabayashi, Y. Yamada, M. Suzuki, M. Sato, A. Sakata, K. Akiyama, T. Sakurai, F. Matsuda, T. Aoki, M. Y. Hirai, and K. Saito, "RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database." Phytochemistry, vol. 82, pp. 38–45, oct 2012. [Online]. Available: http://dx.doi.org/10.1016/j.phytochem.2012.07.007

[93] R. Jansen, G. Lachatre, and P. Marquet, "LC-MS/MS systematic toxicological analysis: comparison of MS/MS spectra obtained with different instruments and settings." Clinical biochemistry, vol. 38, no. 4, pp. 362–72, apr 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15766737

[94] J.-l. Wolfender, S. Rudaz, Y. H. Choi, and H. K. Kim, "Plant metabolomics: from holistic data to relevant biomarkers." Current medicinal chemistry, vol. 20, no. 8, pp. 1056–90, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23210790

[95] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler, "On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study." Journal of mass spectrometry : JMS, vol. 44, no. 4, pp. 485–93, apr 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19165818

[96] ——, "On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm." Journal of mass spectrometry : JMS, vol. 44, no. 4, pp. 494–502, apr 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19152368

[97] H. Oberacher, F. Pitterl, E. Siapi, B. R. Steele, T. Letzel, S. Grosse, B. Poschner, F. Tagliaro, R. Gottardo, S. A. Chacko, and J. L. Josephs, "On the inter-instrument and the inter-laboratory transferability of a tandem mass

spectral reference library. 3. Focus on ion trap and upfront CID." Journal of mass spectrometry : JMS, vol. 47, no. 2, pp. 263–70, feb 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22359338

[98] X. Yang, P. Neta, and S. E. Stein, "Quality control for building libraries from electrospray ionization tandem mass spectra." Analytical chemistry, vol. 86, no. 13, pp. 6393–400, jul 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24896981

[99] T. Kind and O. Fiehn, "Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry." Phytochemistry letters, vol. 21, pp. 313–319, sep 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29225718

[100] G. J. Patti, O. Yanes, and G. Siuzdak, "Innovation: Metabolomics: the apogee of the omics trilogy." Nature reviews. Molecular cell biology, vol. 13, no. 4, pp. 263–9, mar 2012. [Online]. Available: http://dx.doi.org/10.1038/nrm3314

[101] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," Analytica Chimica Acta, vol. 185, pp. 1–17, 1986. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0003267086800289

[102] A. Höskuldsson, "PLS regression methods," Journal of Chemometrics, vol. 2, no. 3, pp. 211–228, jun 1988. [Online]. Available: http://doi.wiley.com/10.1002/cem.1180020306

[103] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," Chemometrics and Intelligent Laboratory Systems, vol. 58, no. 2, pp. 109–130, oct 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743901001551

[104] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data." Briefings in bioinformatics, vol. 8, no. 1, pp. 32–44, jan 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16772269

[105] V. N. Vapnik, Statistical Learning Theory. New York, NY: John Wiley & Sons, Ltd., 1998, vol. 2. [Online]. Available: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471030031.html

[106] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky, "Analysis of metabolomic data using support vector machines." Analytical chemistry, vol. 80, no. 19, pp. 7562–70, oct 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18767870

[107] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/article/10.1023/A:1010933404324

[108] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Boston: Addison-Wesley, 1989. [Online]. Available: https://dl.acm.org/citation.cfm?id=534133

[109] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, and D. B. Kell, "Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry," Analytica Chimica Acta, vol. 348, no. 1-3, pp. 71–86, aug 1997. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0003267097000652

[110] M. Brown, W. B. Dunn, D. I. Ellis, R. Goodacre, J. Handl, J. D. Knowles, S. O'Hagan, I. Spasić, and D. B. Kell, "A metabolome pipeline: from concept to data to knowledge," Metabolomics, vol. 1, no. 1, pp. 39–51, mar 2005. [Online]. Available: http://link.springer.com/10.1007/s11306-005-1106-4

[111] E. Saccenti, H. C. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. Hendriks, "Reflections on univariate and multivariate analysis of metabolomics data," Metabolomics, vol. 10, no. 3, pp. 361–374, 2014. [Online]. Available: https://link.springer.com/article/10.1007/s11306-013-0598-6

[112] R. L. Last, A. D. Jones, and Y. Shachar-Hill, "Towards the plant metabolome and beyond." Nature reviews. Molecular cell biology, vol. 8, no. 2, pp. 167–74, feb 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17213843

[113] N. D. Yuliana, A. Khatib, Y. H. Choi, and R. Verpoorte, "Metabolomics for bioactivity assessment of natural products." Phytotherapy research: PTR, vol. 25, no. 2, pp. 157–69, feb 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20658470

[114] E. K. Kemsley, G. Le Gall, J. R. Dainty, A. D. Watson, L. J. Harvey, H. S. Tapp, and I. J. Colquhoun, "Multivariate techniques and their application in nutrition: a metabolomics case study." The British journal of nutrition, vol. 98, no. 1, pp. 1–14, jul 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17381968

[115] M. Amargianitaki and A. Spyros, "NMR-based metabolomics in wine quality control and authentication," Chemical and Biological Technologies in Agriculture, vol. 4, no. 1, pp. 1–12, 2017. [Online]. Available: https://link.springer.com/article/10.1186/s40538-017-0092-x

[116] D. I. Broadhurst and D. B. Kell, "Statistical strategies for avoiding false discoveries in metabolomics and related experiments," Metabolomics, vol. 2, no. 4, pp. 171–196, nov 2006. [Online]. Available: http://link.springer.com/10.1007/s11306-006-0037-z

[117] P. S. Gromski, Y. Xu, E. Correa, D. I. Ellis, M. L. Turner, and R. Goodacre, "A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data." Analytica chimica acta, vol. 829, pp. 1–8, jun 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24856395

[118] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, and R. Goodacre, "A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding." Analytica chimica acta, vol. 879, pp. 10–23, jun 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26002472

[119] S. Ren, A. A. Hinzman, E. L. Kang, R. D. Szczesniak, and L. J. Lu, "Computational and statistical analysis of metabolomics data," Metabolomics, vol. 11, no. 6, pp. 1492–1513, 2015. [Online]. Available: https://link.springer.com/article/10.1007/s11306-015-0823-6

[120] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, and Y. Liang, "Chemometric methods in data processing of mass spectrometry-based metabolomics: A review." Analytica chimica acta, vol. 914, pp. 17–34, mar 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26965324

[121] M. Karelson, V. S. Lobanov, and A. R. Katritzky, "Quantum-Chemical Descriptors in QSAR/QSPR Studies," Chemical Reviews, vol. 96, no. 3, pp. 1027–1044, 1996. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/cr950202r

[122] A. Z. Dudek, T. Arodz, and J. Gálvez, "Computational methods in developing quantitative structure-activity relationships (QSAR): a review." Combinatorial chemistry & high throughput screening, vol. 9, no. 3, pp. 213–28, mar 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16533155

[123] H. Xu, Z. Liu, W. Cai, and X. Shao, "A wavelength selection method based on randomization test for near-infrared spectral analysis," Chemometrics and Intelligent Laboratory Systems, vol. 97, no. 2, pp. 189–193, jul 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743909000951

[124] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences, vol. 43, no. 6, pp. 1947–58, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14632445

[125] A. Liaw and V. Svetnik, "QSAR modeling: prediction of biological activity from chemical structure," in Statistical Methods for Evaluating Safety in Medical Product Development, A. Lawrence Gould., Ed. Chichester, UK: John Wiley & Sons, Ltd, dec 2014, pp. 66–83. [Online]. Available: http://doi.wiley.com/10.1002/9781118763070.ch3

[126] X.-W. Zhu, Y.-J. Xin, and H.-L. Ge, "Recursive Random Forests Enable Better Predictive Performance and Model Interpretation than Variable Selection by LASSO." Journal of chemical information and modeling, vol. 55, no. 4, pp. 736–46, apr 2015. [Online]. Available: http://dx.doi.org/10.1021/ci500715e

[127] M. K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, and C. Lemmen, "Active learning with support vector machines in the drug discovery process." Journal of chemical information and computer sciences, vol. 43, no. 2, pp. 667–73, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12653536

[128] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, H. Ishwaran, K. Knight, J. M. Loubes, P. Massart, D. Madigan, G. Ridgeway, S. Rosset, J. I. Zhu, R. A. Stine, B. A. Turlach, S. Weisberg, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," Annals of Statistics, vol. 32, no. 2, pp. 407–499, 2004. [Online]. Available: https://www.jstor.org/stable/3448465

[129] Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih, and M. Aziz, "High-dimensional QSAR modelling using penalized linear regression model with L1/2 -norm," SAR and QSAR in Environmental Research, vol. 27, no. 9, pp. 1–17, 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1062936X.2016.1228696

[130] M. Goodarzi, B. Dejaegher, and Y. Vander Heyden, "Feature selection methods in QSAR studies." Journal of AOAC International, vol. 95, no. 3, pp. 636–51, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22816254

[131] G. Ghasemi, S. Arshadi, A. N. Rashtehroodi, M. Nirouei, S. Shariati, and Z. Rastgoo, "QSAR Investigation on Quinolizidinyl Derivatives in Alzheimer's Disease," Journal

of Computational Medicine, vol. 2013, no. Article ID 312728, pp. 1–8, 2013. [Online]. Available: http://www.hindawi.com/archive/2013/312728/

[132] M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, "Benchmarking Variable Selection in QSAR." Molecular informatics, vol. 31, no. 2, pp. 173–9, feb 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27476962

[133] ——, "Choosing feature selection and learning algorithms in QSAR." Journal of chemical information and modeling, vol. 54, no. 3, pp. 837–43, mar 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24460242

[134] Y. Cheng, Y. Wang, and X. Wang, "A causal relationship discovery-based approach to identifying active components of herbal medicine." Computational biology and chemistry, vol. 30, no. 2, pp. 148–54, apr 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16542877

[135] Y. Wang, X. Wang, and Y. Cheng, "A computational approach to botanical drug design by modeling quantitative composition-activity relationship." Chemical biology & drug design, vol. 68, no. 3, pp. 166–72, sep 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17062014

[136] N. J. C. Bailey, Y. Wang, J. Sampson, W. Davis, I. Whitcombe, P. J. Hylands, S. L. Croft, and E. Holmes, "Prediction of anti-plasmodial activity of Artemisia annua extracts: application of 1H NMR spectroscopy and chemometrics." Journal of pharmaceutical and biomedical analysis, vol. 35, no. 1, pp. 117–26, apr 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15030886

[137] A. T. Cardoso-Taketa, R. Pereda-Miranda, H. C. Young, R. Verpoorte, and M. L. Villarreal, "Metabolic profiling of the Mexican anxiolytic and sedative plant Galphimia glauca using nuclear magnetic resonance spectroscopy and multivariate data analysis," Planta Medica, vol. 74, no. 10, pp. 1295–1301, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18612944

[138] M. Dumarey, A. van Nederkassel, E. Deconinck, and Y. Vander Heyden, "Exploration of linear multivariate calibration techniques to predict the total antioxidant capacity of green tea from chromatographic fingerprints." Journal of chromatography. A, vol. 1192, no. 1, pp. 81–8, may 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18395730

[139] N. Nguyen Hoai, B. Dejaegher, C. Tistaert, V. Nguyen Thi Hong, C. Rivière, G. Chataigné, K. Phan Van, M. Chau Van, J. Quetin-Leclercq, and Y. Vander

Heyden, "Development of HPLC fingerprints for Mallotus species extracts and evaluation of the peaks responsible for their antioxidant activity." Journal of pharmaceutical and biomedical analysis, vol. 50, no. 5, pp. 753–63, dec 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19586736

[140] T. Rajalahti and O. M. Kvalheim, "Multivariate data analysis in pharmaceutics: a tutorial review." International journal of pharmaceutics, vol. 417, no. 1-2, pp. 280–90, sep 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21335075

[141] J. Trygg and S. Wold, "Orthogonal projections to latent structures (O-PLS)," Journal of Chemometrics, vol. 16, no. 3, pp. 119–128, mar 2002. [Online]. Available: http://doi.wiley.com/10.1002/cem.695

[142] C. Tistaert, G. Chataigné, B. Dejaegher, C. Rivière, N. Nguyen Hoai, M. Chau Van, J. Quetin-Leclercq, and Y. Vander Heyden, "Multivariate data analysis to evaluate the fingerprint peaks responsible for the cytotoxic activity of Mallotus species," Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, vol. 910, pp. 103–113, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.jchromb.2012.10.001

[143] J. Xia, I. V. Sinelnikov, B. Han, and D. S. Wishart, "MetaboAnalyst 3.0–making metabolomics more meaningful." Nucleic acids research, vol. 43, no. W1, pp. W251–7, jul 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25897128

[144] J. Xia and D. S. Wishart, "Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis." Current protocols in bioinformatics, vol. 55, no. September, pp. 14.10.1–14.10.91, sep 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27603023

[145] M. N. Triba, L. Le Moyec, R. Amathieu, C. Goossens, N. Bouchemal, P. Nahon, D. N. Rutledge, and P. Savarin, "PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters." Molecular bioSystems, vol. 11, no. 1, pp. 13–9, jan 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25382277

[146] N. D. Yuliana, A. Khatib, R. Verpoorte, and Y. H. Choi, "Comprehensive extraction method integrated with NMR metabolomics: a new bioactivity screening method for plants, adenosine A1 receptor binding compounds in Orthosiphon stamineus Benth." Analytical chemistry, vol. 83, no. 17, pp. 6902–6, sep 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21800886

[147] N. D. Yuliana, S. Budijanto, R. Verpoorte, and Y. H. Choi, "NMR metabolomics for identification of adenosine A1 receptor binding compounds from Boesenbergia rotunda rhizomes extract." Journal of ethnopharmacology, vol. 150, no. 1, pp. 95–9, oct 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jep.2013.08.012

[148] Y. Fujimura, K. Kurihara, M. Ida, R. Kosaka, D. Miura, H. Wariishi, M. Maeda-Yamamoto, A. Nesumi, T. Saito, T. Kanda, K. Yamada, and H. Tachibana, "Metabolomics–driven nutraceutical evaluation of diverse green tea cultivars." PloS one, vol. 6, no. 8, p. e23426, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21853132

[149] D. M. Kulakowski, S.-B. Wu, M. J. Balick, and E. J. Kennelly, "Merging bioactivity with liquid chromatography-mass spectrometry-based chemometrics to identify minor immunomodulatory compounds from a Micronesian adaptogen, Phaleria nisidai." Journal of chromatography. A, vol. 1364, pp. 74–82, oct 2014. [Online]. Available: http://dx.doi.org/10.1016/j.chroma.2014.08.049

[150] J. Maree, G. Kamatou, S. Gibbons, A. Viljoen, and S. Van Vuuren, "The application of GC-MS combined with chemometrics for the identification of antimicrobial compounds from selected commercial essential oils," Chemometrics and Intelligent Laboratory Systems, vol. 130, pp. 172–181, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.chemolab.2013.11.004

[151] S. Wold, E. Johansson, and M. Cocchi, "PLS: Partial Least Squares Projections to Latent Structures, 3D QSAR in drug design," in 3D QSAR in Drug Design, Volume 1: Theory Methods and Applications, 1st ed., H. Kubinyi, Ed., 1993, pp. 523–550.

[152] S. Wiklund, E. Johansson, L. Sjöström, E. J. Mellerowicz, U. Edlund, J. P. Shockcor, J. Gottfries, T. Moritz, and J. Trygg, "Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models." Analytical chemistry, vol. 80, no. 1, pp. 115–22, jan 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18027910

[153] O. M. Kvalheim and T. V. Karstang, "Interpretation of latent-variable regression models," Chemometrics and Intelligent Laboratory Systems, vol. 7, no. 1-2, pp. 39–51, dec 1989. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0169743989801108

[154] O. M. Kvalheim, "Latent-variable regression models with higher-order terms: An extension of response modelling by orthogonal design and multiple linear regression,"

Chemometrics and Intelligent Laboratory Systems, vol. 8, no. 1, pp. 59–67, may 1990. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0169743990800414

[155] R. Ergon, "PLS post-processing by similarity transformation (PLS + ST): A simple alternative to OPLS," Journal of Chemometrics, vol. 19, no. 1, pp. 1–4, 2005. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/cem.899/abstract

[156] O. M. Kvalheim, T. Rajalahti, and R. Arneberg, "X-tended target projection (XTP)-comparison with orthogonal partial least squares (OPLS) and PLS post-processing by similarity transformation," Journal of Chemometrics, vol. 23, no. 1, pp. 49–55, jan 2009. [Online]. Available: http://doi.wiley.com/10.1002/cem.1193

[157] O. M. Kvalheim, "Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots," Journal of Chemometrics, vol. 24, no. 7-8, pp. 496–504, jul 2010. [Online]. Available: http://doi.wiley.com/10.1002/cem.1289

[158] T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik, and O. M. Kvalheim, "Biomarker discovery in mass spectral profiles by means of selectivity ratio plot," Chemometrics and Intelligent Laboratory Systems, vol. 95, no. 1, pp. 35–48, jan 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743908001548

[159] T. Rajalahti, R. Arneberg, A. C. Kroksveen, M. Berle, K.-M. Myhr, and O. M. Kvalheim, "Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles," Analytical Chemistry, vol. 81, no. 7, pp. 2581–2590, apr 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19228047

[160] F.-T. Chau, H.-Y. Chan, C.-Y. Cheung, C.-J. Xu, Y. Liang, and O. M. Kvalheim, "Recipe for uncovering the bioactive components in herbal medicine." Analytical chemistry, vol. 81, no. 17, pp. 7217–25, sep 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19634860

[161] O. M. Kvalheim, H.-Y. Chan, I. F. Benzie, Y.-T. Szeto, A. H.-C. Tzang, D. K.-W. Mok, and F.-T. Chau, "Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products," Chemometrics and Intelligent Laboratory Systems, vol. 107, no. 1, pp. 98–105, may 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743911000189

[162] J. J. Kellogg, D. A. Todd, J. M. Egan, H. A. Raja, N. H. Oberlies, O. M. Kvalheim, and N. B. Cech, "Biochemometrics for Natural Products Research: Comparison of Data Analysis Approaches and Application to Identification of Bioactive Compounds." Journal of natural products, vol. 79, no. 2, pp. 376–86, feb 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26841051

[163] E. R. Britton, J. J. Kellogg, O. M. Kvalheim, and N. B. Cech, "Biochemometrics to Identify Synergists and Additives from Botanical Medicines: A Case Study with Hydrastis canadensis (Goldenseal)." Journal of natural products, p. acs.jnatprod.7b00654, nov 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29091439

[164] C. Tistaert, B. Dejaegher, G. Chataigné, C. Rivière, N. N. Hoai, M. C. Van, J. Quetin-Leclercq, and Y. V. Heyden, "Potential antioxidant compounds in Mallotus species fingerprints. Part II: fingerprint alignment, data analysis and peak identification." Analytica chimica acta, vol. 721, pp. 35–43, apr 2012. [Online]. Available: http://dx.doi.org/10.1016/j.aca.2012.01.058

[165] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634

[166] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. Mcclean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance," Ecography, vol. 36, no. 1, pp. 027–046, 2013. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2012.07348.x/full

[167] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267–288, nov 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[168] A. Gordinsky, "New Facts in Regression Estimation under Conditions of Multicollinearity," Open Journal of Statistics, vol. 06, no. 05, pp. 842–861, 2016. [Online]. Available: http://www.scirp.org/journal/PaperDownload.aspx?DOI=10.4236/ojs.2016.65070

[169] J. Engel, L. Buydens, and L. Blanchet, "An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics,"

Journal of Chemometrics, vol. 31, no. 4, p. e2880, apr 2017. [Online]. Available: http://doi.wiley.com/10.1002/cem.2880

[170] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics." Statistical applications in genetics and molecular biology, vol. 4, no. 1, p. Article32, 2005. [Online]. Available: http://ideas.repec.org/a/bpj/sagmbi/v4y2005i1n32.htmlhttp://www.ncbi.nlm.nih.gov/pubmed/16646851

[171] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho, "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions." BMC proceedings, vol. 6 Suppl 2, no. Suppl 2, p. S10, may 2012. [Online]. Available: http://www.biomedcentral.com/1753-6561/6/S2/S10

[172] Z. M. Hira, D. F. Gillies, Z. M. Hira, and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," Advances in Bioinformatics, vol. 2015, no. 1, pp. 1–13, 2015. [Online]. Available: http://www.hindawi.com/journals/abi/2015/198363/

[173] S. Mittal, K. Arora, A. R. Rao, M. G. Mallikarjuna, H. S. Gupta, and N. Thirunavukkarasu, "Genomic selection for drought tolerance using genome-wide SNPs in maize," Frontiers in plant science, vol. 8, no. April, pp. 1–12, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5399777/

[174] S. Wahl, C. Holzapfel, Z. Yu, M. Breier, I. Kondofersky, C. Fuchs, P. Singmann, C. Prehn, J. Adamski, H. Grallert, T. Illig, R. Wang-Sattler, and T. Reinehr, "Metabolomics reveals determinants of weight loss during lifestyle intervention in obese children," Metabolomics, vol. 9, no. 6, pp. 1157–1167, 2013. [Online]. Available: https://link.springer.com/article/10.1007/s11306-013-0550-9

[175] F. Z. Zhang and D. Hong, "Elastic net-based framework for imaging mass spectrometry data biomarker selection and classification." Statistics in medicine, vol. 30, no. 7, pp. 753–68, mar 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21394751

[176] J. A. Westerhuis, H. C. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. Velzen, J. P. Duijnhoven, and F. A. Dorsten, "Assessment of PLSDA cross validation," Metabolomics, vol. 4, no. 1, pp. 81–89, 2008.

[177] K. H. Esbensen and P. Geladi, "Principles of proper validation: Use and abuse of re-sampling for validation," Journal of Chemometrics, vol. 24, no. 3-4, pp. 168–187,

2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/cem.1310/abstract

[178] B. Worley and R. Powers, "PCA as a Practical Indicator of OPLS-DA Model Reliability," Current Metabolomics, vol. 4, no. 2, pp. 97–103, jun 2016. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4990351

[179] R. G. Brereton, "Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data," TrAC - Trends in Analytical Chemistry, vol. 25, no. 11, pp. 1103–1111, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165993606002330

[180] E. Anderssen, K. Dyrstad, F. Westad, and H. Martens, "Reducing over-optimism in variable selection by cross-model validation," Chemometrics and Intelligent Laboratory Systems, vol. 84, no. 1-2, pp. 69–74, dec 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743906001109

[181] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: Taking the magic away," Journal of Chemometrics, vol. 28, no. 4, pp. 213–225, 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/cem.2609/abstract

[182] H. Brückner, J. Maisch, C. Reinecke, and A. Kimonyo, "Use of alpha-aminoisobutyric acid and isovaline as marker amino acids for the detection of fungal polypeptide antibiotics. Screening of Hypocrea," Amino Acids, vol. 1, no. 2, pp. 251–257, 1991.

[183] T. Degenkolb, J. Kirschbaum, and H. Brückner, "New sequences, constituents, and producers of peptaibiotics: an updated review." Chemistry & biodiversity, vol. 4, no. 6, pp. 1052–67, jun 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17589876

[184] C. Toniolo, M. Crisma, F. Formaggio, C. Peggion, R. F. Epand, and R. M. Epand, "Lipopeptaibols, a novel family of membrane active, antimicrobial peptides." Cellular and molecular life sciences : CMLS, vol. 58, no. 9, pp. 1179–88, aug 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11577977

[185] T. Degenkolb, T. Gräfenhan, A. Berg, H. I. Nirenberg, W. Gams, and H. Brückner, "Peptaibiomics: Screening for polypeptide antibiotics (peptaibiotics) from plant-protective Trichoderma species." Chemistry & biodiversity, vol. 3, no. 6, pp. 593–610, jun 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17193294

[186] T. Degenkolb, A. Berg, W. Gams, B. Schlegel, and U. Gräfe, "The occurrence of peptaibols and structurally related peptaibiotics in fungi and their mass spectrometric identification via diagnostic fragment ions." Journal of peptide science : an official publication of the European Peptide Society, vol. 9, no. 11-12, pp. 666–78, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14658788

[187] J. K. Chugh and B. Wallace, "Peptaibols: models for ion channels." Biochemical Society transactions, vol. 29, no. 4, pp. 565–70, aug 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11498029

[188] C. E. Meyer and F. Reusser, "A polypeptide antibacterial agent isolated from Trichoderma viride." Experientia, vol. 23, no. 2, pp. 85–6, feb 1967. [Online]. Available: http://link.springer.com/10.1007/BF02135929

[189] N. K. N. Neumann, N. Stoppacher, S. Zeilinger, T. Degenkolb, H. Brückner, and R. Schuhmacher, "The peptaibiotics database–a comprehensive online resource." Chemistry & biodiversity, vol. 12, no. 5, pp. 743–51, may 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26010663

[190] C. Krause, J. Kirschbaum, and H. Brückner, "Peptaibiomics: an advanced, rapid and selective analysis of peptaibiotics/peptaibols by SPE/LC-ES-MS." Amino acids, vol. 30, no. 4, pp. 435–43, jun 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16622603

[191] C. R. Röhrich, A. Iversen, W. M. Jaklitsch, H. Voglmayr, A. Vilcinskas, K. F. Nielsen, U. Thrane, H. von Döhren, H. Brückner, and T. Degenkolb, "Screening the biosphere: the fungicolous fungus Trichoderma phellinicola, a prolific source of hypophellins, new 17-, 18-, 19-, and 20-residue peptaibiotics." Chemistry & biodiversity, vol. 10, no. 5, pp. 787–812, may 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23681726

[192] P. W. Crous, W. Gams, J. A. Stalpers, V. Robert, G. Stegehuis, C. Schimmelcultures, P. O. Box, and A. D. Utrecht, "MycoBank : an online initiative to launch mycology into the 21st century," Studies in Mycology, vol. 50, pp. 19–22, 2004. [Online]. Available: http://www.westerdijkinstitute.nl/images/ResearchGroups/Phytopathology/pdf/PDFOPNUMMER/203.pdf

[193] V. Robert, D. Vu, A. B. H. Amor, N. van de Wiele, C. Brouwer, B. Jabas, S. Szoke, A. Dridi, M. Triki, S. Ben Daoud, O. Chouchen, L. Vaas, A. de Cock, J. a. Stalpers, D. Stalpers, G. J. M. Verkley, M. Groenewald, F. B. Dos

Santos, G. Stegehuis, W. Li, L. Wu, R. Zhang, J. Ma, M. Zhou, S. P. Gorjón, L. Eurwilaichitr, S. Ingsriswang, K. Hansen, C. Schoch, B. Robbertse, L. Irinyi, W. Meyer, G. Cardinali, D. L. Hawksworth, J. W. Taylor, and P. W. Crous, "MycoBank gearing up for new horizons." IMA fungus, vol. 4, no. 2, pp. 371–9, dec 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24563843

[194] I. L. Karle and P. Balaram, "Structural characteristics of alpha-helical peptide molecules containing Aib residues." Biochemistry, vol. 29, no. 29, pp. 6747–56, jul 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2204420

[195] M. S. Sansom, "Alamethicin and related peptaibols–model ion channels." European biophysics journal : EBJ, vol. 22, no. 2, pp. 105–24, 1993. [Online]. Available: https://link.springer.com/article/10.1007/BF00196915

[196] R. Anders, O. Ohlenschläger, V. Soskic, H. Wenschuh, B. Heise, and L. R. Brown, "The NMR solution structure of the ion channel peptaibol chrysospermin C bound to dodecylphosphocholine micelles." European journal of biochemistry, vol. 267, no. 6, pp. 1784–94, mar 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10712611

[197] M. Kronen, H. Görls, H.-H. Nguyen, S. Reissmann, M. Bohl, J. Sühnel, and U. Gräfe, "Crystal structure and conformational analysis of ampullosporin A." Journal of peptide science : an official publication of the European Peptide Society, vol. 9, no. 11-12, pp. 729–44, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14658792

[198] E. S. Salnikov, H. Friedrich, X. Li, P. Bertani, S. Reissmann, C. Hertweck, J. D. J. O'Neil, J. Raap, and B. Bechinger, "Structure and alignment of the membrane-associated peptaibols ampullosporin A and alamethicin by oriented 15N and 31P solid-state NMR spectroscopy." Biophysical journal, vol. 96, no. 1, pp. 86–100, jan 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18835909

[199] S. Aravinda, N. Shamala, and P. Balaram, "Aib residues in peptaibiotics and synthetic sequences: analysis of nonhelical conformations." Chemistry & biodiversity, vol. 5, no. 7, pp. 1238–62, jul 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18649312

[200] A. D. Milov, Y. D. Tsvetkov, J. Raap, M. De Zotti, F. Formaggio, and C. Toniolo, "Conformation, self-aggregation, and membrane interaction of peptaibols as studied by pulsed electron double resonance spectroscopy."

Biopolymers, vol. 106, no. 1, pp. 6–24, jan 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26270729

[201] H. Sato and J. B. Feix, "Peptide-membrane interactions and mechanisms of membrane destruction by amphipathic alpha-helical antimicrobial peptides." Biochimica et biophysica acta, vol. 1758, no. 9, pp. 1245–56, sep 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16697975

[202] H. W. Huang, "Molecular mechanism of antimicrobial peptides: The origin of cooperativity," Biochimica et Biophysica Acta-Biomembranes, vol. 1758, no. 9, pp. 1292–1302, sep 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16542637

[203] R. O. Fox and F. M. Richards, "A voltage-gated ion channel model inferred from the crystal structure of alamethicin at 1.5-A resolution." Nature, vol. 300, no. 5890, pp. 325–30, nov 1982. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6292726

[204] P. Pieta, J. Mirza, and J. Lipkowski, "Direct visualization of the alamethicin pore formed in a planar phospholipid matrix." Proceedings of the National Academy of Sciences of the United States of America, vol. 109, no. 52, pp. 21 223–7, dec 2012. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1201559110

[205] F. Abbasi, J. Jay Leitch, Z. Su, G. Szymanski, and J. Lipkowski, "Direct visualization of alamethicin ion pores formed in a floating phospholipid membrane supported on a gold electrode surface," Electrochimica Acta, feb 2018. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0013468618303396

[206] G. Bocchinfuso, A. Palleschi, B. Orioni, G. Grande, F. Formaggio, C. Toniolo, Y. Park, K.-S. Hahm, and L. Stella, "Different mechanisms of action of antimicrobial peptides: insights from fluorescence spectroscopy experiments and molecular dynamics simulations." Journal of peptide science : an official publication of the European Peptide Society, vol. 15, no. 9, pp. 550–8, sep 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19455510

[207] S. Iftemi, M. De Zotti, F. Formaggio, C. Toniolo, L. Stella, and T. Luchian, "Electrophysiology investigation of Trichogin GA IV activity in planar lipid membranes reveals ion channels of well-defined size." Chemistry & biodiversity, vol. 11, no. 7, pp. 1069–77, jul 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25044592

[208] Z. O. Shenkarev, A. S. Paramonov, E. N. Lyukmanova, A. K. Gizatullina, A. V. Zhuravleva, A. A. Tagaev, Z. A. Yakimenko, I. N. Telezhinskaya, M. P. Kirpichnikov, T. V. Ovchinnikova, and A. S. Arseniev, "Peptaibol antiamoebin I: spatial structure, backbone dynamics, interaction with bicelles and lipid-protein nanodiscs, and pore formation in context of barrel-stave model." Chemistry & biodiversity, vol. 10, no. 5, pp. 838–63, may 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23681729

[209] D. Sengupta, H. Leontiadou, A. E. Mark, and S.-J. Marrink, "Toroidal pores formed by antimicrobial peptides show significant disorder." Biochimica et biophysica acta, vol. 1778, no. 10, pp. 2308–17, oct 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18602889

[210] L. Yang, T. A. Harroun, T. M. Weiss, L. Ding, and H. W. Huang, "Barrel-stave model or toroidal model? A case study on melittin pores." Biophysical journal, vol. 81, no. 3, pp. 1475–85, sep 2001. [Online]. Available: http://www.cell.com/biophysj/fulltext/S0006-3495(01)75802-X

[211] S. J. Ludtke, K. He, W. T. Heller, T. A. Harroun, L. Yang, and H. W. Huang, "Membrane pores induced by magainin." Biochemistry, vol. 35, no. 43, pp. 13 723–8, oct 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8901513

[212] Y. Huang, J. Huang, and Y. Chen, "Alpha-helical cationic antimicrobial peptides: relationships of structure and function." Protein & cell, vol. 1, no. 2, pp. 143–52, feb 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21203984

[213] H. F. Christoffersen, S. K. Hansen, B. S. Vad, E. H. Nielsen, J. T. Nielsen, T. Vosegaard, T. Skrydstrup, and D. E. Otzen, "The natural, peptaibolic peptide SPF-5506-A4 adopts a beta-bend spiral structure, shows low hemolytic activity and targets membranes through formation of large pores." Biochimica et biophysica acta, vol. 1854, no. 8, pp. 882–9, aug 2015. [Online]. Available: http://dx.doi.org/10.1016/j.bbapap.2015.03.003

[214] M. Eid, S. Rippa, S. Castano, B. Desbat, J. Chopineau, C. Rossi, and L. Béven, "Exploring the membrane mechanism of the bioactive peptaibol ampullosporin a using lipid monolayers and supported biomimetic membranes." Journal of biophysics (Hindawi Publishing Corporation : Online), vol. 2010, no. Article ID 179641, p. 179641, jan 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21403824

[215] M. Bortolus, A. Dalzini, F. Formaggio, C. Toniolo, M. Gobbo, and A. L. Maniero, "An EPR study of ampullosporin A, a medium-length peptaibiotic, in bicelles and vesicles." Physical chemistry chemical physics : PCCP, vol. 18, no. 2, pp. 749–60, jan 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26627901

[216] S. Ayers, B. M. Ehrmann, A. F. Adcock, D. J. Kroll, E. J. Carcache de Blanco, Q. Shen, S. M. Swanson, J. O. Falkinham, M. C. Wani, S. M. Mitchell, C. J. Pearce, and N. H. Oberlies, "Peptaibols from two unidentified fungi of the order Hypocreales with cytotoxic, antibiotic, and anthelmintic activities," Journal of Peptide Science, vol. 18, no. 8, pp. 500–510, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22744757

[217] S. B. Krasnoff, R. F. Reátegui, M. M. Wagenaar, J. B. Gloer, and D. M. Gibson, "Cicadapeptins I and II: new Aib-containing peptides from the entomopathogenic fungus Cordyceps heteropoda." Journal of natural products, vol. 68, no. 1, pp. 50–5, jan 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15679316

[218] K. Dornberger, W. Ihn, M. Ritzau, U. Gräfe, B. Schlegel, W. F. Fleck, and J. W. Metzger, "Chrysospermins, new peptaibol antibiotics from Apiocrea chrysosperma Ap101." The Journal of antibiotics, vol. 48, no. 9, pp. 977–89, sep 1995. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7592066

[219] S.-U. Oh, B.-S. Yun, S.-J. Lee, J.-H. Kim, and I.-D. Yoo, "Atroviridins A-C and neoatroviridins A-D, novel peptaibol antibiotics produced by Trichoderma atroviride F80317. I. Taxonomy, fermentation, isolation and biological activities." The Journal of antibiotics, vol. 55, no. 6, pp. 557–64, jun 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12195961

[220] S. J. Lee, B. S. Yun, D. H. Cho, and I. D. Yoo, "Tylopeptins A and B, new antibiotic peptides from Tylopilus neofelleus." The Journal of antibiotics, vol. 52, no. 11, pp. 998–1006, nov 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10656572

[221] M. I. Mitova, A. C. Murphy, G. Lang, J. W. Blunt, A. L. Cole, G. Ellis, and M. H. Munro, "Evolving trends in the dereplication of natural product extracts. 2. The isolation of chrysaibol, an antibiotic peptaibol from a New Zealand sample of the mycoparasitic fungus Sepedonium chrysospermum." Journal of natural products, vol. 71, no. 9, pp. 1600–3, sep 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18702471

[222] A. Otto, A. Laub, L. Wendt, A. Porzel, J. Schmidt, G. Palfner, J. Becerra, D. Krüger, M. Stadler, L. Wessjohann, B. Westermann, and N. Arnold, "Chilenopeptins A and B, Peptaibols from the Chilean Sepedonium aff. chalcipori KSH 883," Journal of Natural Products, vol. 79, no. 4, pp. 929–938, 2016. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acs.jnatprod.5b01018

[223] A. Otto, A. Laub, M. Haid, A. Porzel, J. Schmidt, L. Wessjohann, and N. Arnold, "Tulasporins A–D, 19-Residue Peptaibols from the Mycoparasitic Fungus Sepedonium tulasneanum," Natural Product Communications, vol. 11, no. 12, pp. 1821–1824, 2016.

[224] M. Stadler, S. Seip, H. Müller, T. Henkel, A. Lagojda, and G. Kleymann, "New antiviral peptaibols from the mycoparasitic fungus Sepedonium microspermum." in 13. Irseer Naturstofftage der DECHEMA, Irsee, 2001.

[225] Y. H. Kim, W. H. Yeo, Y. S. Kim, S. Y. Chae, and K. S. Kim, "Antiviral activity of antibiotic peptaibols, chrysospemins B and D, produced by Apiocrea sp. 14T against TMV infection," Journal of Microbiology and Biotechnology, vol. 10, no. 4, pp. 522–528, 2000.

[226] I. Berek, A. Becker, H. Schröder, A. Härtl, V. Höllt, and G. Grecksch, "Ampullosporin A, a peptaibol from Sepedonium ampullosporum HKI-0053 with neuroleptic-like activity," Behavioural Brain Research, vol. 203, no. 2, pp. 232–239, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166432809003118

[227] M. Ritzau, S. Heinze, K. Dornberger, A. Berg, W. Fleck, B. Schlegel, A. Härtl, and U. Gräfe, "Ampullosporin, a new peptaibol-type antibiotic from Sepedonium ampullosporum HKI-0053 with neuroleptic activity in mice." The Journal of antibiotics, vol. 50, no. 9, pp. 722–8, sep 1997. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9360615

[228] M. Kronen, P. Kleinwächter, B. Schlegel, a. Härtl, and U. Gräfe, "Ampullosporines B,C,D,E1,E2,E3 and E4 from Sepedonium ampullosporum HKI-0053: structures and biological activities." The Journal of antibiotics, vol. 54, no. 2, pp. 175–8, feb 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11302491

[229] M. Schiell, J. Hofmann, M. Kurz, F. R. Schmidt, L. Vértesy, M. Vogel, J. Wink, and G. Seibert, "Cephaibols, new peptaibol antibiotics with anthelmintic properties from Acremonium tubakii DSM 12774." The Journal of antibiotics, vol. 54, no. 3, pp. 220–33, mar 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11372779

[230] M. Iijima, M. Amemiya, R. Sawa, Y. Kubota, T. Kunisada, I. Momose, M. Kawada, and M. Shibasaki, "Acremopeptin, a new peptaibol from Acremonium sp. PF1450," The Journal of Antibiotics, vol. 70, pp. 1–4, feb 2017. [Online]. Available: https://www.nature.com/articles/ja201715

[231] R. Tavano, G. Malachin, M. De Zotti, C. Peggion, B. Biondi, F. Forrnaggio, and E. Papini, "The peculiar N- and (-termini of trichogin GA IV are needed for membrane interaction and human cell death induction at doses lacking antibiotic activity." Biochimica et biophysica acta, vol. 1848, no. 1 Pt A, pp. 134–44, jan 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25306964

[232] L. Du, A. L. Risinger, C. A. Mitchell, J. You, B. W. Stamps, N. Pan, J. B. King, J. C. Bopassa, S. I. V. Judge, Z. Yang, B. S. Stevenson, and R. H. Cichewicz, "Unique amalgamation of primary and secondary structural elements transform peptaibols into potent bioactive cell-penetrating peptides." Proceedings of the National Academy of Sciences of the United States of America, vol. 114, no. 43, pp. E8957–E8966, oct 2017. [Online]. Available: http://www.pnas.org/content/114/43/E8957

[233] M. Shi, H.-N. Wang, S.-T. Xie, Y. Luo, C.-Y. Sun, X.-L. Chen, and Y.-Z. Zhang, "Antimicrobial peptaibols, novel suppressors of tumor cells, targeted calcium-mediated apoptosis and autophagy in human hepatocellular carcinoma cells." Molecular cancer, vol. 9, p. 26, feb 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20122248

[234] S. Krasnoff, S. Gupta, R. Leger, J. Renwick, and D. Roberts, "Antifungal and insecticidal properties of the efrapeptins: Metabolites of the fungus Tolypocladium niveum," Journal of Invertebrate Pathology, vol. 58, no. 2, pp. 180–188, sep 1991. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/002220119190062U

[235] A. E. Papathanassiu, N. J. MacDonald, D. R. Emlet, and H. A. Vu, "Antitumor activity of efrapeptins, alone or in combination with 2-deoxyglucose, in breast cancer in vitro and in vivo." Cell stress & chaperones, vol. 16, no. 2, pp. 181–93, mar 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20927616

[236] C. P. Kubicek, M. Komoń-Zelazowska, E. Sándor, and I. S. Druzhinina, "Facts and challenges in the understanding of the biosynthesis of peptaibols by Trichoderma." Chemistry & biodiversity, vol. 4, no. 6, pp. 1068–82, jun 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17589877

[237] K. Reiber, T. Neuhof, J. H. Ozegowski, H. von Döhrend, and T. Schwecke, "A nonribosomal peptide synthetase involved in the biosynthesis of ampullosporins in Sepedonium ampullosporum." Journal of peptide science : an official publication of the European Peptide Society, vol. 9, no. 11-12, pp. 701–13, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14658790

[238] T. A. Keating and C. T. Walsh, "Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis." Current opinion in chemical biology, vol. 3, no. 5, pp. 598–606, oct 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10508662

[239] R. D. Süssmuth and A. Mainz, "Nonribosomal Peptide Synthesis–Principles and Prospects," Angewandte Chemie - International Edition, vol. 56, no. 14, pp. 3770–3821, 2017. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/anie.201609079/abstract

[240] ——, "Nicht-ribosomale Peptidsynthese - Prinzipien und Perspektiven," Angewandte Chemie, vol. 129, no. 14, pp. 3824–3878, 2017. [Online]. Available: http://doi.wiley.com/10.1002/ange.201609079

[241] T. Stachelhaus, H. D. Mootz, and M. A. Marahiel, "The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases." Chemistry & biology, vol. 6, no. 8, pp. 493–505, aug 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10421756

[242] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, and D. H. Huson, "Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs)." Nucleic acids research, vol. 33, no. 18, pp. 5799–808, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16221976

[243] T. Degenkolb, R. Karimi Aghcheh, R. Dieckmann, T. Neuhof, S. E. Baker, I. S. Druzhinina, C. P. Kubicek, H. Brückner, and H. von Döhren, "The production of multiple small peptaibol families by single 14-module Peptide synthetases in Trichoderma/Hypocrea." Chemistry & biodiversity, vol. 9, no. 3, pp. 499–535, mar 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22422521

[244] T. Neuhof, R. Dieckmann, I. S. Druzhinina, C. P. Kubicek, and H. von Döhren, "Intact-cell MALDI-TOF mass spectrometry analysis of peptaibol formation by the genus Trichoderma/Hypocrea: can molecular phylogeny of species predict peptaibol

structures?" Microbiology (Reading, England), vol. 153, no. Pt 10, pp. 3417–37, oct 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17906141

[245] P. K. Mukherjee, A. Wiest, N. Ruiz, A. Keightley, M. E. Moran-Diez, K. McCluskey, Y. F. Pouchus, and C. M. Kenerley, "Two classes of new peptaibols are synthesized by a single non-ribosomal peptide synthetase of Trichoderma virens." The Journal of biological chemistry, vol. 286, no. 6, pp. 4544–54, feb 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21123172

[246] M. A. Marahiel, "A structural model for multimodular NRPS assembly lines." Natural product reports, vol. 33, no. 2, pp. 136–40, feb 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26429504

[247] H. Link, "Observationes in ordines plantarum naturales," Magazin der Gesellschaft Naturforschender Freunde, vol. 3, pp. 3–42, 1809.

[248] C. T. Rogerson and G. J. Samuels, "Boleticolous species of Hypomyces," Mycologia, vol. 81, no. 3, pp. 413–432, 1989. [Online]. Available: https://www.jstor.org/stable/3760079

[249] H. Ammer, H. Besl, and S. Vilsmeier, "Der Flaschensporige Goldschimmel, Sepedonium ampullosporum - ein thermophiler Parasit an Pilzfruchtkörpern der Ordnung Boletales," Zeitschrift für Mykologie, vol. 63, no. 2, pp. 127–132, 1997.

[250] H. Besl, A. Hagn, A. Jobst, and U. Lange, "Der Kleinsporige Goldschimmel, Sepedonium microspermum - ein Parasit an Röhrlingen der Xerocomus-chrysenteron-Gruppe," Zeitschrift für Mykologie, vol. 64, no. 1, pp. 45–52, 1998.

[251] S. C. Damon, "Two Noteworthy Species of Sepedonium," Mycologia, vol. 44, no. 1, pp. 86–96, 1952. [Online]. Available: http://www.jstor.org/stable/4547568

[252] G. R. Arnold, "Über Apiocrea tulasneana (Plowr.) Syd." Westfälische Pilzbriefe, vol. 7, p. 80, 1969.

[253] G. J. Samuels and K. A. Seifert, "Taxonomic Implications of Variation among Hypocrealean Anamorphs," in Pleomorphic Fungi: The Diversity and Its Taxonomic Implications, J. Sugiyama, Ed. Elsevier, 1987, ch. Taxonomic, pp. 29–56. [Online]. Available: http://dx.doi.org/10.1016/B978-0-444-98966-6.50007-1

[254] T. Sahr, H. Ammer, H. Besl, M. Fischer, and A. H. B. H. Sahr T., "Infrageneric classification of the boleticolous genus Sepedonium: species delimitation and phylogenetic relationships," Mycologia, vol. 91, no. 6, pp. 935–943, nov 1999. [Online]. Available: https://www.jstor.org/stable/3761625

[255] W. Helfer, "Pilze auf Pilzfruchtkoerpern: Untersuchungen zur Oekologie, Systematik und Chemie," in Libri Botanici 1. Eching: IHW-Verlag, 1991, pp. 1–157. [Online]. Available: https://www.schweitzer-online.de/buch/Helfer/Pilze-Pilzfruchtkoerpern/9783980273220/A19950465/

[256] J. Arellano-Galindo, V.-M. Eugenio, J.-H. Elva, R.-S. Jesús, M.-R. María de Los Ángeles, J.-J. Rodolfo Norberto, X.-C. Juan, O. Sara A, and C.-C. Ariadna, "A saprophytic fungus (Sepedonium) associated with fatal pneumonia in a patient undergoing stem cell transplantation." The Journal of international medical research, vol. 25, no. 3, p. 300060517708103, jan 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28606018

[257] K. P. Ng, T. S. Soo-Hoo, S. L. Na, and L. H. Tan, "Sepedonium species: An emerging opportunistic fungal infection in a patient with AIDS," Clinical Microbiology Newsletter, vol. 25, no. 3, pp. 20–22, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0196439903800259

[258] S. Shibata, J. Shoji, A. Ohta, and M. Watanabe, "Metabolic products of fungi. XI. Some observation on the occurrence of skyrin and rugulosin in mold metabolites, with a reference to structural relationship between penicilliopsin and skyrin." Pharmaceutical bulletin, vol. 5, no. 4, pp. 380–2, aug 1957. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/13494139

[259] A. Kato, K. Ando, K. Kodama, S. Suzuki, K. Suzuki, G. Tamura, and K. Arima, "Production, isolation and purification of antitumor active monoglycerides and other antibiotics from Sepedonium ampullosporum." The Journal of Antibiotics, vol. 22, no. 2, pp. 71–76, 1969. [Online]. Available: http://joi.jlc.jst.go.jp/JST.Journalarchive/antibiotics1968/22.71?from=CrossRef

[260] P. Divekar, H. Raistrick, T. Dobson, and L. C. Vining, "Sepedonin, a tropolone metabolite of Sepedonium chrysosperumum Fries," Canadian Journal of Chemistry, vol. 43, no. October 1966, pp. 1835–1848, 1965.

[261] D. N. Quang, J. Schmidt, A. Porzel, L. Wessjohann, M. Haid, and N. Arnold, "Ampullosine, a new isoquinoline alkaloid from Sepedonium ampullosporum (Ascomycetes)." Natural product communications, vol. 5, no. 6, pp. 869–872, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20614812

[262] A. Closse and D. Hauser, "Isolierung und Konstitutionsermittlung von Chrysodin,"

Helvetica Chimica Acta, vol. 56, no. 8, pp. 2694–2698, 1973. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/hlca.19730560803/abstract

[263] A. G. Brown, T. C. Smale, T. J. King, R. Hasenkamp, and R. H. Thompson, "Crystal and molecular structure of compactin, a new antifungal metabolite from Penicillium brevicompactum," Journal of the Chemical Society, Perkin Transactions 1, vol. 0, no. 11, p. 1165, 1976. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/945291http://xlink.rsc.org/?DOI=p19760001165

[264] A. Endo, K. Hasumi, A. Yamada, R. Shimoda, and H. Takeshima, "The synthesis of compactin (ML-236B) and monacolin K in fungi," The Journal of antibiotics, vol. XXXIX, no. 11, pp. 1609–1610, 1986. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/3793631

[265] M. I. Mitova, B. G. Stuart, G. H. Cao, J. W. Blunt, A. L. J. Cole, and M. H. G. Munro, "Chrysosporide, a cyclic pentapeptide from a New Zealand sample of the fungus Sepedonium chrysospermum." Journal of natural products, vol. 69, no. 10, pp. 1481–4, oct 2006. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/np060137o

[266] T. Suzuki, C. Okada, K. Arai, A. Awaji, T. Shimizu, K. Tanemura, and T. Horaguchi, "Synthesis of 7-acetyloxy-3, 7-dimethyl-7, 8-dihydro-6H-isochromene-6, 8-dione and its analogues," Journal of heterocyclic chemistry, vol. 38, pp. 1409–1418, 2001. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/jhet.5570380625/full

[267] H.-H. Nguyen, D. Imhof, M. Kronen, B. Schlegel, A. Härtl, U. Gräfe, L. Gera, and S. Reissmann, "Synthesis and biological evaluation of analogues of the peptaibol ampullosporin A." Journal of medicinal chemistry, vol. 45, no. 13, pp. 2781–7, jun 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12061880

[268] H.-H. Nguyen, D. Imhof, M. Kronen, U. Gräfe, and S. Reissmann, "Circular dichroism studies of ampullosporin-A analogues." Journal of peptide science : an official publication of the European Peptide Society, vol. 9, no. 11-12, pp. 714–28, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14658791

[269] P. A. Grigoriev, M. Kronen, B. Schlegel, A. Härtl, and U. Gräfe, "Differences in ion-channel formation by ampullosporins B, C, D and semisynthetic desacetyltryptophanyl ampullosporin A." Bioelectrochemistry (Amsterdam,

Netherlands), vol. 57, no. 2, pp. 119–21, sep 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12160607

[270] P. A. Grigoriev, B. Schlegel, M. Kronen, A. Berg, A. Härtl, and U. Gräfe, "Differences in membrane pore formation by peptaibols." Journal of peptide science : an official publication of the European Peptide Society, vol. 9, no. 11-12, pp. 763–8, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14658795

[271] H. Hülsmann, S. Heinze, M. Ritzau, B. Schlegel, and U. Gräfe, "Isolation and structure of peptaibolin, a new peptaibol from Sepedonium strains." The Journal of antibiotics, vol. 51, no. 11, pp. 1055–8, nov 1998. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9918401

[272] T. Neuhof, A. Berg, H. Besl, T. Schwecke, R. Dieckmann, and H. von Döhren, "Peptaibol production by sepedonium strains parasitizing boletales." Chemistry & biodiversity, vol. 4, no. 6, pp. 1103–15, jun 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17589879

[273] T. Degenkolb, L. Götze, H. von Döhren, A. Vilcinskas, and H. Brückner, "Sequences of stilboflavin C: towards the peptaibiome of the filamentous fungus Stilbella (= Trichoderma) flavipes." Journal of peptide science : an official publication of the European Peptide Society, vol. 22, no. 8, pp. 517–24, aug 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27443977

[274] N. Ruiz, G. Wielgosz-Collin, L. Poirier, O. Grovel, K. E. Petit, M. Mohamed-Benkada, T. R. du Pont, J. Bissett, P. Vérité, G. Barnathan, and Y. F. Pouchus, "New Trichobrachins, 11-residue peptaibols from a marine strain of Trichoderma longibrachiatum." Peptides, vol. 28, no. 7, pp. 1351–8, jul 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17629355

[275] B. Biondi, C. Peggion, M. De Zotti, C. Pignaffo, A. Dalzini, M. Bortolus, S. Oancea, G. Hilma, A. Bortolotti, L. Stella, J. Z. Pedersen, V. N. Syryamina, Y. D. Tsvetkov, S. A. Dzuba, C. Toniolo, and F. Formaggio, "Conformational properties, membrane interaction, and antibacterial activity of the peptaibiotic chalciporin A: Multitechnique spectroscopic and biophysical investigations on the natural compound and labeled analogs," Peptide Science, vol. 108, no. July, p. e23083, feb 2018. [Online]. Available: http://doi.wiley.com/10.1002/bip.23083

[276] S. J. Lee, W. H. Yeo, B. S. Yun, and I. D. Yoo, "Isolation and sequence analysis of new peptaibol, boletusin, from Boletus spp." Journal of peptide science : an official

publication of the European Peptide Society, vol. 5, no. 8, pp. 374–8, aug 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10507687

[277] K. Michels, "Entwicklung einer LC-MS basierten Methode zur Identifizierung von aktivitätsrelevanten Metaboliten in komplexen Mischungen," Dissertation, Martin-Luther-Universität Halle-Wittenberg, 2011. [Online]. Available: https://d-nb.info/1025351746/34

[278] K. Michels, R. Heinke, P. Schöne, O. P. Kuipers, N. Arnold, and L. A. Wessjohann, "A fluorescence-based bioassay for antibacterials and its application in screening natural product extracts." The Journal of antibiotics, vol. 68, no. 12, pp. 734–40, dec 2015. [Online]. Available: http://www.nature.com/articles/ja201571

[279] J.-w. Veening, W. K. Smits, L. W. Hamoen, J. D. Jongbloed, and O. P. Kuipers, "Visualization of differential gene expression by improved cyan fluorescent protein and yellow fluorescent protein production in Bacillus subtilis." Applied and environmental microbiology, vol. 70, no. 11, pp. 6809–15, nov 2004. [Online]. Available: http://aem.asm.org/content/70/11/6809.long

[280] J. Fogh and G. Trempe, "New Human Tumor Cell Lines," in Human Tumor Cells in Vitro, J. Fogh, Ed. Boston, MA: Springer US, 1975, ch. 5, pp. 115–159. [Online]. Available: http://www.springer.com/us/book/9781475716498

[281] K. D. Paull, R. H. Shoemaker, M. R. Boyd, J. L. Parsons, P. A. Risbood, W. A. Barbera, M. N. Sharma, D. C. Baker, E. Hand, D. A. Scudiero, A. Monks, M. C. Alley, and M. Grote, "The synthesis of XTT: A new tetrazolium reagent that is bioreducible to a water-soluble formazan," Journal of Heterocyclic Chemistry, vol. 25, no. 3, pp. 911–914, may 1988. [Online]. Available: http://doi.wiley.com/10.1002/jhet.5570250340

[282] M. V. Berridge, A. S. Tan, K. D. McCoy, and R. Wang, "The biochemical and cellular basis of cell proliferation assays that use tetrazolium salts," Biochemica, vol. 4, no. 1, pp. 14–19, 1996.

[283] F. Commo and B. M. Bot, "nplr: N-Parameter Logistic Regression," 2016. [Online]. Available: http://cran.r-project.org/package=nplr

[284] F. J. Richards, "A Flexible Growth Function for Empirical Use," Journal of Experimental Botany, vol. 10, no. 2, pp. 290–301, 1959. [Online]. Available: https://academic.oup.com/jxb/article/10/2/290/528209

[285] J. Giraldo, N. M. Vivas, E. Vila, and A. Badia, "Assessing the (a)symmetry of concentration-effect curves: empirical versus mechanistic models." Pharmacology & therapeutics, vol. 95, no. 1, pp. 21–45, jul 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12163126

[286] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem Substance and Compound databases." Nucleic acids research, vol. 44, no. D1, pp. D1202–13, jan 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26400175

[287] J. Buckingham, "Dictionary of natural products on CD-ROM," London, 2005.

[288] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka, "MassBank: a public repository for sharing mass spectral data for life sciences." Journal of mass spectrometry : JMS, vol. 45, no. 7, pp. 703–14, jul 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20623627

[289] R Core Team, "R: A Language and Environment for Statistical Computing," Computer Program, Vienna, Austria, 2013. [Online]. Available: http://www.r-project.org/

[290] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching." Bioinformatics (Oxford, England), vol. 22, no. 17, pp. 2059–65, sep 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16820428

[291] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification." Analytical chemistry, vol. 78, no. 3, pp. 779–87, feb 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16448051

[292] R. Tautenhahn, C. Böttcher, and S. Neumann, "Highly sensitive feature detection for high resolution LC/MS." BMC bioinformatics, vol. 9, p. 504, nov 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19040729

[293] R. Heinke, "Mass Spectrometry, Biological Screening and Informatics of Prenylated Natural Products," Ph.D. dissertation, Martin-Luther-Universität Halle-Wittenberg, 2015. [Online]. Available: http://digital.bibliothek.uni-halle.de/id/2332987

[294] T. Wei and V. Simko, "corrplot: Visualization of a Correlation Matrix," 2016. [Online]. Available: http://cran.r-project.org/package=corrplot

[295] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, vol. 58, no. 301, pp. 236–244, mar 1963. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845

[296] F. Schlünzen, R. Zarivach, J. Harms, A. Bashan, A. Tocilj, R. Albrecht, A. Yonath, and F. Franceschi, "Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria." Nature, vol. 413, no. 6858, pp. 814–21, oct 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11677599

[297] J. Xia, D. I. Broadhurst, M. Wilson, and D. S. Wishart, "Translational biomarker discovery in clinical metabolomics: an introductory tutorial." Metabolomics : Official journal of the Metabolomic Society, vol. 9, no. 2, pp. 280–299, apr 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23543913

[298] Haibo He and E. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, sep 2009. [Online]. Available: http://ieeexplore.ieee.org/document/5128907/

[299] D. Ruiz-Perez and G. Narasimhan, "So you think you can PLS-DA?" bioRxiv preprint, vol. 99, no. 1, p. 17, 2017. [Online]. Available: https://www.biorxiv.org/content/early/2017/10/21/207225

[300] C. Christin, H. C. J. Hoefsloot, A. K. Smilde, B. Hoekman, F. Suits, R. Bischoff, and P. Horvatovich, "A critical assessment of feature selection methods for biomarker discovery in clinical proteomics." Molecular & cellular proteomics : MCP, vol. 12, no. 1, pp. 263–76, jan 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23115301

[301] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R." Bioinformatics (Oxford, England), vol. 21, no. 20, pp. 3940–3941, oct 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16096348

[302] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "pROC: an open-source package for R and S+ to analyze and compare

ROC curves." BMC bioinformatics, vol. 12, no. 1, p. 77, mar 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/77

[303] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." Biometrics, vol. 44, no. 3, pp. 837–45, sep 1988. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3203132

[304] H. Wickham, ggplot2, R. Gentleman, K. Hornik, and G. Parmigiani, Eds. New York, NY: Springer New York, 2009. [Online]. Available: http://link.springer.com/10.1007/978-0-387-98141-3

[305] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, "pcaMethods–a bioconductor package providing PCA methods for incomplete data." Bioinformatics (Oxford, England), vol. 23, no. 9, pp. 1164–7, may 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17344241

[306] R. Wehrens, Chemometrics with R, R. Gentleman, K. Hornik, and G. Parmigiani, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. [Online]. Available: http://link.springer.com/10.1007/978-3-642-17841-2

[307] P. Filzmoser and K. Varmuza, "chemometrics: Multivariate Statistical Analysis in Chemometrics," 2015. [Online]. Available: http://cran.r-project.org/package=chemometrics

[308] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," Chemometrics and Intelligent Laboratory Systems, vol. 18, no. 3, pp. 251–263, mar 1993. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/016974399385002X

[309] F. Lindgren, P. Geladi, and S. Wold, "The kernel algorithm for PLS," Journal of Chemometrics, vol. 7, no. 1, pp. 45–59, jan 1993. [Online]. Available: http://doi.wiley.com/10.1002/cem.1180070104

[310] W. Kessler, Multivariate Datenanalyse: für die Pharma-, Bio- und Prozessanalytik. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, oct 2006. [Online]. Available: http://doi.wiley.com/10.1002/9783527610037

[311] E. A. Thévenot, A. Roux, Y. Xu, E. Ezan, and C. Junot, "Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical

Analyses," Journal of Proteome Research, vol. 14, no. 8, pp. 3322–3335, aug 2015. [Online]. Available: http://pubs.acs.org/doi/10.1021/acs.jproteome.5b00354

[312] B.-H. Mevik and R. Wehrens, "The pls Package: Principal Component and Partial Least Squares Regression in R," Journal of Statistical Software, vol. 18, no. 2, 2007. [Online]. Available: http://www.jstatsoft.org/v18/i02/

[313] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 12, no. 1, pp. 55–67, feb 1970.

[314] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2013, vol. 103. [Online]. Available: http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdfhttp://link.springer.com/10.1007/978-1-4614-7138-7

[315] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," Journal of Statistical Software, vol. 33, no. 1, pp. 1–22, may 2010.

[316] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, apr 2005. [Online]. Available: http://www.jstor.org/stable/3647580

[317] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical Learning with Sparsity: The Lasso and Generalizations," Crc, p. 362, 2015. [Online]. Available: https://www.crcpress.com/Statistical-Learning-with-Sparsity-The-Lasso-and-Generalizations/Hastie-Tibshirani-Wainwright/p/book/9781498712163

[318] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen, Classification and Regression Trees. Chapman and Hall/CRC, 1984.

[319] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY: Springer New York, 2013. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-6849-3

[320] R. Genuer, J.-m. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," Pattern Recognition Letters, vol. 31, no. 14, pp. 2225–2236, oct 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167865510000954

[321] W. Li, J. Wang, and Z.-Y. Yan, "Development of a sensitive and rapid method for rifampicin impurity analysis using supercritical fluid chromatography." Journal of pharmaceutical and biomedical analysis, vol. 114, pp. 341–7, oct 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26103526

[322] H. M. Bolt and H. Remmer, "Implication of rifampicin-quinone in the irreversible binding of rifampicin to macromolecules." Xenobiotica; the fate of foreign compounds in biological systems, vol. 6, no. 1, pp. 21–32, jan 1976. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/5822

[323] B. Prasad and S. Singh, "In vitro and in vivo investigation of metabolic fate of rifampicin using an optimized sample preparation approach and modern tools of liquid chromatography-mass spectrometry." Journal of pharmaceutical and biomedical analysis, vol. 50, no. 3, pp. 475–90, oct 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19535209

[324] D. Volmer and J. Hui, "Study of erythromycin A decomposition products in aqueous solution by solid-phase microextraction/liquid chromatography/tandem mass spectrometry." Rapid communications in mass spectrometry : RCM, vol. 12, no. 3, pp. 123–9, jan 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9474800

[325] A. Deubel, A. Fandiño, F. Sörgel, and U. Holzgrabe, "Determination of erythromycin and related substances in commercial samples using liquid chromatography/ion trap mass spectrometry." Journal of chromatography. A, vol. 1136, no. 1, pp. 39–47, dec 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17014855

[326] G. C. Kearney, P. J. Gates, P. F. Leadlay, J. Staunton, and R. Jones, "Structural elucidation studies of erythromycins by electrospray tandem mass spectrometry II." Rapid communications in mass spectrometry : RCM, vol. 13, no. 16, pp. 1650–6, jan 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10440983

[327] J. P. Shaffer, "Multiple Hypothesis Testing," Annual Review of Psychology, vol. 46, no. 1, pp. 561–584, jan 1995. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev.ps.46.020195.003021

[328] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society. Series B (Methodological), vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[329] T. A. Knijnenburg, L. F. A. Wessels, M. J. T. Reinders, and I. Shmulevich, "Fewer permutations, more accurate P-values." Bioinformatics (Oxford, England), vol. 25, no. 12, pp. i161–8, jun 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19477983

[330] N. Altman and M. Krzywinski, "Points of significance: Sources of variation." Nature methods, vol. 12, no. 1, pp. 5–6, jan 2015. [Online]. Available: http://www.nature.com/articles/nmeth.4210

[331] R. Nuzzo, "Scientific method: Statistical errors," Nature, vol. 506, no. 7487, pp. 150–152, feb 2014. [Online]. Available: http://www.nature.com/news/scientific-method-statistical-errors-1.14700

[332] A. C. Davison and D. V. Hinkley, Bootstrap methods and their application, 1st ed. Cambridge: Cambridge University Press, 1997. [Online]. Available: http://ebooks.cambridge.org/ref/id/CBO9780511802843

[333] K. Kjeldahl and R. Bro, "Some common misunderstandings in chemometrics," Journal of Chemometrics, vol. 24, no. 7-8, pp. 558–564, jul 2010. [Online]. Available: http://doi.wiley.com/10.1002/cem.1346

[334] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in Partial Least Squares Regression," Chemometrics and Intelligent Laboratory Systems, vol. 118, pp. 62–69, aug 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169743912001542

[335] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest." BMC bioinformatics, vol. 7, p. 3, jan 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16398926

[336] X. Chen and H. Ishwaran, "Random forests for genomic data analysis." Genomics, vol. 99, no. 6, pp. 323–9, jun 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22546560

[337] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." Psychological methods, vol. 14, no. 4, pp. 323–48, dec 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19968396

[338] A. Dasgupta, Y. V. Sun, I. R. König, J. E. Bailey-Wilson, and J. D. Malley, "Brief review of regression-based and machine learning methods in genetic epidemiology:

the Genetic Analysis Workshop 17 experience." Genetic epidemiology, vol. 35 Suppl 1, no. Suppl 1, pp. S5–11, 2011.

[339] H. C. V. Houwelingen and W. Sauerbrei, "Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited," Open Journal of Statistics, vol. 03, no. 02, pp. 79–102, 2013. [Online]. Available: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=30157{&}{#}abstract

[340] R. Heinke, P. Schöne, N. Arnold, L. Wessjohann, J. Schmidt, and J. Schmidt, "Metabolite profiling and fingerprinting of Suillus species (Basidiomycetes) by electrospray mass spectrometry." European journal of mass spectrometry (Chichester, England), vol. 20, no. 1, pp. 85–97, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24881458

[341] H. A. Gad, S. H. El-Ahmady, M. I. Abou-Shoer, and M. M. Al-Azizi, "Application of chemometrics in authentication of herbal medicines: a review." Phytochemical analysis : PCA, vol. 24, no. 1, pp. 1–24, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22678654

[342] A. M. van Nederkassel, M. Daszykowski, D. L. Massart, and Y. Vander Heyden, "Prediction of total green tea antioxidant capacity from chromatograms by multivariate modeling." Journal of chromatography. A, vol. 1096, no. 1-2, pp. 177–86, nov 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16301079

[343] S. Thiangthum, B. Dejaegher, M. Goodarzi, C. Tistaert, A. Y. Gordien, N. Nguyen Hoai, M. Chau Van, J. Quetin-Leclercq, L. Suntornsuk, and Y. Vander Heyden, "Potentially antioxidant compounds indicated from Mallotus and Phyllanthus species fingerprints." Journal of chromatography. B, Analytical technologies in the biomedical and life sciences, vol. 910, pp. 114–21, dec 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22795556

[344] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1-3, pp. 37–52, aug 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0169743987800849

[345] O. M. Kvalheim, "Scaling of analytical data," Analytica Chimica Acta, vol. 177, no. C, pp. 71–79, 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003267000829396

[346] L. Eriksson, J. Trygg, and S. Wold, "A chemometrics toolbox based on projections

and latent variables," Journal of Chemometrics, vol. 28, no. 5, pp. 332–346, may 2014. [Online]. Available: http://doi.wiley.com/10.1002/cem.2581

[347] R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data." BMC genomics, vol. 7, p. 142, jun 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16762068

[348] C. D. Brown and R. L. Green, "Critical factors limiting the interpretation of regression vectors in multivariate calibration," TrAC Trends in Analytical Chemistry, vol. 28, no. 4, pp. 506–514, apr 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165993609000363

[349] M. C.-H. Ng, T.-Y. Lau, K. Fan, Q.-S. Xu, J. Poon, S. K. Poon, M. K. Lam, F.-T. Chau, and D. M.-Y. Sze, "Prediction of Radix Astragali Immunomodulatory Effect of CD80 Expression from Chromatograms by Quantitative Pattern-Activity Relationship." BioMed research international, vol. 2017, p. 3923865, 2017. [Online]. Available: https://www.hindawi.com/journals/bmri/2017/3923865/

[350] F.-T. Chau, Q.-S. Xu, D. M.-Y. Sze, H.-Y. Chan, T.-Y. Lau, D.-L. Yuan, M. C.-H. Ng, K. Fan, D. K.-W. Mok, and Y.-Z. Liang, "A New Methodology for Uncovering the Bioactive Fractions in Herbal Medicine Using the Approach of Quantitative Pattern-Activity Relationship," in Data Analytics for Traditional Chinese Medicine Research, J. Poon and S. K. Poon, Eds. Cham: Springer International Publishing, 2014, no. December, pp. 155–172.

[351] M. Farrés, S. Platikanov, S. Tsakovski, and R. Tauler, "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation," Journal of Chemometrics, vol. 29, no. 10, pp. 528–536, oct 2015. [Online]. Available: http://doi.wiley.com/10.1002/cem.2736

[352] T. N. Tran, N. L. Afanador, L. M. Buydens, and L. Blanchet, "Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC)," Chemometrics and Intelligent Laboratory Systems, vol. 138, pp. 153–160, nov 2014. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0169743914001786

[353] B. Campos, N. Garcia-Reyero, C. Rivetti, L. Escalon, T. Habib, R. Tauler, S. Tsakovski, B. Piña, and C. Barata, "Identification of metabolic pathways in Daphnia magna explaining hormetic effects of selective serotonin reuptake inhibitors

and 4-nonylphenol using transcriptomic and phenotypic responses." Environmental science & technology, vol. 47, no. 16, pp. 9434–43, aug 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23855649

[354] B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk, and M. Sznajder, "Detection of discoloration in diesel fuel based on gas chromatographic fingerprints." Analytical and bioanalytical chemistry, vol. 407, no. 4, pp. 1159–70, feb 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25407430

[355] A. Acharjee, R. Finkers, R. G. Visser, and C. Maliepaard, "Comparison of Regularized Regression Methods for ~Omics Data," Metabolomics, vol. 3, no. 3, p. 126, 2013.

[356] D. Lin, G. Cohen Freue, Z. Hollander, G. B. John Mancini, M. Sasaki, A. Mui, J. Wilson-McManus, A. Ignaszewski, C. Imai, A. Meredith, R. Balshaw, R. T. Ng, P. A. Keown, W. Robert McMaster, R. Carere, J. G. Webb, B. M. McManus, Biomarkers in Transplantation Team, C. o. E. f. C. Networks of Centres of Excellence, and R.-P. of Organ Failure Centre of Excellence, "Plasma protein biosignatures for detection of cardiac allograft vasculopathy." The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation, vol. 32, no. 7, pp. 723–33, jul 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23796154

[357] M. Lankinen, U. Schwab, P. V. Gopalacharyulu, T. Seppänen-Laakso, L. Yetukuri, M. Sysi-Aho, P. Kallio, T. Suortti, D. E. Laaksonen, H. Gylling, K. Poutanen, M. Kolehmainen, and M. Oresic, "Dietary carbohydrate modification alters serum metabolic profiles in individuals with the metabolic syndrome." Nutrition, metabolism, and cardiovascular diseases : NMCD, vol. 20, no. 4, pp. 249–57, may 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19553094

[358] F. Mostajabi, S. Datta, and S. Datta, "Predicting patient survival from proteomic profile using mass spectrometry data: An empirical study," Communications in Statistics: Simulation and Computation, vol. 42, no. 3, pp. 485–498, 2013. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03610918.2011.636165

[359] F. Lu and E. Petkova, "A comparative study of variable selection methods in the context of developing psychiatric screening instruments." Statistics in medicine, vol. 33, no. 3, pp. 401–21, feb 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23934941

[360] A. C. Alves, M. Rantalainen, E. Holmes, J. K. Nicholson, and T. M. D. Ebbels, "Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets." Analytical chemistry, vol. 81, no. 6, pp. 2075–84, mar 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19220030

[361] O. Cloarec, M.-E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, and J. Nicholson, "Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets." Analytical chemistry, vol. 77, no. 5, pp. 1282–9, mar 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15732908

[362] C. Andersen and R. Bro, "Variable selection in regression-a tutorial," Journal of Chemometrics, vol. 24, no. 11-12, pp. 728–737, nov 2010. [Online]. Available: http://doi.wiley.com/10.1002/cem.1360

[363] C. M. Rubingh, S. Bijlsma, E. P. P. a. Derks, I. Bobeldijk, E. R. Verheij, S. Kochhar, and A. K. Smilde, "Assessing the performance of statistical validation tools for megavariate metabolomics data." Metabolomics : Official journal of the Metabolomic Society, vol. 2, no. 2, pp. 53–61, jul 2006. [Online]. Available: http://link.springer.com/10.1007/s11306-006-0022-6

[364] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," NeuroImage, jun 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28655633

[365] H. S. Tapp and E. K. Kemsley, "Notes on the practical utility of OPLS," TrAC Trends in Analytical Chemistry, vol. 28, no. 11, pp. 1322–1327, dec 2009. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165993609001873

[366] T. Chen, Y. Cao, Y. Zhang, J. Liu, Y. Bao, C. Wang, W. Jia, and A. Zhao, "Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection." Evidence-based complementary and alternative medicine : eCAM, vol. 2013, p. 298183, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23573122

[367] Z. Lin, C. M. Vicente Gonçalves, L. Dai, H.-m. Lu, J.-h. Huang, H. Ji, D.-s. Wang, L.-z. Yi, and Y.-z. Liang, "Exploring metabolic syndrome serum profiling based on gas chromatography mass spectrometry and random forest models." Analytica chimica acta, vol. 827, pp. 22–7, may 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24832990

[368] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution." BMC bioinformatics, vol. 8, p. 25, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17254353

[369] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, no. 4, pp. 221–232, 2016. [Online]. Available: http://link.springer.com/10.1007/s13748-016-0094-0

[370] H. S. Tapp, M. Radonjic, E. Kate Kemsley, and U. Thissen, "Evaluation of multiple variate selection methods from a biological perspective: a nutrigenomics case study." Genes & nutrition, vol. 7, no. 3, pp. 387–97, jul 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22382778

[371] S. Hielscher-Michael, "Glutaminylzyklase (QC) inhibierende Verbindungen aus Mikroalgen – neue Leitstrukturen für den Einsatz in der Therapie der Alzheimer Erkrankung," Ph.D. dissertation, Martin-Luther-Universität Halle-Wittenberg, 2017. [Online]. Available: http://digital.bibliothek.uni-halle.de/id/2515415

[372] S. Hielscher-Michael, C. Griehl, M. Buchholz, H.-U. Demuth, N. Arnold, and L. A. Wessjohann, "Natural Products from Microalgae with Potential against Alzheimer's Disease: Sulfolipids Are Potent Glutaminyl Cyclase Inhibitors." Marine drugs, vol. 14, no. 11, nov 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27827845

[373] T. Inui, Y. Wang, S. M. Pro, S. G. Franzblau, and G. F. Pauli, "Unbiased evaluation of bioactive secondary metabolites in complex matrices." Fitoterapia, vol. 83, no. 7, pp. 1218–25, oct 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22766306

[374] M. B. Abdullah, "On a Robust Correlation Coefficient," Journal of the Royal Statistical Society. Series D (The Statistician), vol. 39, no. 4, pp. 455–460, 1990. [Online]. Available: https://www.jstor.org/stable/2349088

[375] J. H. Friedman, "Multivariate Adaptive Regression Splines," The Annals of Statistics, vol. 19, no. 1, pp. 1–67, 1991. [Online]. Available: http://www.jstor.org/stable/2241837

[376] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proceedings of the National

Academy of Sciences of the United States of America, vol. 99, no. 10, pp. 6567–72, may 2002. [Online]. Available: http://www.pnas.org/content/99/10/6567

[377] Y.-H. Yun, B.-C. Deng, D.-S. Cao, W.-T. Wang, and Y.-Z. Liang, "Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery," Analytica Chimica Acta, vol. 911, no. June 2015, pp. 27–34, mar 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267016300216?via{%}3Dihub

[378] A. Jaworski and H. Brückner, "Detection of new sequences of peptaibol antibiotics trichotoxins A-40 by on-line liquid chromatography-electrospray ionization mass spectrometry." Journal of chromatography. A, vol. 862, no. 2, pp. 179–89, nov 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10596975

[379] B. Paizs and S. Suhai, "Fragmentation pathways of protonated peptides," Mass Spectrometry Reviews, vol. 24, no. 4, pp. 508–548, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15389847

[380] N. Stoppacher, N. K. N. Neumann, L. Burgstaller, S. Zeilinger, T. Degenkolb, H. Brückner, and R. Schuhmacher, "The comprehensive peptaibiotics database." Chemistry & biodiversity, vol. 10, no. 5, pp. 734–43, may 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23681723

[381] H. Edelhoch, "Spectroscopic determination of tryptophan and tyrosine in proteins." Biochemistry, vol. 6, no. 7, pp. 1948–1954, 1967. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/bi00859a010

[382] A. Carroux, A.-I. Van Bohemen, C. Roullier, T. Robiou du Pont, M. Vansteelandt, A. Bondon, A. Zalouk-Vergnoux, Y. F. Pouchus, N. Ruiz, A.-i. V. Bohemen, C. Roullier, T. Robiou, M. Vansteelandt, A. Bondon, A. Zalouk-Vergnoux, and N. Ruiz, "Unprecedented 17-residue peptaibiotics produced by marine-derived Trichoderma atroviride." Chemistry & biodiversity, vol. 10, no. 5, pp. 772–86, may 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23681725

[383] R. Tavano, G. Malachin, M. De Zotti, C. Peggion, B. Biondi, F. Formaggio, and E. Papini, "Comparison of bactericidal and cytotoxic activities of trichogin analogs." Data in brief, vol. 6, pp. 359–67, mar 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26862583

[384] P. Banerjee, J. Erehman, B.-O. Gohlke, T. Wilhelm, R. Preissner, and M. Dunkel, "Super Natural II– a database of natural products." Nucleic acids

research, vol. 43, no. Database issue, pp. D935–9, jan 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25300487

[385] A. Jaworski and H. Brückner, "New sequences and new fungal producers of peptaibol antibiotics antiamoebins." Journal of peptide science : an official publication of the European Peptide Society, vol. 6, no. 4, pp. 149–67, apr 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10809388

[386] A. Berg, M. Ritzau, W. Ihn, B. Schlegel, W. F. Fleck, S. Heinze, and U. Gräfe, "Isolation and structure of bergofungin, a new antifungal peptaibol from Emericellopsis donezkii HKI 0059." The Journal of antibiotics, vol. 49, no. 8, pp. 817–20, aug 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8823517

[387] A. D. Argoudelis and L. E. Johnson, "Emerimicins II, III and IV, antibiotics produced by Emericellopsis microspora in media supplemented with trans-4-n-propyl-L-proline." The Journal of antibiotics, vol. 27, no. 4, pp. 274–82, apr 1974. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/4851844

[388] A. D. Argoudelis, A. Dietz, and L. E. Johnson, "Zervamicins I and II, polypeptide antibiotics produced by emericellopsis salmosynnemata." The Journal of antibiotics, vol. 27, no. 5, pp. 321–8, may 1974. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/4855438

[389] T. V. Ovchinnikova, N. G. Levitskaya, O. G. Voskresenskaya, Z. A. Yakimenko, A. A. Tagaev, A. Y. Ovchinnikova, A. N. Murashev, and A. A. Kamenskii, "Neuroleptic properties of the ion-channel-forming peptaibol zervamicin: Locomotor activity and behavioral effects," Chemistry and Biodiversity, vol. 4, no. 6, pp. 1374–1387, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17589870

[390] Y. Hayakawa, H. Adachi, J. W. Kim, K. Shin-ya, and H. Seto, "Adenopeptin, a new apoptosis inducer in transformed cells from Chrysosporium sp." Tetrahedron, vol. 54, no. 52, pp. 15 871–15 878, dec 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S004040209800996X

[391] S. B. Krasnoff and S. Gupta, "Efrapeptin production by Tolypocladium fungi (Deuteromycotina: Hyphomycetes): Intra- and interspecific variation." Journal of chemical ecology, vol. 18, no. 10, pp. 1727–41, oct 1992. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24254715

[392] A. Fredenhagen, L.-P. Molleyres, B. Böhlendorf, and G. Laue, "Structure determination of neoefrapeptins A to N: peptides with insecticidal activity produced

by the fungus Geotrichum candidum." The Journal of antibiotics, vol. 59, no. 5, pp. 267–80, may 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16883776

[393] T. Degenkolb, C. R. Röhrich, A. Vilcinskas, H. von Döhren, and H. Brückner, "A new family of N-terminally truncated peptaibols from the biocontrol fungus Trichoderma harzianum," in Proceedings - 34th European Peptide Symposium, 2016, pp. PPVI–093.

[394] C. Theis, T. Degenkolb, and H. Brückner, "Studies on the selective trifluoroacetolytic scission of native peptaibols and model peptides using HPLC and ESI-CID-MS." Chemistry & biodiversity, vol. 5, no. 11, pp. 2337–55, nov 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19035563

Teil IV.

Anhang

A. Antibiotika

Tabelle A.1.: Liste der Antibiotikapeaks im Positiv-Ionen-Modus

m/z	Antibiotikum	Annotation
mz	Antibiotikum	Annotation
789.37314	Rif	$[Rif - H_2 - MeOH]^+$
790.37660	Rif	$[Rif - H_2 - MeOH]^+$, 1. Isotop
791.38776	Rif	$[Rif - MeOH]^+$
792.39245	Rif	$[Rif - MeOH]^+$, 1. Isotop
821.39844	Rif	$[Rif - H_2 + H]^+$
822.40266	Rif	$[Rif - H_2 + H]^+$, 1. Isotop
823.41301	Rif	$[Rif + H]^+$
824.41803	Rif	$[Rif + H]^+$, 1. Isotop
825.41986	Rif	$[Rif + H]^+$, 2. Isotop
843.38046	Rif	$[Rif - H_2 + Na]^+$
845.39545	Rif	$[Rif + Na]^+$
846.39878	Rif	$[Rif + Na]^+$, 1. Isotop
847.40226	Rif	$[Rif + Na]^+$, 2. Isotop
859.35407	Rif	$[Rif - H_2 + K]^+$
860.35701	Rif	$[Rif - H_2 + K]^+$, 1. Isotop
861.36967	Rif	$[Rif + K]^+$
862.37371	Rif	$[Rif + K]^+$, 1. Isotop
863.37184	Rif	$[Rif + K]^+$, 2. Isotop
864.37215	Rif	$[Rif + K]^+$, 3. Isotop
865.37791	Rif	$[Rif + K]^+$, 3. Isotop (?)
877.36099	Rif	N.A. aus Rif
926.50933	Rif	N.A. aus Rif
927.51288	Rif	N.A. aus Rif
158.11778	Ery	$[Desosamin + H]^+$ aus Ery

Tabelle A.1.: Liste der Antibiotikapeaks im Positiv-Ionen-Modus (Fortsetzung)

m/z	Antibiotikum	Annotation
166.26314	Ery	N.A. aus Ery
192.13888	Ery	N.A. aus Ery
244.83707	Ery	N.A. aus Ery
244.95506	Ery	N.A. aus Ery
245.17896	Ery	N.A. aus Ery
522.34538	Ery	[Fragment EryA + H]$^+$
540.35613	Ery	[Fragment EryA + H]$^+$
558.36686	Ery	[AEryA - Cladinose + H]$^+$
559.37040	Ery	[AEryA - Cladinose + H]$^+$, 1. Isotop
576.37727	Ery	[EryA - Cladinose + H]$^+$
577.38130	Ery	[EryA - Cladinose + H]$^+$, 1. Isotop
578.38406	Ery	[EryA - Cladinose + H]$^+$, 2. Isotop
716.46228	Ery	N.A. aus Ery
718.47814	Ery	[EryB + H]$^+$
720.45752	Ery	[EryC / NdeMeEryA + H]$^+$
734.46781	Ery	[EryA + H]$^+$
735.47191	Ery	[EryA + H]$^+$, 1. Isotop
736.47801	Ery	[EryA + H]$^+$, 2. Isotop
737.48154	Ery	[EryA + H]$^+$, 3. Isotop
748.44816	Ery	[EryE + H]$^+$
750.46766	Ery	[EryF / EryAEO + H]$^+$
772.42250	Ery	[EryA + K]$^+$
773.42633	Ery	[EryA + K]$^+$, 1. Isotop

[a] Diese Fragmente können bei der Fragmentierung verschiedener Erythromycin-Spezies entstehen ([324])

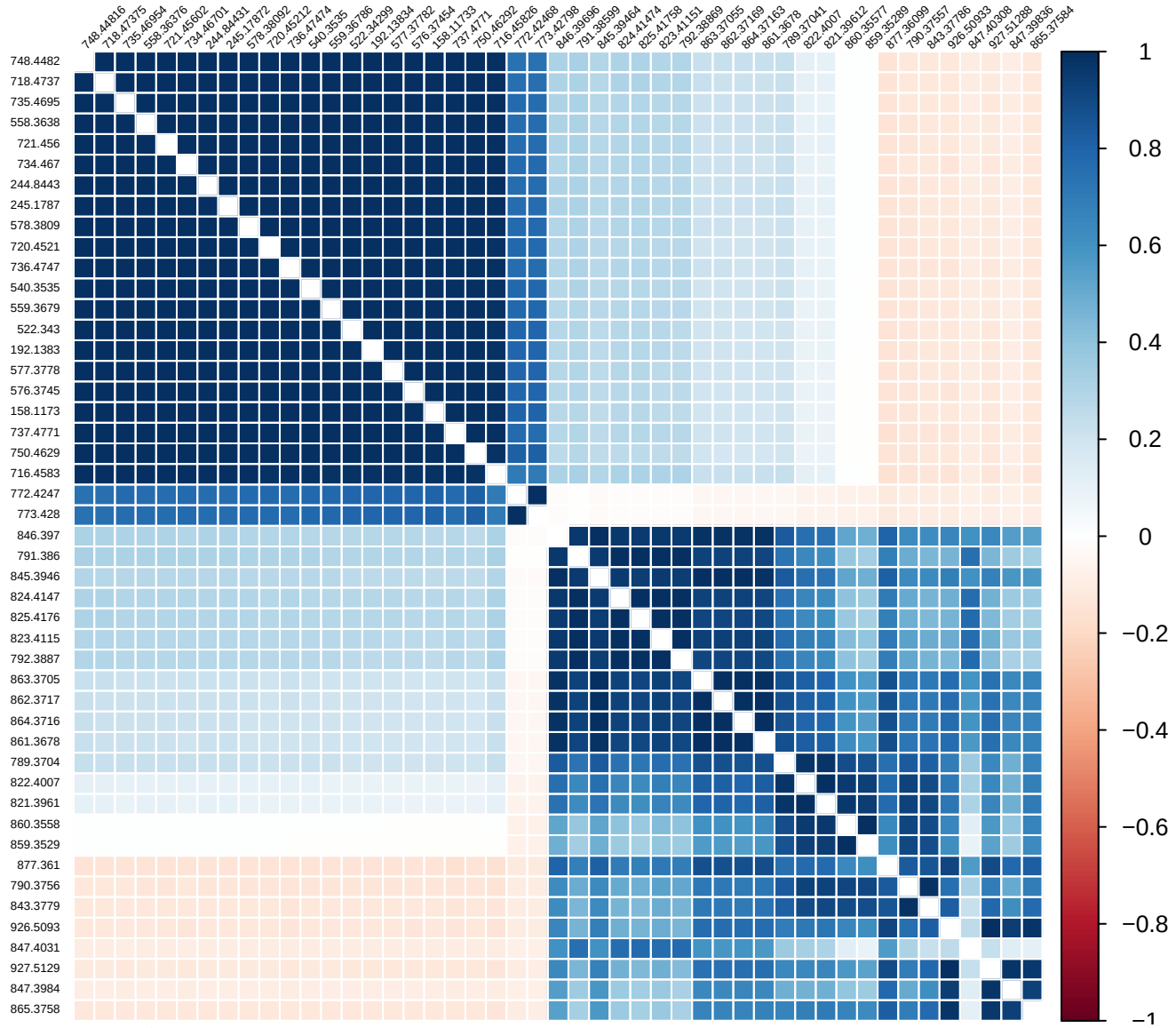


Abbildung A.1.: Pearson-Korrelation der Antibiotikapeaks mit hierarchischem Clustering nach der Ward-Methode

B. Ergebnistabellen Proof of Concept

Auf den folgenden Seiten sind die Ergebnisse des Proof-Of-Concept Experimentes dargestellt. Zur Berechnung der statistischen Signifikanz wurden 10000 Permutationen durchgeführt. Das Signifikanzniveau beträgt 5 %.

- EryA - Erythromycin A
- AEryA - Anhydromerythromycin A
- EryAEO - Erythromycin A *N*-oxid
- NdMeEryA - *N*-demethylerythromycin A
- EryB - Erythromycin B
- EryC - Erythromycin C
- EryE - Erythromycin E
- EryF - Erythromycin F
- Rif - Rifampicin

B.1. Positiv Modus

Tabelle B.1.: Hitliste des Proof of Concept Experiments - Positiv-Ionen-Modus

Rang	Korrelations-koeffizient	m/z	Probennummer [Gesamtzahl der Peaks]	Annotation	Δ ppm
1	0.687	822.4007	1, 2, 4, 13, 15, 16	$[Rif - H_2 + H]^+$, 1. Isotop	0,7
2	0.674	821.39612	1, 2, 4, 5, 15, 16	$[Rif - H_2 + H]^+$	0,8
3	0.670	862.37169	1, 2, 4, 5, 15, 16	$[Rif + K]^+$, 1. Isotop	0,07
4	0.660	845.39464	1, 2, 4, 15, 16	$[Rif + Na]^+$	0,4
5	0.660	846.39695	1, 2, 4, 15, 16	$[Rif + Na]^+$, 1. Isotop	0,9
6	0.658	791.38599	1, 2, 4, 15, 16	$[Rif - MeOH]^+$	0,3

Tabelle B.1.: Hitliste des Proof of Concept Experiments - Positiv-Ionen-Modus (Fortsetzung)

Peak	Korrelations-koeffizient	m/z	Probennummer [Gesamtzahl der Peaks]	Annotation	Δ ppm
7	0.658	792.38699	1, 2, 4, 15, 16	[Rif - MeOH]$^+$, 1. Isotop	3,2
8	0.658	859.35289	1, 2, 4, 5, 14, 15, 16	[Rif - H_2 + K]$^+$	0,3
9	0.658	860.35577	1, 2, 4, 5, 14, 15, 16	[Rif - H_2 + K]$^+$, 1. Isotop	0,2
10	0.657	287.15266	1, 2, 4, 15, 16	?	-
11	0.657	863.37055	1, 2, 4, 15, 16	[Rif + K]$^+$, 2. Isotop	5,1
12	0.657	864.37163	1, 2, 4, 15, 16	[Rif + K]$^+$, 3. Isotop	7,8
13	0.655	861.3678	1, 2, 4, 5, 14, 15, 16	[Rif + K]$^+$	0,6
14	0.653	823.41151	1, 2, 4, 15, 16	[Rif + H]$^+$	1,1
15	0.653	824.41474	1, 2, 4, 15, 16	[Rif + H]$^+$, 1. Isotop	1,2
16	0.653	825.41758	1, 2, 4, 15, 16	[Rif + H]$^+$, 2. Isotop	1,8
17	0.603	789.37041	1, 2, 15, 16	[Rif - H_2 - MeOH]$^+$	0,2
18	0.592	287.48765	1, 2, 15, 16	[$C_{12}H_{24}O_6$ + Na]$^+$	-
19	0.592	877.36099	1, 2, 15, 16	NA aus Rif	-
20	0.592	899.32288	1, 2, 15, 16	NA	-
21	0.591	1043.4458	1, 2, 15, 16	NA	-
22	0.581	924.495	1, 2, 5, 15, 16	?	-
23	0.580	926.50933	1, 2, 15, 16	NA aus Rif	-
24	0.580	927.51288	1, 2, 15, 16	NA aus Rif *m/z* 926,50933 , 1. Isotop	-
25	0.554	1041.43219	2, 15, 16	NA	-
26	0.545	790.37557	2, 15 ,15	[Rif - H_2 - MeOH]$^+$, 1. Isotop	2,0
27	0.545	843.37786	2, 15 ,15	[Rif - H_2 + Na]$^+$	1,0
28	0.499	900.32745	1, 2, 16	NA	-
29	0.498	1044.44626	1, 2, 16	NA	-
30	0.463	859.34924	1, 2, 5, 14, 16	NA	-
31	0.458	865.37021	2, 15	NA	-
32	0.458	1041.43823	2, 15	NA	-
33	0.454	925.49872	1, 15, 16	NA	-
34	0.449	865.37584	1, 15, 16	[Rif + K]$^+$, 4. Isotop	6,7
35	0.429	735.46954	2, 4, 7, 9, 13, 14	[EryA + H]$^+$, 1. Isotop	3,3
36	0.420	773.42798	4, 9	[EryA + K]$^+$, 1. Isotop	0,3
37	0.411	158.11733	4, 9	[Desosamin + H]$^+$ aus Ery	1,5

Tabelle B.1.: Hitliste des Proof of Concept Experiments - Positiv-Ionen-Modus (Fortsetzung)

Peak	Korrelations-koeffizient	m/z	Probennummer [Gesamtzahl der Peaks]	Annotation	Δ ppm
38	0.411	192.13834	4, 9	NA aus Ery	-
39	0.411	244.84432	4, 9	NA aus Ery	-
40	0.411	245.17872	4, 9	NA Ery	-
41	0.411	522,34299[a]	4, 9	[Fragment Ery + H]$^+$	-
42	0.411	540,35349[a]	4, 9	[Fragment Ery + H]$^+$	-
43	0.411	558.36376	4, 9	[AEryA - Cladinose + H]$^+$	0,2
44	0.411	559.36786	4, 9	[AEryA - Cladinose + H]$^+$, 1. Isotop	1,5
45	0.411	578.38092	4, 9	[EryA - Cladinose + H]$^+$, 2. Isotop	0
46	0.411	718.47376	4, 9	[EryB + H]$^+$	1,0
47	0.411	720.45212	4, 9	[EryC / NdeMeEryA + H]$^+$	0,3
48	0.411	721.45602	4, 9	[EryC / NdeMeEryA + H]$^+$, 1. Isotop	0,5
49	0.411	748.44816	4, 9	[EryE + H]$^+$	0,5
50	0.411	750.46292	4, 9	[EryF / EryAEO + H]$^+$	0,7
51	0.402	195.79219	15, 16	NA	-
52	0.402	286.48032	15, 16	NA	-
53	0.402	891.37364	15, 16	NA	-
54	0.402	1171.62964	15, 16	NA	-
55	0.398	175.1187	HE1, HE2, HE3, HE4, HE5, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16		-
56	0.394	274.49867	1, 2	NA	-
57	0.394	847.40308	1, 2	[Rif + K]$^+$, 2. Isotop	2,3
58	0.389	396.25982	HE1, HE3, 5, 6, 7, 9, 1, 12, 13, 15, 16	NA	-
59	0.387	365.15754	HE1, 5, 9, 14, 15	NA	-
60	0.383	1045.45418	1, 2	NA	-
61	0.381	772.42468	4, 9, 13, 14	[EryA + K]$^+$	0,3
62	0.375	576.37454	4, 9, 13, 14	[EryA - Cladinose + H]$^+$	0,6
63	0.371	734.46701	2, 3, 4, 5, 6, 7, 9, 12, 13, 14	[EryA + H]$^+$	2,1

[a] Diese Fragmente können bei der Fragmentierung verschiedener Erythromycin-Spezies entstehen (ref Volmer)

B.2. Negativ Modus

Tabelle B.2.: Hitliste des Proof of Concept Experiments - Negativ-Ionen-Modus

Rang	Korrelations-koeffizient	m/z	Probennummer [Gesamtzahl der Peaks]	Annotation	Δ ppm
1	0.690802455	822.40186	1, 2, 4, 5, 8, 9, 13, 14, 15, 16	$[Rif - H]^-$ 1. Isotop	0.8
2	0.656835972	821.39715	1, 2, 4, 5, 8, 9, 15, 16	$[Rif - H]^-$	0.8
3	0.654285174	819.38452	1, 2, 4, 15, 16	$[Rif - H_2 -H]^-$	2.8
4	0.64906134	273.83055	1, 2, 4, 15, 16	NA aus Rif	-
5	0.64906134	274.16403	1, 2, 4, 15, 16	*m/z* 273.83055, 1. Isotop	-
6	0.64906134	824.40924	1, 2, 4, 15, 16	$[Rif - H]^-$, 3. Isotop	1.6
7	0.647320063	837.39495	1, 2, 4, 15, 16	$[Rif + O - H]^-$	2.5
8	0.647320063	838.39782	1, 2, 4, 15, 16	$[Rif + O - H]^-$, 1. Isotop	1.9
9	0.645578785	724.29981	1, 2, 4, 15, 16	NA	-
10	0.645578785	725.30323	1, 2, 4, 15, 16	NA	-
11	0.628823351	820.39017	1, 2, 4, 14, 15, 16	$[Rif - H_2 -H]^-$, 1. Isotop	5.5
12	0.584519157	721.30026	1, 2, 15, 16	NA	-
13	0.564592368	893.4415	1, 2, 4, 15	NA	-
14	0.562694578	853.42885	1, 2, 4, 15	NA aus Rif	-
15	0.557950104	851.4135	1, 2, 4, 15	NA aus Rif	-
16	0.540530594	823.40575	1, 2, 4, 5, 8, 14, 15, 16	$[Rif - H]^-$, 2. Isotop	1.5
17	0.47793837	377.08646	E4, 2, 3, 9, 13, 16	NA	-
18	0.457846704	164.30216	1, 2, 4	NA	-
19	0.452504269	838.40225	1, 2, 4	NA	-
20	0.452504269	839.40663	1, 2, 4	$[Rif + O -H]^-$, 2. Isotop	8.4
21	0.452504269	853.43427	1, 2, 4	NA	-
22	0.450367295	274.50057	1, 2, 4	NA	-
23	0.447626097	315.09332	HE1, HE2, HE3, HE4, HE5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 11, 12, 13, 14, 15, 16	NA	-
24	0.420003803	854.43646	2, 4	NA	-
25	0.411067552	852.42177	2, 4	NA	-
26	0.411067552	885.45894	2, 4	NA	-

Tabelle B.2.: Hitliste des Proof of Concept Experiments - Negativ-Ionen-Modus (Fortsetzung)

Peak	Korrelations-koeffizient	m/z	Probennummer [Gesamtzahl der Peaks]	Annotation	Δ ppm
27	0.41002196	457.12092	HE1, HE2, 2, 3, 4, 6, 8, 9, 1, 11, 12, 13, 14, 15, 16	NA	-
28	0.406025082	795.21706	2, 3, 9	NA	-
29	0.39474021	855.20077	HE1, 2, 4, 5	NA	-
30	0.394471657	273.94216	1, 2	NA	-
31	0.363833081	835.37337	4, 15	NA	-
32	0.363833081	885.45326	4, 16	NA	-

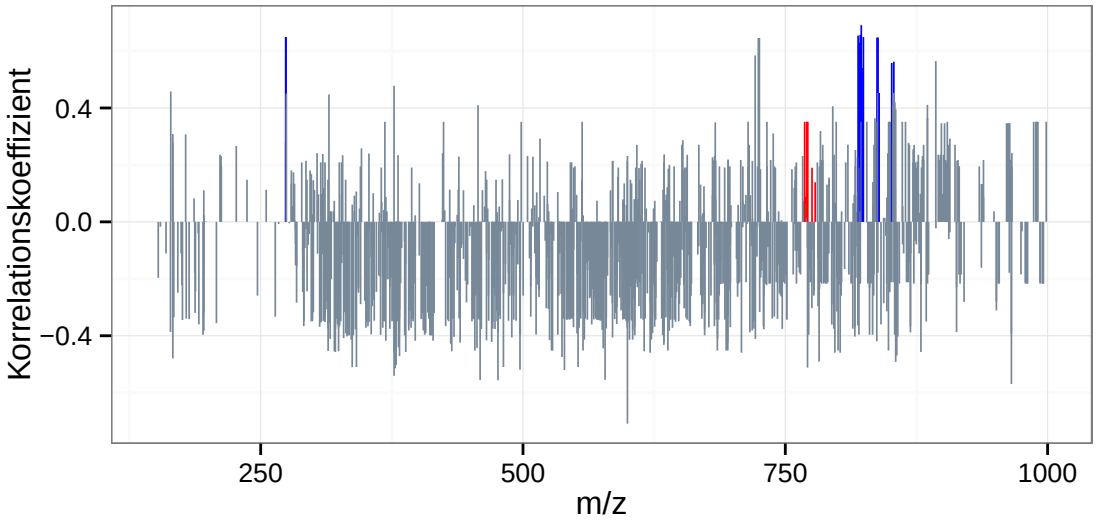


Abbildung B.1.: Plot der Korrelationskoeffizienten aus der Aktivitäts-Korrelations-Analyse gegen *m/z*. Peaks aus Erythromycin (Rot), Rifampicin (Blau) und den nicht-aktiven Substanzen (Grau) sind farblich gekennzeichnet.

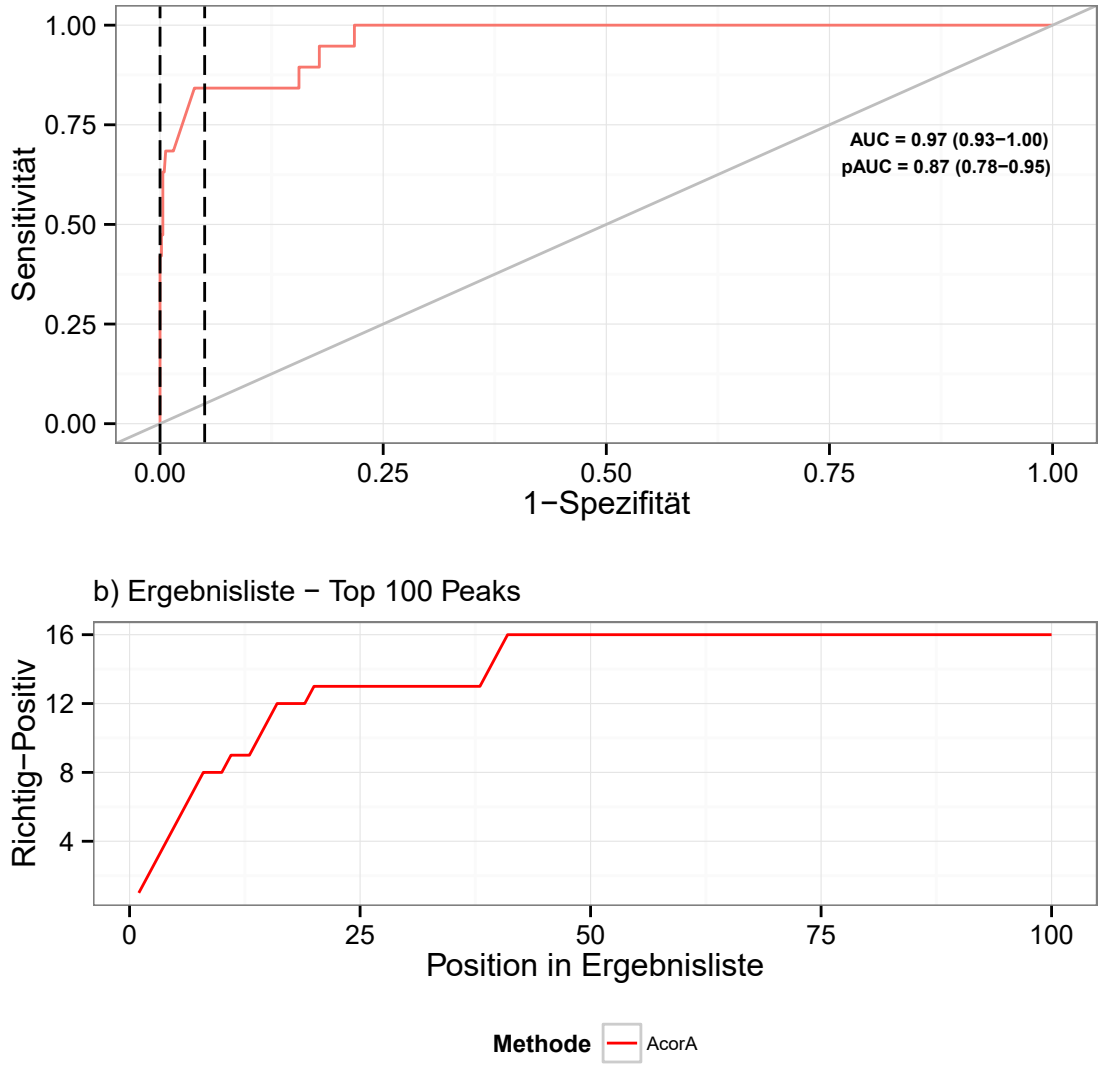


Abbildung B.2.: ROC-Kurve der AcorA Ergebnisliste des Negativ-Ionen-Modus.

B.3. Kovarianz-Korrelations-Diagramm Proof of Concept

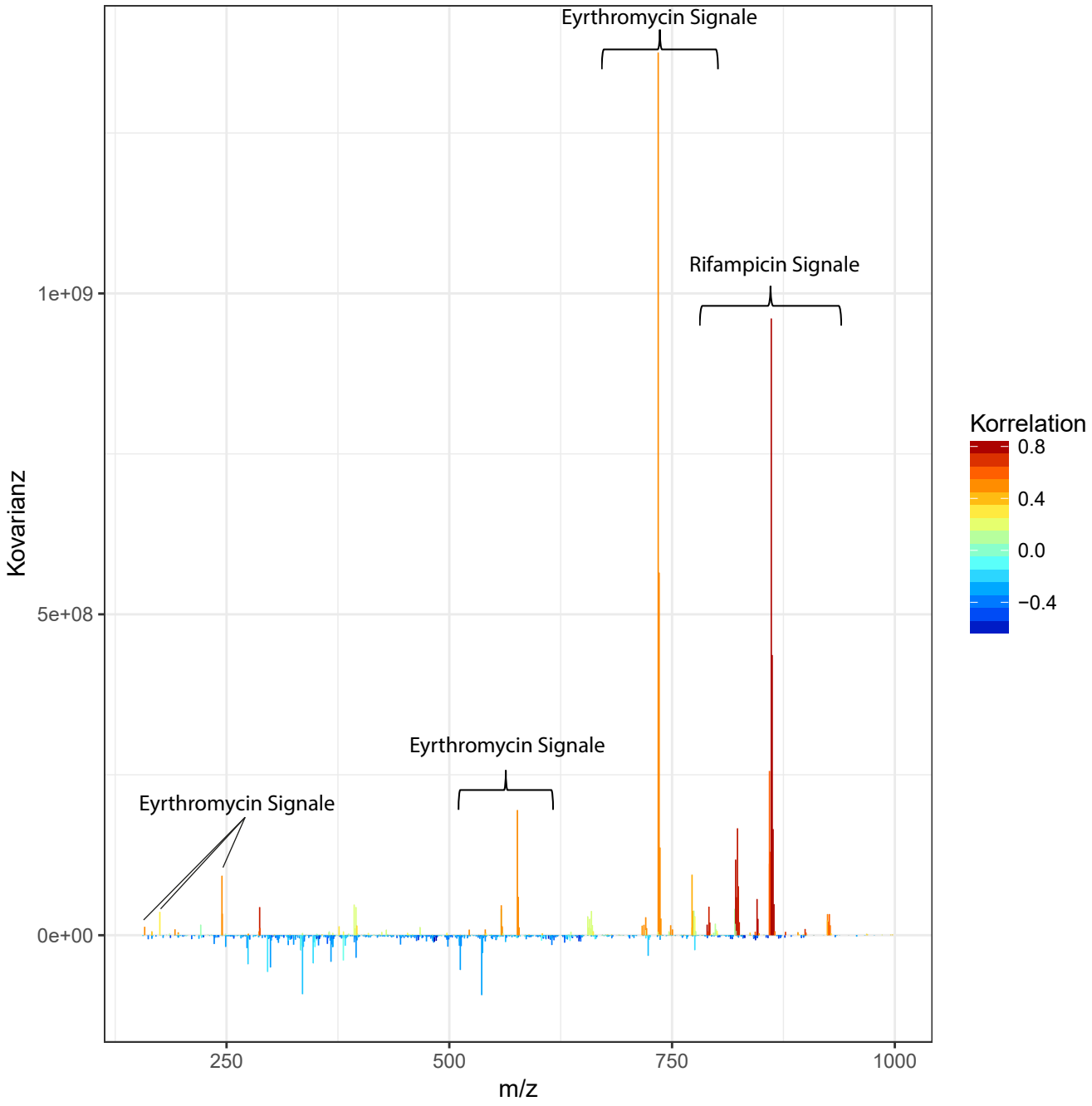


Abbildung B.3.: Kovarianz-Korrelations-Diagramm Proof of Concept.

C. AcorA Ergebnisse Sepedonium ampullosporum

C.1. Hitliste aller Extrakte

Tabelle C.1.: Hitliste *S. ampullosporum*

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
1	0.638	327.20266	$C_{18}H_{28}O_2Na_1^+$	
2	0.613	756.28916	NA	NA
3	0.583	274.35241	$C_{16}H_{36}NO_2^+$	
4	0.573	815.96545	Ampullosporin B/C/D $[M+H+Na]^{2+}$	1
5	0.542	398.25115	NA	NA
6	0.535	367.26508	NA, $[344.279565+Na]^+$	NA
7	0.517	714.93095	Peptaibol	
8	0.510	653.39805	**63**,**64** b_7^+	
9	0.506	698.94252	**63**,**64**,**65**,**66** $[M+2H]^{2+}$	2
10	0.506	1646.93261	**61** $[M+Na]^+$	
11	0.505	212.13931	NA	
12	0.503	502.28992	NA	NA
13	0.500	722.43992	**63**,**64**,**65**,**66** $[M + 2Na]^{2+}$,Isotop	2
14	0.500	687.13245	NA	NA
15	0.499	764.46312	**61**, **63**, **65** $[y_7+Na]^+$	1,2
16	0.495	719.92310	**63**,**64**,**65**,**66** $[M+ 2Na]^{2+}$	2
17	0.495	742.93932	Peptaibol	2
18	0.494	236.33030	NA	
19	0.494	716.91735	**63**,**64**,**65**,**66** $[M+H+K]^{2+}$	2
20	0.494	721.94010	**63**,**64**,**65**,**66** $[M+2Na]^{2+}$	2

Tabelle C.1.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
21	0.494	733.92769	Peptaibol	2
22	0.493	668.37645	**61**, **62** , b_6^+	1
23	0.489	824.97949	**61**, **62** $[M+H+Na]^{2+}$, Isotop	1
24	0.489	860.51560	Peptaibol	
25	0.487	727.39291	**61**, **62** $[M+Na+K]^{2+}$	1
26	0.487	679.41109	NA	
27	0.487	978.57298	**61**, **63**, **66** $[y_9+Na]^+$	1,2
28	0.484	811.98432	**61**, **62** $[M+2H]^{2+}$	
29	0.482	812.48674	**61**, **62** $[M+2H]^{2+}$, Isotop	
30	0.480	355.23395	**63**, **64** b_3^+	
31	0.479	717.41788	**63**,**64**,**65**,**66** $[M+H+K]^{2+}$, Isotop	
32	0.474	743.40390	Peptaibol	2
33	0.474	743.90703	Peptaibol	2
34	0.473	793.22648	NA	NA
35	0.472	742.43595	Peptaibol	2
36	0.470	1045.12727	NA	
37	0.469	812.98671	**61** $[M+2H]^{2+}$, Isotop	s
38	0.469	1623.97827	**61**, **62** $[M+H]^+$, Isotop	
39	0.467	955.59093	**61**, **63**, **66** y_9^+	
40	0.466	697.93985	**63**,**64**,**65**,**66** $[M+2H]^{2+}$	
41	0.466	698.44062	**63**,**64**,**65**,**66** $[M+2H]^{2+}$, Isotop	
42	0.463	1415.82756	**63**,**64**,**65**,**66** $[M+Na]^+$	2
43	0.462	708.41525	**63**,**64**,**65**,**66** $[M+H+Na]^{2+}$	
44	0.461	594.35893	**61**, **63**, **65** $[y_5+Na]^+$	1,2
45	0.461	860.02054	Peptaibol	
46	0.460	507.24257	NA	NA
47	0.460	709.93188	**63**,**64**,**65**,**66** $[M+H+Na]^{2+}$, Isotop	2
48	0.460	710.43349	**63**,**64**,**65**,**66** $[M+H+Na]^{2+}$, Isotop	2
49	0.457	831.45911	**61**, **62** $[M+H+K]^{2+}$, Isotop	1
50	0.457	831.96091	**61**, **62** $[M+H+K]^{2+}$, Isotop	1
51	0.454	718.33736	NA	NA

Tabelle C.1.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
52	0.452	246.18145	**61**, **63**, **65** y_2^+	1,2
53	0.452	1395.87035	**63**,**64**,**65**,**66** $[M+H]^+$, Isotop	2
54	0.451	671.13412	NA	NA
55	0.449	977.57270	**61**, **63**, **66** $[y_9+Na]^+$	1,2
56	0.445	440.28633	**63**, **64** b_5^+	2
57	0.444	344.27957	NA	NA
58	0.443	170.13219	NA	NA
59	0.443	1447.84605	NA	
60	0.441	641.87611	NA	
61	0.439	1417.84262	**63**,**64**,**65**,**66** $[M+Na]^+$, Isotop	2
62	0.439	485.26105	NA	NA
63	0.438	830.95756	**61**, **62** $[M+H+K]^{2+}$	1
64	0.436	326.37863	NA	NA
65	0.435	735.07813	NA	NA
66	0.432	745.06826	NA	NA
	0.430	482.27548	**61**, **63**, **66** $[y_4+K]^+$	1,2
68	0.428	822.97537	**61**, **62**$[M+H+Na]^{2+}$	1
69	0.428	736.92402	Peptaibol	2
70	0.426	859.51663	Peptaibol	1
71	0.424	261.12334	NA	NA
72	0.423	734.41956	Peptaibol	2
73	0.422	823.47613	**61**, **62**$[M+H+Na]^{2+}$, Isotop	1
74	0.422	823.97743	**61**, **62**$[M+H+Na]^{2+}$, Isotop	1
75	0.413	719.10436	NA	NA
76	0.413	741.93452	Peptaibol	2
77	0.412	361.27815	NA	NA
78	0.412	708.93175	**63**,**64**,**65**,**66** [M+H+Na]2+, Isotop	2
79	0.412	709.43160	**63**,**64**,**65**,**66** [M+H+Na]2+, Isotop	2
80	0.412	426.30772	NA	
81	0.410	824.47774	**61**, **62**$[M+H+Na]^{2+}$, Isotop	1
82	0.410	590.39130	NA	NA

Tabelle C.1.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
83	0.407	1418.85151	**63**,**64**,**65**,**66** $[M+Na]^+$, Isotop	2
84	0.406	1645.93715	**61**, **62** $[M+H]^+$, Isotop	1
85	0.405	1644.93916	**61**, **62** $[M+H]^+$	1
86	0.404	470.23953	**61**, **62** b_4^+	1
87	0.400	331.23420	**61**, **63**, **65** y_3^+	1,2
88	0.400	362.29012	Na, Isotop von 361.27815	NA
89	0.399	827.95858	NA	NA
90	0.398	478.29224	NA	NA
91	0.398	842.45599	**61**, **62** [M+Na+K]2+, Isotop	1
92	0.398	842.95677	**61**, **62** [M+Na+K]2+, Isotop	1
93	0.398	375.28581	NA	NA
94	0.397	383.26940	NA	NA
95	0.393	274.51954	NA	
96	0.393	572.37688	**61**, **63**, **65** y_5^+	1,2
97	0.387	741.32094	NA	NA
98	0.387	843.95768	**61**, **62** $[M+Na+K]^{2+}$, Isotop	1
99	0.383	1044.62590	NA	
100	0.382	720.42318	**63**,**64**,**65**,**66** $[M+2Na]^{2+}$, Isotop	2
101	0.382	720.92410	**63**,**64**,**65**,**66** $[M+2Na]^{2+}$, Isotop	2
102	0.382	732.93118	Peptaibol	2
103	0.382	733.43159	Peptaibol	2
104	0.380	345.28311	NA, Isotop von 344.27956	NA
105	0.380	740.31870	NA	NA
106	0.380	841.95419	**61**, **62** $[M+Na+K]^{2+}$	1
107	0.380	843.45603	**61**, **62** $[M+Na+K]^{2+}$, 3. Isotop	1
108	0.378	1062.62764	**61**, **63**, **65** y_{10}^+	1,2
109	0.375	727.91130	**63**,**64**,**65**,**66** $[M+Na+K]^{2+}$	1,2
110	0.372	839.44027	Peptaibol	
111	0.372	896.50414	NA	NA
112	0.371	703.10591	NA	
113	0.370	244.07773	NA	NA

Tabelle C.1.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
114	0.370	671.40891	NA	
115	0.368	1416.84993	**63**,**64**,**65**,**66** $[M+Na]^+$	2

C.2. Hitliste 2

Tabelle C.2.: Ergebnisse AcorA Amp II. Assoziierte Peaks sind farblich gekennzeichnet. Peakcluster 3: rot, Peakcluster 4: blau, Peakcluster 5: grün.

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
1	0.7503	591.35007	**67**, **68**, **69**, **70**, **71**, **72** y_5^+	3,4
2	0.7501	327.20218	**69**, **70**, b_4^+	NA1
3	0.7160	409.18379	NA	NA1
4	0.7138	488.34299	NA	NA1
5	0.7017	371.22767	NA	NA1
6	0.6853	454.21436	NA	NA1
7	0.6746	756.28912	NA	
8	0.6491	896.00835	**69**, **71**, **72** $[M + H + Na]^{2+}$	4
9	0.6479	375.28581	NA	
10	0.6469	897.00433	**69**, **70**, **71**, **72** $[M + H + Na]^{2+}$	4
11	0.6469	896.50441	**69**, **70**, **71**, **72** $[M + H + Na]^{2+}$	4
12	0.6363	576.39650	NA	NA1
13	0.6286	701.39288	NA	
14	0.6242	453.21021	NA	NA1
15	0.6108	400.29082	NA	NA1
16	0.6101	1218.67084	NA	
17	0.6089	724.89603	**73** $[M + H+ Na]^{2+}$	5
18	0.6081	755.42104	NA	NA1
19	0.6056	512.37069	**24** $[M+ H]^+$, Isotop	6
20	0.6048	365.15773	NA	NA1
21	0.6010	718.33775	NA	
22	0.6010	585.28791	NA	NA1

Tabelle C.2.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
23	0.5953	478.29233	NA	
24	0.5944	541.26208	NA	NA1
25	0.5921	713.09809	NA	NA1
26	0.5916	885.51575	**69**, **72**$[M + 2H]^{2+}$	4
27	0.5907	440.33532	439.33229, Isotop	
29	0.5895	1021.54011	**67**, **68** $[M + 2Na]^{2+}$	3
28	0.5895	459.28107	NA	NA1
30	0.5861	503.30680	NA	NA1
31	0.5852	1022.04101	**67**, **68** $[M + 2Na]^{2+}$, Isotop	3
32	0.5835	613.33231	[**67**, **68**, **69**, **70**, **71**, **72** y_5+Na]$^+$	3,4
33	0.5780	629.31307	NA	NA1
34	0.5777	398.25109	NA	
35	0.5734	388.25470	NA	NA1
36	0.5726	1011.54994	**67**, **68** $[M + H + Na]^{2+}$	3
37	0.5680	367.26525	NA	
38	0.5667	627.48061	NA	
39	0.5661	367.19554	NA	
41	0.5576	1012.05208	**67**, **68** $[M + H + Na]^{2+}$	3
40	0.5576	641.19779	NA	
42	0.5565	415.25443	NA	NA1
45	0.5551	498.23978	NA	
44	0.5551	365.21625	**67**, **69**, **71** [y_3+Na]$^+$	3,4
43	0.5551	306.16080	305.157222, 1. Istotop	NA1
46	0.5505	735.88374	**73** $[M + 2Na]^{2+}$, Isotop	5
47	0.5498	886.01703	**69**, **70**, **71**, **72** $[M + 2H]^{2+}$	4
49	0.5468	999.56134	**68** $[M + 2H]^{2+}$	3
48	0.5468	885.01784	**69**, **72**, **71** $[M + 2H]^{2+}$	4
50	0.5445	362.29017	NA	
51	0.5417	343.23418	**67**, **69**, **71** y_3^+	3,4
52	0.5378	904.48825	**69**, **70**, **71**, **72** $[M + H + K]^{2+}$	4
53	0.5331	366.20046	**68**, **70**, **72** [y_3+Na]$^+$	

Tabelle C.2.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
54	0.5299	597.12326	NA	NA1
55	0.5298	991.58120	NA	
56	0.5213	745.06859	NA	
57	0.5138	410.18817	NA	NA1
58	0.5125	510.36517	**24** $[M+ H]^+$	6
60	0.5122	767.05210	NA	NA1
59	0.5122	629.14878	NA	
64	0.5110	1447.76615	NA	
63	0.5110	1446.77090	NA	
62	0.5110	734.87859	**73** $[M + 2Na]^{2+}$	5
61	0.5110	344.23764	**68**, **70**, **72** y_3^+	3,4
65	0.5094	305.15722	NA	NA1
66	0.5085	907.99831	**69**, **70**, **71**, **72** $[M + 2Na]^{2+}$	4
68	0.5081	476.30699	NA	NA1
67	0.5081	432.28096	NA	NA1
69	0.5046	907.00073	**69**, **70**, **71**, **72** $[M + 2Na]^{2+}$	4
70	0.5032	549.32416	**24** $[M+ K]^+$, Isotop	
71	0.4942	519.33542	NA	
72	0.4915	439.33229	NA	
73	0.4914	1010.54763	**67**, **68** $[M +H + Na]^{2+}$	3
75	0.4839	688.36012	NA	
74	0.4839	564.35990	NA	NA1
76	0.4823	657.36315	NA	
77	0.4815	511.36895	**24** $[M+ H]^+$, Isotop	6
78	0.4805	735.07776	NA	NA1
79	0.4788	725.39727	**73** $[M + H + Na]^{2+}$, Isotop	5
80	0.4788	344.21821	NA	
81	0.4781	526.29229	NA	NA1
82	0.4769	566.29801	NA	
83	0.4678	361.27822	NA	
84	0.4661	1010.05116	**67** $[M +H + Na]^{2+}$	3

Tabelle C.2.: Hitliste *S. ampullosporum* (Fortsetzung)

Rang	Korrelations-koeffizient	m/z	Annotation	Cluster
85	0.4653	608.38631	NA	NA1
86	0.4647	393.20972	NA	NA1
87	0.4597	509.27221	NA	
88	0.4592	574.14855	NA	
89	0.4517	345.28323	NA	
90	0.4513	592.33314	**68**, **70**, **72** y_5^+	3,4
92	0.4490	1031.02868	**67**, **68** $[M + Na + K]^{2+}$, Isotop	3
91	0.4490	898.00448	**69**, **70** $[M + H + Na]^{2+}$	4
93	0.4486	719.10415	NA	NA1
94	0.4484	346.29544	NA	

C.3. Ampullosporin A (61)

C.3.1. Annotation der AmpA Signale aus Hitlisten

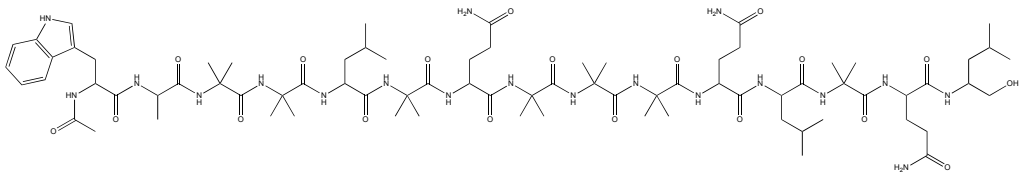
Tabelle C.3.: AcorA Ergebnisse Ampullosporin A

Peak	Korrelations-koeffizient	m/z	Annotation	Δ ppm	Modus	Probennummer
10	0,506	1646.93261	$[AmpA + Na]^+$, 2. Isotop	11,5	+	531, 533, 537, 561
23	0,489	824.97949	$[AmpA + H + Na]^{2+}$, 4. Isotop	3,9	+	537, 561
28	0,484	811.98432	$[AmpA + 2H]^{2+}$	0,9	+	531, 533, 537, 561, 635, 641
29	0,482	812.48674	$[AmpA + 2H]^{2+}$, 1. Isotop	0	+	531, 533, 537, 561, 635, 641
31	0,441	1658.91413	$[AmpA + {}^{37}Cl]^-$	4,4	-	??
37	0,469	812.98671	$[AmpA + 2H]^{2+}$, 2. Isotop	2,1	+	533, 561, 635
38	0,469	1623.97827	$[AmpA + H]^+$, 1. Isotop	7,4	+	533, 561, 635
48	0,377	809.97326	$[AmpA - 2H]^{2-}$	3,4	-	??
49	0,457	831.45911	$[AmpA + H + K]^{2+}$, 1. Isotop	6,7	+	531, 533, 537, 561, 635, 641
50	0,457	831.96091	$[AmpA + H + K]^{2+}$, 2. Isotop	6,5	+	531, 533, 537, 561, 635, 641

Tabelle C.3.: AcorA Ergebnisse Ampullosporin A (Fortsetzung)

Peak	Korrelations-koeffizient	m/z	Annotation	Δ ppm	Modus	Probennummer
55	0,367	1622.97959	$[AmpA - H]^-$, 2. Isotop	15,2	-	??
58	0,360	811.47141	$[AmpA - 2H]^{2-}$, 3. Isotop	5,1	-	??
63	0,438	830.95756	$[AmpA + H + K]^{2+}$	6,5	+	531, 533, 537, 561, 635, 641
68	0,428	822.97537	$[AmpA + H + Na]^{2+}$	0,8	+	531, 533, 534, 537, 559, 561, 635, 641
73	0,422	823.47613	$[AmpA + H + Na]^{2+}$, 1. Isotop	1,9	+	531, 533, 534, 537, 559, 561, 635, 641
74	0,422	823.97743	$[AmpA + H + Na]^{2+}$, 2. Isotop	2,4	+	531, 533, 534, 537, 559, 561, 635, 641
81	0,410	824.47774	$[AmpA + H + Na]^{2+}$, 3. Isotop	4,0	+	531, 533, 534, 537, 559, 561, 635, 641
84	0,406	1645.93715	$[AmpA + Na]^+$, 1. Isotop	6,7	+	531, 533, 534, 537, 559, 561, 635, 641
85	0,405	1644.93916	$[AmpA + Na]^+$	3,4	+	531, 533, 534, 537, 559, 561, 635, 641
91	0,398	842.45599	$[AmpA + Na + K]^{2+}$, 1. Isotop	0,4	+	531, 533, 534, 537, 559, 561, 635, 641
92	0,398	842.95677	$[AmpA + Na + K]^{2+}$, 2. Isotop	0,6	+	531, 533, 534, 537, 559, 561, 635, 641
98	0,387	843.95768	$[AmpA + Na + K]^{2+}$, 4. Isotop	3,5	+	537, 561, 641
106	0,380	841.95419	$[AmpA + Na + K]^{2+}$	0,3	+	531, 533, 534, 537, 559, 561, 635, 641
107	0,380	843.45603	$[AmpA + Na + K]^{2+}$, 3. Isotop	3,5	+	531, 533, 534, 537, 559, 561, 635, 641

C.3.2. Charakterisierung Ampullosporin A (61)



Habitus	amorpher Feststoff
Summenformel, exakte Masse	$C_{77}H_{127}N_{19}O_{19}$, 1621.955563 g mol^{-1}
UV-Vis (MeOH)	λ_{max} (logϵ) 220 (2.1), 281 (1.4), 290 (1.3) nm
ORD	$[\alpha]_D^{25}$ = -26.3 (c 0.7, MeOH)
IR	(ATR) ν_{max}: 3300, 2945, 2871, 2829, 2360, 2158, 1650, 1530, 1455, 1385, 1363, 1288, 1221, 1172, 1102, 1023, 926, 741 cm^{-1}
ESI-FT-ICR-MS	*m/z* (Ion, rel. Int.) 811.98432 ([M+2H]2+, 2), 822.97500 ([M+H+Na]2+, 36), 830.95565 ([M+H+K]2+, 83), 833.96517 ([M+2Na]2+, 60), 1622.97393 (M+H]+, 26), 1644.94488 ([M+Na]+, 26), 1660.90217 ([M+K]+, 9)
ESI-IT-MS	*m/z* (Ion, rel. Int.) 1622.76 ([M+H]+, 100), 1644.94 ([M+Na]+, 11), 812.28 ([M+2H]2+, 50)
ESI-IT-MS^2 [812 (45 eV)]	*m/z* (Ion, rel. Int.) 1377.44 (b13+, 50), 1292.66 (b12+, 0.6), 1051.36 (b10+, 60), 966.34 (b9+, 46), 881.22 (b8+, 44),668.19 (b6+, 100), 583.04 (b5+, 14), 470.04 (b4+, 17), 384.96 (b3+, 6), 299.87 (b2+, 4), 229.11 (b1+, 0.4), 1153.64 (y11+, 10), 1040.31 (y10+, 8), 955.48 (y9+, 63), 826.77 (y8+, 6), 742.37 (y7+, 40), 657.35 (y6+, 47)
ESI-IT-MS^3 [812 (45 eV) → 881 (30 eV)]	*m/z* (Ion, rel. Int.) 796.14 (b7+, 3), 668.05 (b6+, 100), 583.23 (b5+), 470.04 (b4+, 13), 384.89 (b3+, 4)
ESI-IT-MS^3 [812 (45 eV) → 668 (35 eV)]	*m/z* (Ion, rel. Int.) 583.11 (b5+, 100), 470.01 (b4+, 43), 384.91 (b3+, 7), 300.15 (b2+, 1)
ESI-Qq-TOF-MS/MS	siehe Tabelle 3.6 Seite 173

C.4. Korrelationsanalysen

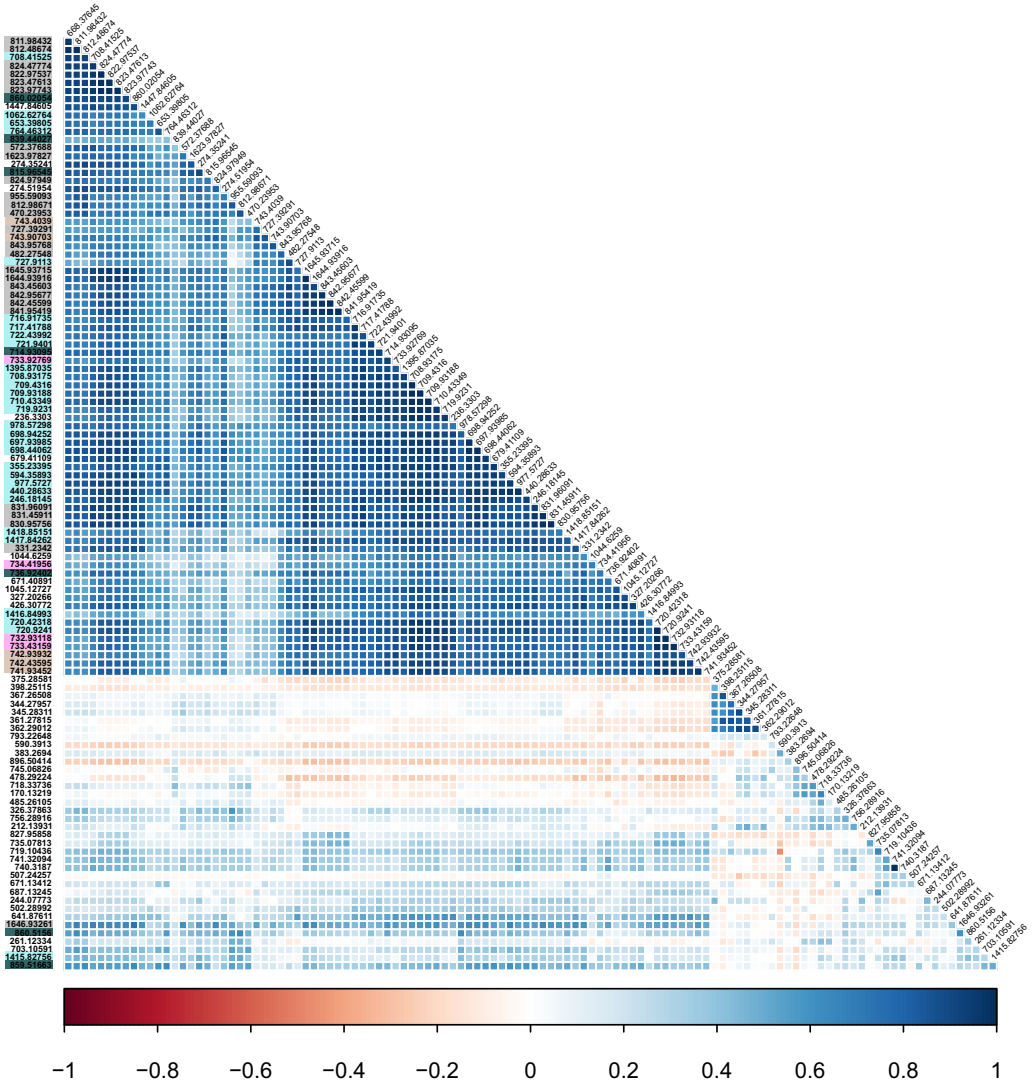


Abbildung C.1.: Pearson-Korrelationen zwischen den Peaks der Hitliste 1. Massensignale, die im Sinne von Isotopen- und Adduktpeaks assoziert sind, sind farblich hervorgehoben. Peakcluster 1: Grau, Peakcluster 2: Blau

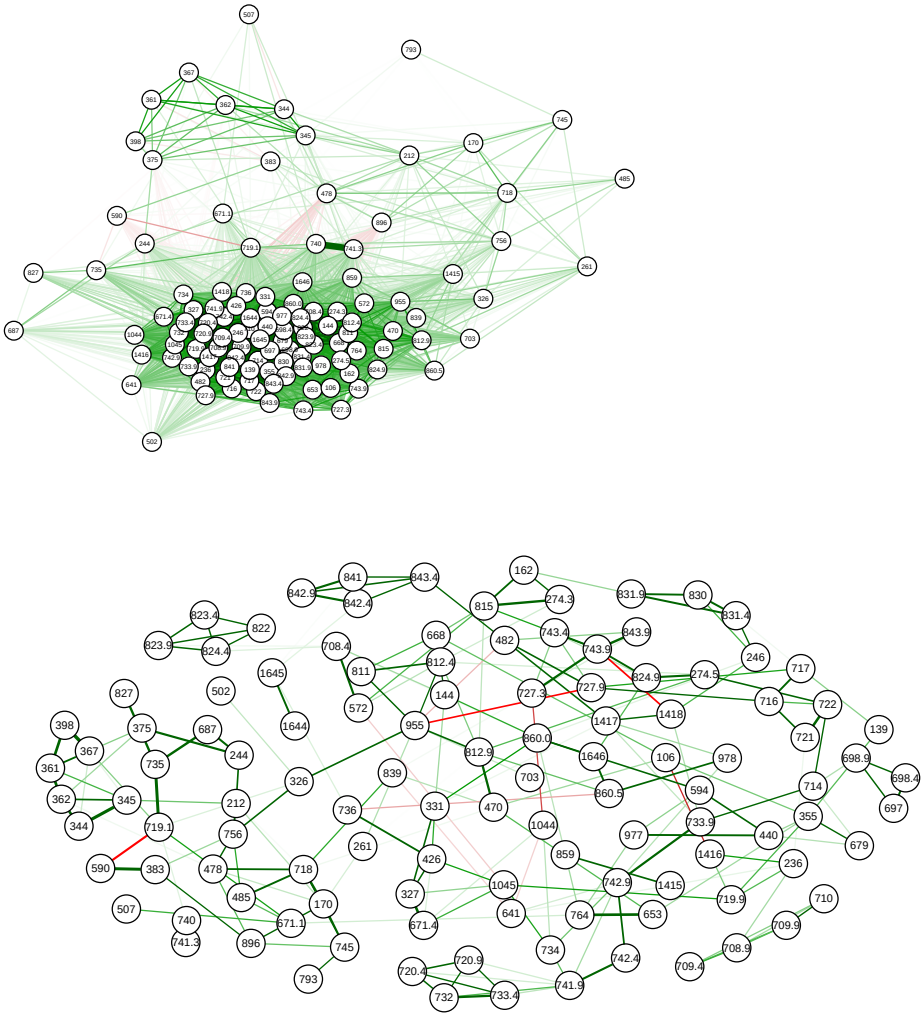


Abbildung C.2.: Pearson-Korrelationen (oben) und GGM-Netzwerk (unten) der Massensignale in der Hitliste 1. Positive Korrelation sind grün, negative Korrelation rot dargestellt.

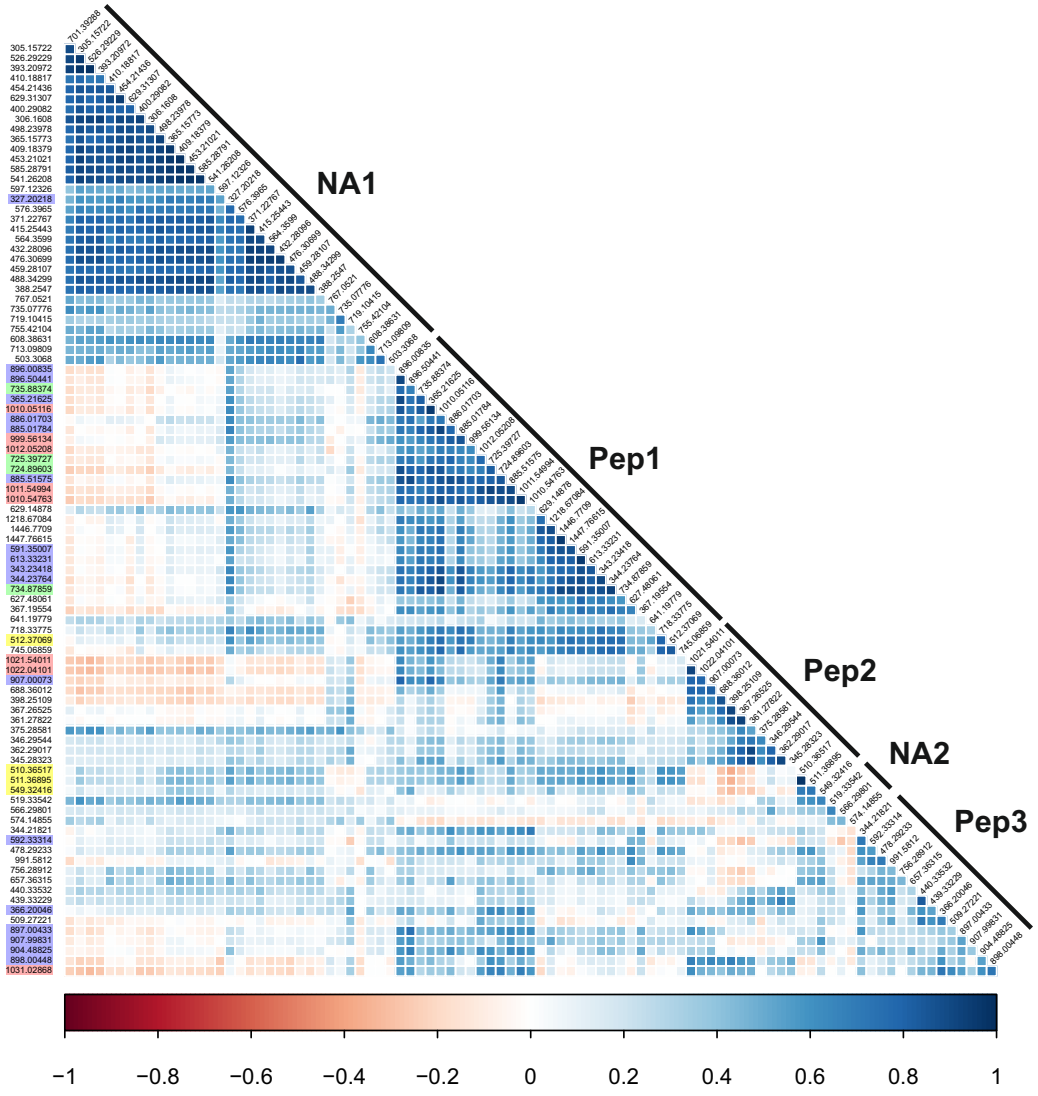


Abbildung C.3.: Pearson-Korrelationen zwischen den Peaks in Hitliste 2.

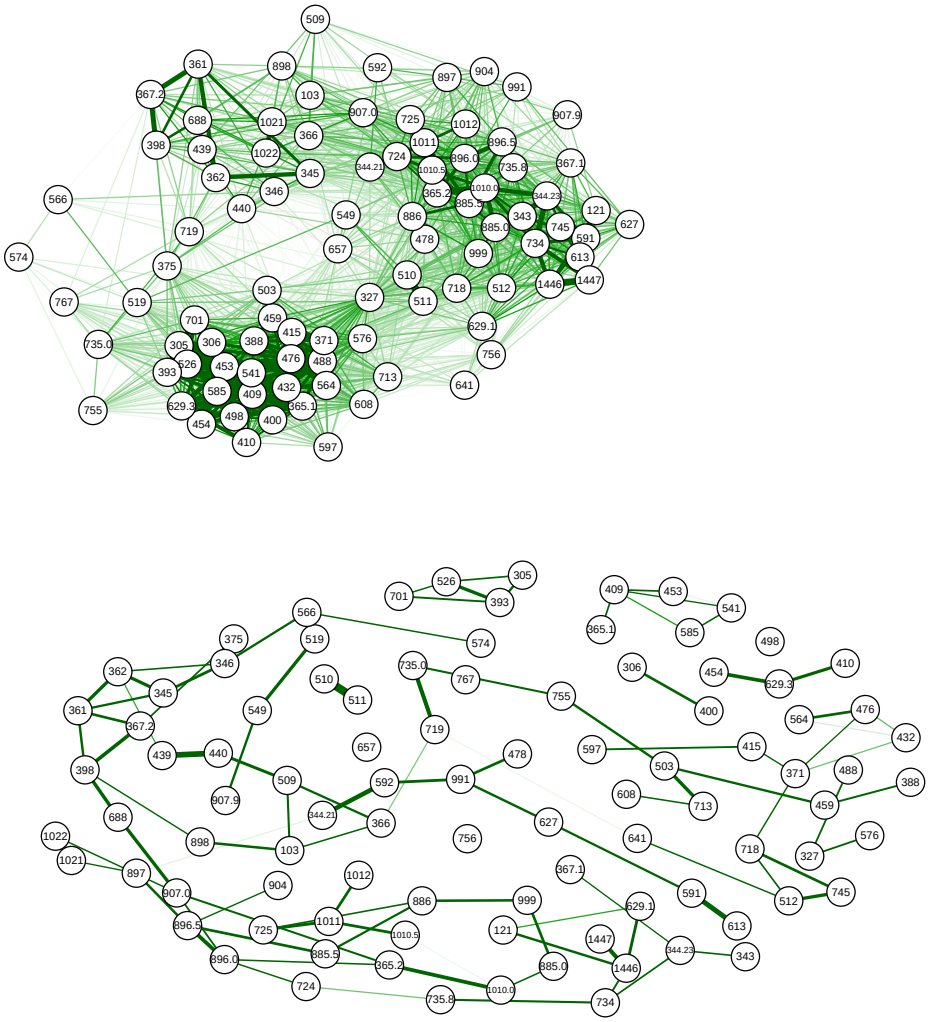
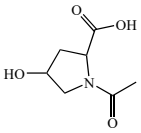
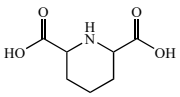


Abbildung C.4.: Pearson-Korrelationen (oben) und GGM-Netzwerk (unten) der Massensignale in Hitliste 2.

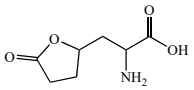
C.5. Strukturvorschläge N-Terminus der Verbindungen 63, 64, 69, 70.



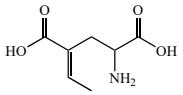
Log P: -1.74
pKa: 13.915, 3.849



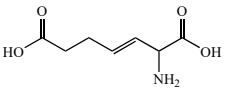
Log P: -0.42
pKa: 9.794, -0.035, -0.035



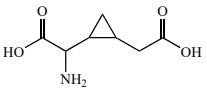
Log P: -1.29
pKa: 2.121, 9.520



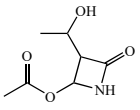
Log P: -0.65
pKa: 2.008, 10.066, 2.765



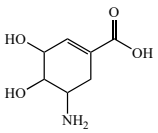
Log P: -0.74
pKa: 1.741, 4.148, 9.728



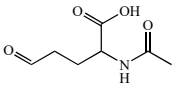
Log P: -1.16
pKa: 3.523, 2.279, 10.081



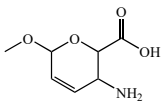
Log P: -0.88
pKa: 13.276



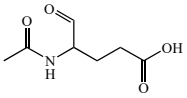
Log P: -1.95
pKa: 13.833, 17.619, 8.902, 2.942



Log P: -1.73
pKa: 3.746



Log P: -0.85
pKa: 9.166, 1.394



Log P: -1.73
pKa: 4.425

Abbildung C.5.: Ergebnisse der Strukturvorschläge in der Super Natural II Datenbank [384] mit der für den *N*-Terminus der Verbindungen **63**, **64**, **69**, **70** vorgeschlagenen Summenformel $C_7H_{11}NO_4$.

C.6. Sequenzalignment der identifizierten Peptaibole

Tabelle C.4.: Sequenzalignment der identifizierten Peptaibole

Peptaibol	P.cluster	RT	Aminosäuresequenz
61	1	7,1	AcTrp Ala Aib Aib **Leu** Aib Gln Aib Aib **Aib** --- --- --- Gln **Leu** Aib --- **Gln** LxxOH
63	2.1	5,1	--- Ala Aib Aib **Leu** Aib Gln Aib Aib **Aib** --- --- --- Gln **Leu** Aib --- **Gln** LxxOH
65	2.2	7,0	--- ??? Aib Aib **Leu** Aib Gln Aib Aib **Aib** --- --- --- Gln **Leu** Aib --- **Gln** LxxOH
67	3	6,9	AcTrp Ala Aib Aib **Aib** Aib Gln Aib Aib **Vxx** Ala Aib Aib Gln **Tyr** Aib Pro **Gln** LxxOH
69	4.1	5,2	--- Ala Aib Aib **Aib** Aib Gln Aib Aib **Vxx** Ala Aib Aib Gln **Tyr** Aib Pro **Gln** LxxOH
71	4.2	6,6	--- ??? Aib Aib **Aib** Aib Gln Aib Aib **Vxx** Ala Aib Aib Gln **Tyr** Aib Pro **Gln** LxxOH
62	1	7,2	AcTrp Ala Aib Aib **Leu** Aib Gln Aib Aib **Aib** --- --- --- Gln **Leu** Aib --- **Glu** LxxOH
64	2.1	5,2	--- Ala Aib Aib **Leu** Aib Gln Aib Aib **Aib** --- --- --- Gln **Leu** Aib --- **Glu** LxxOH
66	2.2	7,1	--- ??? Aib Aib **Leu** Aib Gln Aib Aib **Aib** --- --- --- Gln **Leu** Aib --- **Glu** LxxOH
68	3	7,0	AcTrp Ala Aib Aib **Aib** Aib Gln Aib Aib **Vxx** Ala Aib Aib Gln **Tyr** Aib Pro **Glu** LxxOH
70	4.1	5,3	--- Ala Aib Aib **Aib** Aib Gln Aib Aib **Vxx** Ala Aib Aib Gln **Tyr** Aib Pro **Glu** LxxOH
72	4.2	6,7	--- ??? Aib Aib **Aib** Aib Gln Aib Aib **Vxx** Ala Aib Aib Gln **Tyr** Aib Pro **Glu** LxxOH

D. Multivariate Methoden

D.1. Ergebnisse PCA

Tabelle D.1.: Top 100 Peaks PCA, PC1

Rang	m/z	Regressionskoeffizient	Annotation
1	734.46701	-0.897931437	$[EryA + H]^+$
2	735.46954	-0.368872157	$[EryA + H]^+$, 1. Isotop
3	576.37454	-0.127412575	$[EryA - Cladinose + H]^+$
4	861.3678	-0.102122179	$[Rif + K]^+$
5	736.47474	-0.090325325	$[EryA + H]^+$, 2. Isotop
6	244.84431	-0.060856769	NA
7	772.42468	-0.055528398	$[EryA + K]^+$
8	862.37169	-0.047247261	$[Rif + K]^+$, 1. Isotop
9	577.37782	-0.039116867	$[EryA - Cladinose + H]^+$, 1. Isotop
10	333.21894	-0.030883133	NA
11	558.36376	-0.030670097	$[AEryA - Cladinose + H]^+$
12	381.07934	-0.025811066	NA
13	558.29629	-0.025053054	NA
14	823.41151	-0.023991348	$[Rif + H]^+$
15	393.24035	-0.023430621	NA
16	773.42798	-0.022235406	$[EryA + K]^+$, 1. Isotop
17	245.17872	-0.021929211	NA aus Ery
18	395.25581	-0.020265077	NA
19	720.45212	-0.018276262	$[EryC / NdeMeEryA + H]^+$
20	863.37055	-0.017957363	$[Rif + K]^+$, 2. Isotop
21	737.4771	-0.017337992	$[EryA + H]^+$, 3. Isotop
22	820.52448	-0.017060967	NA
23	774.56495	-0.015015681	NA
24	659.50245	-0.013490132	NA
25	347.19835	-0.012429518	NA

Tabelle D.1.: Top 100 Peaks PCA, PC1 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
26	824.41474	-0.011498224	$[Rif + H]^+$, 1. Isotop
27	175.1187	-0.010687118	NA
28	655.47068	-0.010680118	NA
29	718.47375	-0.010595463	$[EryB + H]^+$
30	748.44816	-0.010244545	$[EryE + H]^+$
31	335.23465	-0.010205015	NA
32	716.45826	-0.010042543	NA aus Ery
33	772.42041	-0.00926043	NA
34	775.56883	-0.00911566	NA
35	559.36786	-0.009067763	$[AEryA - Cladinose + H]^+$, 1. Isotop
36	657.48669	-0.00895477	NA
37	158.11733	-0.008546985	$[Desosamin + H]^+$ aus Ery
38	821.52833	-0.00818907	NA
39	578.38092	-0.007919878	$[EryA - Cladinose + H]^+$, 2. Isotop
40	845.39464	-0.007867263	$[Rif + Na]^+$
41	376.29765	-0.007648917	NA
42	721.45602	-0.007208968	NA
43	798.56445	-0.007207129	NA
44	791.38599	-0.007107763	$[Rif - MeOH]^+$
45	821.39612	-0.007075627	$[Rif - H_2 + H]^+$
46	317.11474	-0.00703325	NA
47	559.29942	-0.006740311	NA
48	467.29255	-0.00672216	NA
49	774.42938	-0.006114305	NA
50	192.13834	-0.005995437	NA aus Ery
51	317.24514	-0.005967642	NA
52	334.22227	-0.005889122	NA
53	540.3535	-0.005878782	$[Fragment EryA + H]^+$
54	750.46292	-0.005869413	$[EryF / EryAEO + H]^+$
55	522.34299	-0.005671126	$[Fragment EryA + H]^+$
56	396.25982	-0.005417046	NA
57	660.50546	-0.005404894	NA
58	287.15266	-0.005317952	NA
59	394.24356	-0.005284889	NA
60	864.37163	-0.005194001	$[Rif + K]^+$, 3. Isotop
61	859.35289	-0.004559041	$[Rif - H_2 + K]^+$

Tabelle D.1.: Top 100 Peaks PCA, PC1 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
62	635.5016	-0.004325554	NA
63	656.47312	-0.004267933	NA
64	636.50365	-0.00412252	NA
65	799.56797	-0.004034427	NA
66	846.39696	-0.003730337	$[Rif + Na]^+$, 1. Isotop
67	822.4007	-0.003725617	$[Rif - H_2 + H]^+$, 1. Isotop
68	658.48969	-0.003587197	NA
69	424.28254	-0.003501833	NA
70	822.53102	-0.003346756	NA
71	397.27138	-0.003332867	NA
72	382.08296	-0.003165043	NA
73	825.41758	-0.003109944	$[Rif + H]^+$, 2. Isotop
74	792.38869	-0.002921286	$[Rif - MeOH]^+$, 1. Isotop
75	244.95555	-0.002841971	NA aus Ery
76	776.56848	-0.002642696	NA
77	860.35577	-0.002545218	$[Rif - H_2 + K]^+$, 1. Isotop
78	663.53274	-0.002527406	NA
79	365.10561	-0.002526792	NA
80	233.04201	-0.002458743	NA
81	166.26624	-0.002409621	NA
82	636.50589	-0.002332121	NA
83	607.47056	-0.002331246	NA
84	166.26025	-0.002271395	NA
85	837.56687	-0.002220392	NA
86	381.24046	-0.002107032	NA
87	800.57146	-0.002088058	NA
88	661.50675	-0.002083474	NA
89	336.238	-0.002071174	NA
90	369.24087	-0.002069546	NA
91	746.53376	-0.00197408	NA
92	497.23737	-0.001969468	NA
93	331.22437	-0.001941574	NA
94	429.24024	-0.001805768	NA
95	789.37041	-0.001769609	$[Rif - H_2 - MeOH]^+$
96	560.29266	-0.001670386	NA
97	720.58993	-0.00163745	NA

Tabelle D.1.: Top 100 Peaks PCA, PC1 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
98	365.15754	-0.001547695	NA
99	383.07689	-0.001482557	NA
100	409.18388	-0.001478757	NA

D.2. Ergebnisse PCR

Tabelle D.2.: Top 100 Peaks PCR, PC2

Rang	m/z	Regressionskoeffizient	Annotation
1	861.3678	4.05E-07	$[Rif + K]^+$
2	862.37169	1.85E-07	$[Rif + K]^+$, 1. Isotop
3	734.46701	1.34E-07	$[EryA + H]^+$
4	859.35289	7.79E-08	$[Rif - H_2 + K]^+$
5	863.37055	6.99E-08	$[Rif + K]^+$, 2. Isotop
6	823.41151	6.92E-08	$[Rif + H]^+$
7	862.36733	5.99E-08	NA
8	859.34924	5.84E-08	NA
9	735.46954	5.57E-08	$[EryA + H]^+$, 1. Isotop
10	821.39612	4.10E-08	$[Rif - H_2 + H]^+$
11	860.35577	4.07E-08	$[Rif - H_2 + K]^+$, 1. Isotop
12	381.07934	3.68E-08	NA
13	824.41474	3.20E-08	$[Rif + H]^+$, 1. Isotop
14	221.04188	3.14E-08	NA
15	845.39464	2.43E-08	$[Rif + Na]^+$

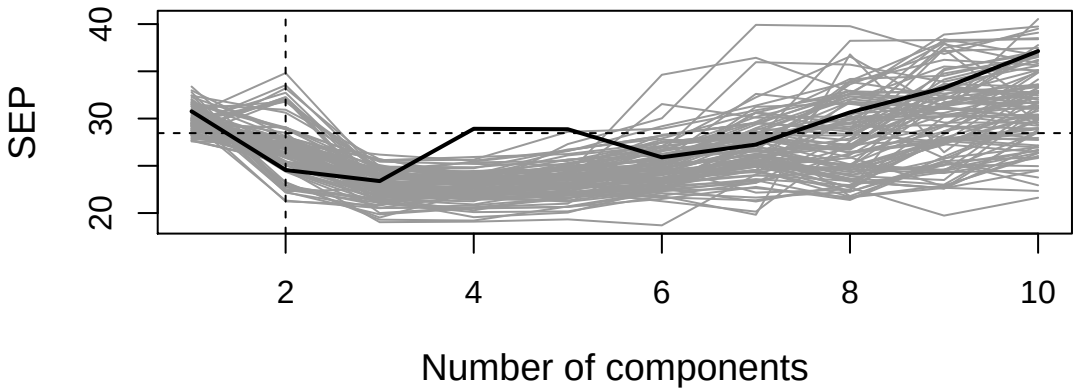


Abbildung D.1.: SEP nach 100-facher doppelter Kreuzvalidierung der PCR.

Tabelle D.2.: Top 100 Peaks PCR, PC2 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
16	822.4007	2.08E-08	[Rif - H_2 + H]$^+$, 1. Isotop
17	864.37163	2.05E-08	[Rif + K]$^+$, 3. Isotop
18	287.15266	1.96E-08	NA
19	791.38599	1.89E-08	[Rif - MeOH]$^+$
20	576.37454	1.68E-08	[EryA - Cladinose + H]$^+$
21	926.50933	1.40E-08	NA aus Rif
22	736.47474	1.33E-08	[EryA + H]$^+$, 2. Isotop
23	351.22963	1.14E-08	NA
24	846.39696	1.09E-08	[Rif + Na]$^+$, 1. Isotop
25	924.495	1.04E-08	NA
26	244.84431	9.26E-09	NA
27	347.19835	9.03E-09	NA
28	349.21377	8.58E-09	NA
29	825.41758	8.56E-09	[Rif + H]$^+$, 2. Isotop
30	792.38869	8.30E-09	[Rif - MeOH]$^+$, 1. Isotop
31	927.51288	6.92E-09	NA aus Rif
32	925.49872	6.12E-09	NA
33	789.37041	5.97E-09	[Rif - H_2 - MeOH]$^+$
34	365.10561	5.85E-09	NA
35	365.20891	5.82E-09	NA
36	376.29765	5.48E-09	NA
37	1043.4458	5.41E-09	NA
38	577.37782	5.19E-09	[EryA - Cladinose + H]$^+$, 1. Isotop
39	287.48765	5.13E-09	NA
40	337.21391	4.98E-09	NA
41	382.08296	4.82E-09	NA
42	558.36376	4.72E-09	[AEryA - Cladinose + H]$^+$
43	351.19341	4.72E-09	NA
44	772.42041	4.29E-09	NA
45	383.2194	4.23E-09	NA
46	397.23509	4.18E-09	NA
47	899.32288	3.99E-09	NA
48	233.04201	3.56E-09	NA
49	723.19559	3.40E-09	NA
50	245.17872	3.33E-09	NA aus Ery
51	335.19834	3.29E-09	NA

Tabelle D.2.: Top 100 Peaks PCR, PC2 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
52	387.14735	3.21E-09	NA
53	352.23306	3.18E-09	NA
54	924.4991	2.99E-09	NA
55	720.45212	2.62E-09	$[EryC / NdeMeEryA + H]^+$
56	734.47015	2.60E-09	NA
57	166.26624	2.50E-09	NA
58	1044.44626	2.50E-09	NA
59	865.37584	2.45E-09	$[Rif + K]^+$, 3. Isotop (?)
60	737.4771	2.44E-09	$[EryA + H]^+$, 3. Isotop
61	1041.43219	2.40E-09	NA
62	877.36099	2.31E-09	NA aus Rif
63	790.37557	2.20E-09	$[Rif - H_2 - MeOH]^+$, 1. Isotop
64	843.37786	2.19E-09	$[Rif - H_2 + Na]^+$
65	348.2016	2.04E-09	NA
66	900.32747	1.97E-09	NA
67	383.07689	1.87E-09	NA
68	724.20159	1.86E-09	NA
69	636.50365	1.86E-09	NA
70	166.25316	1.85E-09	NA
71	350.2173	1.79E-09	NA
72	716.45826	1.70E-09	NA aus Ery
73	331.22437	1.69E-09	NA
74	748.44816	1.67E-09	$[EryE + H]^+$
75	928.51697	1.66E-09	NA
76	718.47375	1.66E-09	$[EryB + H]^+$
77	547.18324	1.60E-09	NA
78	286.48032	1.59E-09	NA
79	317.24514	1.56E-09	NA
80	865.37021	1.53E-09	NA
81	1041.43823	1.44E-09	NA
82	244.95555	1.39E-09	NA aus Ery
83	847.39836	1.38E-09	$[Rif + Na]^+$, 2. Isotop
84	556.58174	1.37E-09	NA
85	633.45005	1.36E-09	NA
86	195.79219	1.33E-09	NA
87	338.21728	1.32E-09	NA

Tabelle D.2.: Top 100 Peaks PCR, PC2 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
88	604.38364	1.31E-09	NA
89	223.04017	1.30E-09	NA
90	847.40308	1.29E-09	$[Rif + Na]^+$, 2. Isotop
91	891.37364	1.27E-09	NA
92	559.36786	1.26E-09	$[AEryA - Cladinose + H]^+$, 1. Isotop
93	1157.60229	1.23E-09	NA
94	578.38092	1.16E-09	$[EryA - Cladinose + H]^+$, 2. Isotop
95	631.43336	1.13E-09	NA
96	721.45602	1.10E-09	NA
97	404.06771	1.10E-09	NA
98	411.25086	1.10E-09	NA
99	584.61285	1.07E-09	NA
100	158.11733	1.07E-09	$[Desosamin + H]^+$ aus Ery

D.3. Ergebnisse PLSR

D.3.1. PLSR Beta

Tabelle D.3.: Top 100 Peaks PLSR Beta, LV1

Rang	m/z	Regressionskoeffizient	Annotation
1	734.46701	2.07E-07	$[EryA + H]^+$
2	861.3678	1.45E-07	$[Rif + K]^+$
3	735.46954	8.52E-08	$[EryA + H]^+$, 1. Isotop
4	862.37169	6.59E-08	$[Rif + K]^+$, 1. Isotop
5	859.35289	3.86E-08	$[Rif - H_2 + K]^+$
6	576.37454	2.94E-08	$[EryA - Cladinose + H]^+$
7	823.41151	2.51E-08	$[Rif + H]^+$
8	863.37055	2.49E-08	$[Rif + K]^+$, 2. Isotop
9	736.47474	2.06E-08	$[EryA + H]^+$, 2. Isotop
10	860.35577	1.96E-08	$[Rif - H_2 + K]^+$, 1. Isotop
11	862.36733	1.82E-08	NA
12	821.39612	1.78E-08	$[Rif - H_2 + H]^+$
13	859.34924	1.68E-08	NA
14	772.42468	1.42E-08	$[EryA + K]^+$
15	244.84431	1.40E-08	NA
16	824.41474	1.15E-08	$[Rif + H]^+$, 1. Isotop
17	577.37782	9.02E-09	$[EryA - Cladinose + H]^+$, 1. Isotop
18	822.4007	8.92E-09	$[Rif - H_2 + H]^+$, 1. Isotop
19	845.39464	8.50E-09	$[Rif + Na]^+$
20	864.37163	7.29E-09	$[Rif + K]^+$, 3. Isotop
21	393.24035	7.18E-09	NA
22	558.36376	7.04E-09	$[AEryA - Cladinose + H]^+$
23	791.38599	6.70E-09	$[Rif - MeOH]^+$
24	287.15266	6.60E-09	NA
25	395.25581	6.58E-09	NA
26	558.29629	6.17E-09	NA
27	820.52448	6.13E-09	NA
28	774.56495	5.75E-09	NA
29	773.42798	5.72E-09	$[EryA + K]^+$, 1. Isotop
30	659.50245	5.69E-09	NA
31	175.1187	5.49E-09	NA

Tabelle D.3.: Top 100 Peaks PLSR Beta, LV1 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
32	245.17872	5.04E-09	NA aus Ery
33	926.50933	4.96E-09	NA aus Rif
34	924.495	4.94E-09	NA
35	775.56883	4.44E-09	NA
36	655.47068	4.40E-09	NA
37	720.45212	4.21E-09	$[EryC / NdeMeEryA + H]^+$
38	737.4771	3.92E-09	$[EryA + H]^+$, 3. Isotop
39	846.39696	3.84E-09	$[Rif + Na]^+$, 1. Isotop
40	657.48669	3.81E-09	NA
41	924.4991	3.48E-09	NA
42	925.49872	3.18E-09	NA
43	825.41758	3.05E-09	$[Rif + H]^+$, 2. Isotop
44	792.38869	2.98E-09	$[Rif - MeOH]^+$, 1. Isotop
45	798.56445	2.69E-09	NA
46	821.52833	2.68E-09	NA
47	789.37041	2.48E-09	$[Rif - H_2 - MeOH]^+$
48	221.04188	2.45E-09	NA
49	718.47375	2.43E-09	$[EryB + H]^+$
50	660.50546	2.37E-09	NA
51	748.44816	2.34E-09	$[EryE + H]^+$
52	927.51288	2.34E-09	NA aus Rif
53	396.25982	2.26E-09	NA
54	716.45826	2.23E-09	NA aus Ery
55	559.36786	2.09E-09	$[AEryA - Cladinose + H]^+$, 1. Isotop
56	376.29765	2.08E-09	NA
57	158.11733	1.98E-09	$[Desosamin + H]^+$ aus Ery
58	1043.4458	1.94E-09	NA
59	467.29255	1.89E-09	NA
60	578.38092	1.82E-09	$[EryA - Cladinose + H]^+$, 2. Isotop
61	772.42041	1.81E-09	NA
62	656.47312	1.80E-09	NA
63	559.29942	1.77E-09	NA
64	287.48765	1.67E-09	NA
65	721.45602	1.66E-09	NA
66	774.42938	1.65E-09	NA
67	658.48969	1.60E-09	NA

Tabelle D.3.: Top 100 Peaks PLSR Beta, LV1 (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
68	394.24356	1.59E-09	NA
69	899.32288	1.45E-09	NA
70	192.13834	1.39E-09	NA aus Ery
71	750.46292	1.36E-09	$[EryF / EryAEO + H]^+$
72	540.3535	1.36E-09	$[Fragment EryA + H]^+$
73	799.56797	1.33E-09	NA
74	522.34299	1.31E-09	$[Fragment EryA + H]^+$
75	429.24024	1.28E-09	NA
76	720.58993	1.24E-09	NA
77	822.53102	1.03E-09	NA
78	800.57146	1.01E-09	NA
79	865.37584	9.39E-10	$[Rif + K]^+$, 3. Isotop (?)
80	776.56848	9.33E-10	NA
81	166.26624	9.21E-10	NA
82	661.50675	9.17E-10	NA
83	286.48032	9.11E-10	NA
84	746.53376	8.85E-10	NA
85	1044.44626	8.74E-10	NA
86	365.20891	8.74E-10	NA
87	790.37557	8.69E-10	$[Rif - H_2 - MeOH]^+$, 1. Isotop
88	381.24046	8.48E-10	NA
89	843.37786	8.34E-10	$[Rif - H_2 + Na]^+$
90	1041.43219	8.34E-10	NA
91	734.47015	8.25E-10	NA
92	636.50365	7.99E-10	NA
93	877.36099	7.99E-10	NA aus Rif
94	424.28254	7.54E-10	NA
95	661.50306	7.21E-10	NA
96	891.37364	7.09E-10	NA
97	195.79219	6.79E-10	NA
98	365.15754	6.63E-10	NA
99	900.32747	6.35E-10	NA
100	865.37021	6.19E-10	NA

D.3.2. PLSR VIP

Tabelle D.4.: Top 100 Peaks PLSR VIP, LV1

Rang	m/z	VIP	Annotation
1	734.46701	26.75184602	$[EryA + H]^+$
2	861.3678	18.69285867	$[Rif + K]^+$
3	735.46954	10.98904809	$[EryA + H]^+$, 1. Isotop
4	862.37169	8.491809477	$[Rif + K]^+$, 1. Isotop
5	859.35289	4.978684115	$[Rif - H_2 + K]^+$
6	576.37454	3.793001988	$[EryA - Cladinose + H]^+$
7	823.41151	3.240039991	$[Rif + H]^+$
8	863.37055	3.216744214	$[Rif + K]^+$, 2. Isotop
9	736.47474	2.658202569	$[EryA + H]^+$, 2. Isotop
10	860.35577	2.528375284	$[Rif - H_2 + K]^+$, 1. Isotop
11	862.36733	2.344274806	NA
12	821.39612	2.291670578	$[Rif - H_2 + H]^+$
13	859.34924	2.164827035	NA
14	772.42468	1.834338363	$[EryA + K]^+$
15	536.33585	1.815123216	NA
16	244.84431	1.802313342	NA
17	335.23465	1.786292614	NA
18	824.41474	1.482383083	$[Rif + H]^+$, 1. Isotop
19	577.37782	1.162785715	$[EryA - Cladinose + H]^+$, 1. Isotop
20	822.4007	1.150658878	$[Rif - H_2 + H]^+$, 1. Isotop
21	296.06584	1.111509757	NA
22	845.39464	1.095561412	$[Rif + Na]^+$
23	512.33515	1.051575297	NA
24	299.28063	0.976215789	NA
25	864.37163	0.940143064	$[Rif + K]^+$, 3. Isotop
26	393.24035	0.925263168	NA
27	558.36376	0.907663149	$[AEryA - Cladinose + H]^+$
28	274.10501	0.87785133	NA
29	791.38599	0.864241474	$[Rif - MeOH]^+$
30	287.15266	0.851522951	NA
31	347.19835	0.849631052	NA
32	395.25581	0.849015631	NA
33	367.22465	0.803926568	NA

Tabelle D.4.: Top 100 Peaks PLSR VIP, LV1 (Fortsetzung)

Rang	m/z	VIP	Annotation
34	558.29629	0.795061262	NA
35	820.52448	0.789987542	NA
36	381.07934	0.76515639	NA
37	774.56495	0.741931082	NA
38	773.42798	0.736944486	$[EryA + K]^+$, 1. Isotop
39	659.50245	0.733334122	NA
40	175.1187	0.707415353	NA
41	395.37432	0.683492552	NA
42	245.17872	0.649495181	NA aus Ery
43	926.50933	0.639348483	NA aus Rif
44	924.495	0.637569252	NA
45	723.19559	0.622224697	NA
46	775.56883	0.572257361	NA
47	655.47068	0.567155382	NA
48	720.45212	0.543001263	$[EryC / NdeMeEryA + H]^+$
49	537.33887	0.539665626	NA
50	737.4771	0.505256023	$[EryA + H]^+$, 3. Isotop
51	846.39696	0.495054585	$[Rif + Na]^+$, 1. Isotop
52	657.48669	0.491216121	NA
53	333.21894	0.454487817	NA
54	775.56538	0.45175591	NA
55	333.18264	0.449916958	NA
56	924.4991	0.448241255	NA
57	925.49872	0.410304445	NA
58	825.41758	0.393793709	$[Rif + H]^+$, 2. Isotop
59	272.97716	0.390593352	NA
60	792.38869	0.383637336	$[Rif - MeOH]^+$, 1. Isotop
61	369.20385	0.367667388	NA
62	349.21377	0.358613564	NA
63	336.238	0.354033357	NA
64	249.03692	0.35392276	NA
65	798.56445	0.347204522	NA
66	821.52833	0.345717374	NA
67	369.23998	0.345450401	NA
68	498.37954	0.344515869	NA
69	513.3383	0.329424516	NA

Tabelle D.4.: Top 100 Peaks PLSR VIP, LV1 (Fortsetzung)

Rang	m/z	VIP	Annotation
70	789.37041	0.320216555	[Rif - H_2 - MeOH]$^+$
71	335.19834	0.319150935	NA
72	511.35476	0.318060846	NA
73	383.2194	0.317919283	NA
74	221.04188	0.315971939	NA
75	718.47375	0.313255386	[EryB + H]$^+$
76	353.20898	0.309384392	NA
77	660.50546	0.305049676	NA
78	748.44816	0.302207377	[EryE + H]$^+$
79	927.51288	0.301572199	NA aus Rif
80	319.20341	0.299090685	NA
81	615.09295	0.295499373	NA
82	396.25982	0.291435951	NA
83	716.45826	0.287802442	NA aus Ery
84	325.2962	0.28762359	NA
85	445.25026	0.28388168	NA
86	368.35233	0.276806251	NA
87	559.36786	0.269912969	[AEryA - Cladinose + H]$^+$, 1. Isotop
88	376.29765	0.268101476	NA
89	236.14903	0.26639092	NA
90	363.19337	0.262629262	NA
91	158.11733	0.25565789	[Desosamin + H]$^+$ aus Ery
92	1043.4458	0.25009006	NA
93	467.29255	0.243164562	NA
94	578.38092	0.235059784	[EryA - Cladinose + H]$^+$, 2. Isotop
95	772.42041	0.233382301	NA
96	656.47312	0.23252067	NA
97	308.22216	0.231068807	NA
98	559.29942	0.227822478	NA
99	629.4544	0.225705164	NA
100	482.46837	0.223225228	NA

D.4. Ergebnisse QPAR

Tabelle D.5.: Top 100 Peaks QPAR Selectivity Ratios

Rang	m/z	Selectivity Ratio	Annotation
1	846.39696	6.21885	$[Rif + Na]^+$, 1. Isotop
2	845.39464	5.86674	$[Rif + Na]^+$
3	863.37055	5.35762	$[Rif + K]^+$, 2. Isotop
4	861.3678	5.35326	$[Rif + K]^+$
5	862.37169	5.33481	$[Rif + K]^+$, 1. Isotop
6	864.37163	5.13949	$[Rif + K]^+$, 3. Isotop
7	287.15266	4.33580	NA
8	823.41151	3.11487	$[Rif + H]^+$
9	824.41474	2.97452	$[Rif + H]^+$, 1. Isotop
10	791.38599	2.96235	$[Rif - MeOH]^+$
11	789.37041	2.86008	$[Rif - H_2 - MeOH]^+$
12	825.41758	2.65770	$[Rif + H]^+$, 2. Isotop
13	792.38869	2.62535	$[Rif - MeOH]^+$, 1. Isotop
14	822.4007	1.53970	$[Rif - H_2 + H]^+$, 1. Isotop
15	821.39612	1.43929	$[Rif - H_2 + H]^+$
16	877.36099	0.97021	NA aus Rif
17	899.32288	0.96402	NA
18	287.48765	0.91530	NA
19	900.32747	0.76886	NA
20	1041.43219	0.76517	NA
21	1043.4458	0.74376	NA
22	926.50933	0.62463	NA aus Rif
23	859.34924	0.60056	NA
24	843.37786	0.60041	$[Rif - H_2 + Na]^+$
25	790.37557	0.59260	$[Rif - H_2 - MeOH]^+$, 1. Isotop
26	927.51288	0.56661	NA aus Rif
27	1044.44626	0.54855	NA
28	860.35577	0.54133	$[Rif - H_2 + K]^+$, 1. Isotop
29	748.44816	0.46615	$[EryE + H]^+$
30	244.95555	0.46576	NA aus Ery
31	718.47375	0.45827	$[EryB + H]^+$
32	859.35289	0.45617	$[Rif - H_2 + K]^+$
33	558.36376	0.45469	$[AEryA - Cladinose + H]^+$

Tabelle D.5.: Top 100 Peaks QPAR, Selectivity Ratios (Fortsetzung)

Rang	m/z	Selectivity Ratio	Annotation
34	721.45602	0.45262	NA
35	735.46954	0.45224	$[EryA + H]^+$, 1. Isotop
36	244.84431	0.45207	NA
37	245.17872	0.45180	NA aus Ery
38	716.45826	0.45043	NA aus Ery
39	734.46701	0.44976	$[EryA + H]^+$
40	772.42041	0.44581	NA
41	578.38092	0.44406	$[EryA - Cladinose + H]^+$, 2. Isotop
42	736.47474	0.44344	$[EryA + H]^+$, 2. Isotop
43	720.45212	0.44013	$[EryC / NdeMeEryA + H]^+$
44	540.3535	0.43517	$[Fragment EryA + H]^+$
45	865.37584	0.43317	$[Rif + K]^+$, 3. Isotop (?)
46	559.36786	0.43301	$[AEryA - Cladinose + H]^+$, 1. Isotop
47	522.34299	0.42786	$[Fragment EryA + H]^+$
48	192.13834	0.42763	NA aus Ery
49	577.37782	0.42713	$[EryA - Cladinose + H]^+$, 1. Isotop
50	576.37454	0.42413	$[EryA - Cladinose + H]^+$
51	737.4771	0.42126	$[EryA + H]^+$, 3. Isotop
52	158.11733	0.41385	$[Desosamin + H]^+$ aus Ery
53	750.46292	0.40795	$[EryF / EryAEO + H]^+$
54	847.39836	0.34945	$[Rif + Na]^+$, 2. Isotop
55	1171.62964	0.31360	NA
56	166.25316	0.28432	NA
57	928.51697	0.28432	NA
58	1157.60229	0.28432	NA
59	166.26624	0.28063	NA
60	925.49872	0.26765	NA
61	924.495	0.26703	NA
62	376.29765	0.25969	NA
63	865.37021	0.23961	NA
64	1041.43823	0.23874	NA
65	274.49867	0.23662	NA
66	847.40308	0.23595	$[Rif + Na]^+$, 2. Isotop
67	195.79219	0.20745	NA
68	1045.45418	0.18797	NA
69	793.39091	0.18320	NA

Tabelle D.5.: Top 100 Peaks QPAR, Selectivity Ratios (Fortsetzung)

Rang	m/z	Selectivity Ratio	Annotation
70	1203.49469	0.18320	NA
71	862.36733	0.18203	NA
72	891.37364	0.16754	NA
73	286.48032	0.16014	NA
74	604.38364	0.09337	NA
75	772.42468	0.08824	$[EryA + K]^+$
76	773.42798	0.08794	$[EryA + K]^+$, 1. Isotop
77	381.24046	0.07965	NA
78	397.23509	0.06488	NA
79	387.14735	0.05538	NA
80	924.4991	0.04150	NA
81	734.47015	0.03392	NA
82	273.82648	0.03217	NA
83	166.25176	0.03217	NA
84	303.72194	0.03217	NA
85	589.40808	0.03217	NA
86	1217.65763	0.03217	NA
87	321.13135	0.03217	NA
88	499.21915	0.03217	NA
89	677.45898	0.03217	NA
90	892.38062	0.03217	NA
91	897.30643	0.03217	NA
92	901.3811	0.03217	NA
93	1169.60187	0.03217	NA
94	1214.6309	0.03217	NA
95	1216.64105	0.03217	NA
96	1218.659	0.03217	NA
97	286.81471	0.03217	NA
98	763.56094	0.03217	NA
99	997.51031	0.03217	NA
100	1215.64779	0.03217	NA

D.5. Ergebnisse Random Forest

Tabelle D.6.: Hitliste Random Forest, *mtry*=139

Rang	m/z	Variable Importance	Annotation
1	483.28759	32.2527	NA
2	861.3678	32.0552	$[Rif + K]^+$
3	822.4007	31.6151	$[Rif - H_2 + H]^+$, 1. Isotop
4	862.37169	31.3341	$[Rif + K]^+$, 1. Isotop
5	821.39612	31.3139	$[Rif - H_2 + H]^+$
6	863.37055	31.2059	$[Rif + K]^+$, 2. Isotop
7	864.37163	30.8794	$[Rif + K]^+$, 3. Isotop
8	846.39696	29.9713	$[Rif + Na]^+$, 1. Isotop
9	789.37041	29.1836	$[Rif - H_2 - MeOH]^+$
10	860.35577	27.3337	$[Rif - H_2 + K]^+$, 1. Isotop
11	859.35289	26.2577	$[Rif - H_2 + K]^+$
12	824.41474	25.6149	$[Rif + H]^+$, 1. Isotop
13	791.38599	25.5408	$[Rif - MeOH]^+$
14	823.41151	25.4552	$[Rif + H]^+$
15	825.41758	25.2991	$[Rif + H]^+$, 2. Isotop
16	792.38869	25.2840	$[Rif - MeOH]^+$, 1. Isotop
17	845.39464	25.2182	$[Rif + Na]^+$
18	287.15266	24.7260	NA
19	319.11288	18.0924	NA
20	354.21225	14.1465	NA
21	645.44894	12.1091	NA
22	647.46518	12.0673	NA
23	485.30326	11.3325	NA
24	899.32288	8.5190	NA
25	210.9403	7.9818	NA
26	877.36099	7.9501	NA aus Rif
27	926.50933	7.5364	NA aus Rif
28	927.51288	7.3156	NA aus Rif
29	843.68403	7.3151	NA
30	1043.4458	6.5977	NA
31	481.27179	6.5154	NA
32	287.48765	6.4669	NA
33	465.27684	5.8182	NA

Tabelle D.6.: Hitliste Random Forest, *mtry*=139 (Fortsetzung)

Rang	m/z	Variable Importance	Annotation
34	831.68245	5.5204	NA
35	323.19822	5.5050	NA
36	236.14903	4.6783	NA
37	368.42505	3.3085	NA
38	351.19341	3.1322	NA
39	1044.44626	2.9123	NA
40	351.21361	2.8614	NA
41	248.89629	2.7069	NA
42	900.32747	2.5909	NA
43	611.44394	2.5667	NA
44	646.45203	2.4737	NA
45	924.495	2.4189	NA
46	547.36084	2.1981	NA
47	337.21391	2.1327	NA
48	349.19849	2.1097	NA
49	865.37584	1.8298	$[Rif + K]^+$, 3. Isotop (?)
50	925.49872	1.8197	NA
51	1041.43219	1.7567	NA
52	869.70115	1.7346	NA
53	441.26066	1.7335	NA
54	632.55772	1.6492	NA
55	734.46701	1.6113	$[EryA + H]^+$
56	187.03641	1.6077	NA
57	859.34924	1.5795	NA
58	576.37454	1.4853	$[EryA - Cladinose + H]^+$
59	737.4771	1.4500	$[EryA + H]^+$, 3. Isotop
60	790.37557	1.4039	$[Rif - H_2 - MeOH]^+$, 1. Isotop
61	843.37786	1.3730	$[Rif - H_2 + Na]^+$
62	736.47474	1.3727	$[EryA + H]^+$, 2. Isotop
63	748.44816	1.3708	$[EryE + H]^+$
64	158.11733	1.3513	$[Desosamin + H]^+$ aus Ery
65	499.28238	1.3475	NA
66	559.36786	1.3260	$[AEryA - Cladinose + H]^+$, 1. Isotop
67	244.84431	1.3072	NA
68	245.17872	1.2937	NA aus Ery
69	192.13834	1.2925	NA aus Ery

Tabelle D.6.: Hitliste Random Forest, *mtry*=139 (Fortsetzung)

Rang	m/z	Variable Importance	Annotation
70	540.3535	1.2855	[Fragment EryA + H]$^+$
71	716.45826	1.2767	NA aus Ery
72	735.46954	1.2703	[EryA + H]$^+$, 1. Isotop
73	469.30879	1.2624	NA
74	721.45602	1.2583	NA
75	718.47375	1.2457	[EryB + H]$^+$
76	817.66847	1.1897	NA
77	396.25982	1.1754	NA
78	857.69785	1.1650	NA
79	577.37782	1.1641	[EryA - Cladinose + H]$^+$, 1. Isotop
80	750.46292	1.1581	[EryF / EryAEO + H]$^+$
81	842.66855	1.1463	NA
82	395.21999	1.1435	NA
83	720.45212	1.1246	[EryC / NdeMeEryA + H]$^+$
84	409.18388	1.0928	NA
85	841.66752	1.0915	NA
86	578.38092	1.0867	[EryA - Cladinose + H]$^+$, 2. Isotop
87	558.36376	1.0548	[AEryA - Cladinose + H]$^+$
88	522.34299	1.0376	[Fragment EryA + H]$^+$
89	605.45433	1.0232	NA
90	845.6961	1.0159	NA
91	465.33423	0.9824	NA
92	258.13093	0.9746	NA
93	301.14107	0.9217	NA
94	486.30685	0.9091	NA
95	453.21054	0.9009	NA
96	627.43626	0.8557	NA
97	429.24024	0.8314	NA
98	876.63312	0.8201	NA
99	511.35476	0.7942	NA
100	357.15925	0.7713	NA
101	366.2122	0.7530	NA
102	603.52322	0.7286	NA
103	629.4544	0.7211	NA
104	830.67064	0.6837	NA
105	274.49867	0.6660	NA

Tabelle D.6.: Hitliste Random Forest, *mtry*=139 (Fortsetzung)

Rang	m/z	Variable Importance	Annotation
106	475.31726	0.6556	NA
107	314.30507	0.6358	NA
108	871.71055	0.6358	NA
109	847.40308	0.6319	$[Rif + Na]^+$, 2. Isotop
110	815.65463	0.6249	NA
111	1045.45418	0.5991	NA
112	591.43854	0.5670	NA
113	895.17781	0.5536	NA
114	477.33547	0.5363	NA
115	604.52595	0.4969	NA
116	241.15441	0.4797	NA
117	631.36279	0.4637	NA
118	175.1187	0.4308	NA
119	224.11267	0.4300	NA
120	564.16118	0.4228	NA
121	503.33477	0.4047	NA
122	576.49469	0.3869	NA
123	383.2194	0.3861	NA
124	638.18592	0.3858	NA
125	643.43275	0.3746	NA
126	337.2328	0.3710	NA
127	702.41606	0.3692	NA
128	855.6866	0.3677	NA
129	470.31206	0.3677	NA
130	613.45844	0.3440	NA
131	847.39836	0.3292	$[Rif + Na]^+$, 2. Isotop
132	631.35992	0.3197	NA
133	845.70572	0.3108	NA
134	801.63965	0.3092	NA
135	419.32848	0.3045	NA
136	604.38364	0.3016	NA
137	501.30056	0.2883	NA
138	320.26385	0.2837	NA
139	803.65393	0.2813	NA
140	1041.43823	0.2807	NA
141	924.4991	0.2794	NA

Tabelle D.6.: Hitliste Random Forest, *mtry*=139 (Fortsetzung)

Rang	m/z	Variable Importance	Annotation
142	627.44032	0.2726	NA
143	536.33585	0.2705	NA
144	409.27177	0.2626	NA
145	554.06346	0.2511	NA
146	430.24398	0.2480	NA
147	665.53378	0.2477	NA
148	334.22227	0.2468	NA
149	933.13443	0.2387	NA
150	384.22288	0.2370	NA
151	575.492	0.2336	NA
152	631.55442	0.2328	NA
153	865.37021	0.2299	NA
154	1171.62964	0.2281	NA
155	293.18183	0.2266	NA
156	638.51084	0.2233	NA
157	553.0573	0.2215	NA

D.6. Ergebnisse Regularisierungsmethoden

D.6.1. Parameteroptimierung

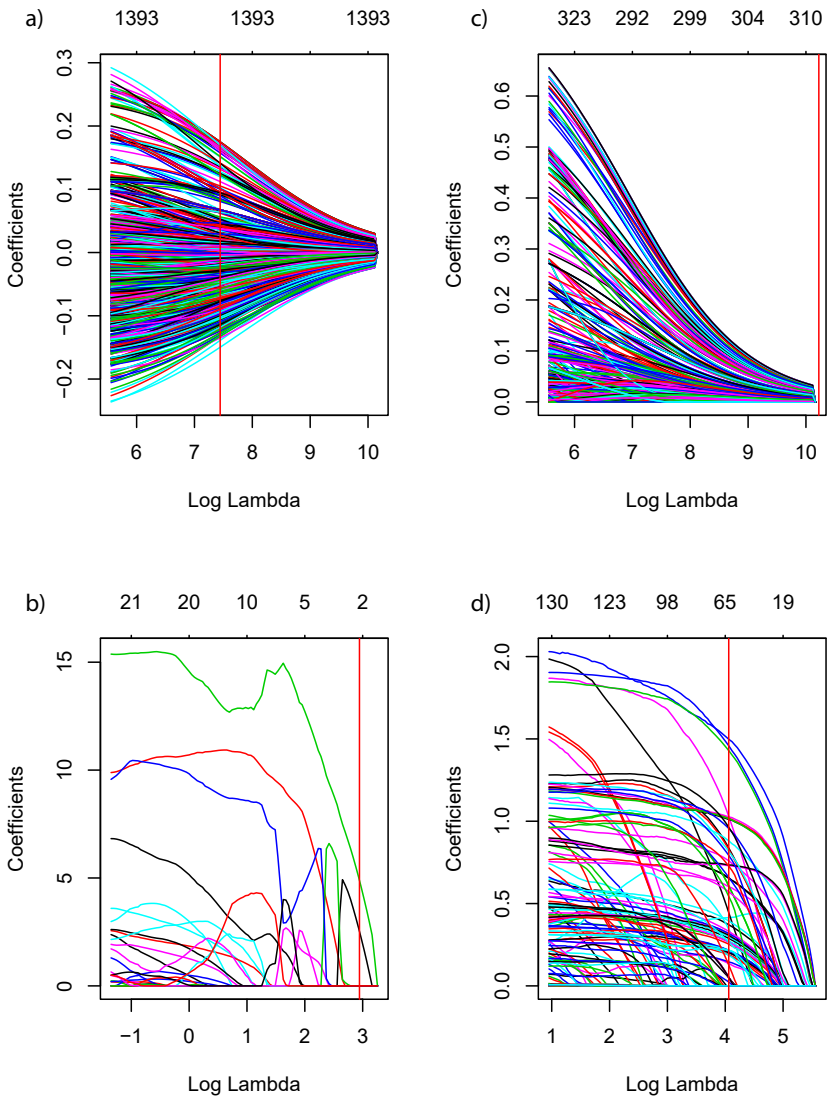


Abbildung D.2.: Regressionskoeffizienten in Abhängigkeit vom Penaliserungsfaktor log λ. Die Anzahl der Variablen im Modell ist jeweils über dem Plot angegeben. a) Ridge Regression mit lower.limits = -Inf, b) Ridge Regression mit lower.limits = 0 c) Lasso mit lower.limits = 0 d) Elastic Net mit lower.limits = 0 für $\alpha = 0.1$

Tabelle D.7.: Ergebnisse Optimierung von α und λ für lower.limit = -Inf.

alpha	lambda.opt	MSEP.lambda.opt	lambda.opt.1SE	MSEP.1SE
0	1710	169.9163286	2994	184.9225901
0.1	104	518.4125708	66	536.172906
0.2	76	796.5909546	94	815.703719
0.3	46	666.5469993	22	684.8994288
0.4	42	556.7922283	45	575.226055
0.5	53	616.8470139	53	635.2259863
0.6	32	710.3384092	21	730.7442634
0.7	25	696.9500653	27	715.8184441
0.8	24	751.9891095	12	770.7889287
0.9	21	671.3081906	21	690.6859527
1	19	744.1955082	20	763.8036063

Tabelle D.8.: Ergebnisse Optimierung von α und λ für lower.limit = 0.

alpha	lambda.opt	MSEP.lambda.opt	lambda.opt.1SE	MSEP.1SE
0	27535	2281.175737	25115	2298.330848
0.1	58	356.1155546	78	375.0324308
0.2	72	547.7111572	81	566.9782107
0.3	51	497.71625	43	516.2434384
0.4	36	550.3250677	23	568.7806654
0.5	30	653.2507155	38	671.6172823
0.6	25	714.0331001	28	734.4934786
0.7	22	521.2368897	7	539.9180559
0.8	23	685.4567726	8	704.503308
0.9	20	600.4673173	22	620.3020947
1	19	808.1143084	9	828.0427641

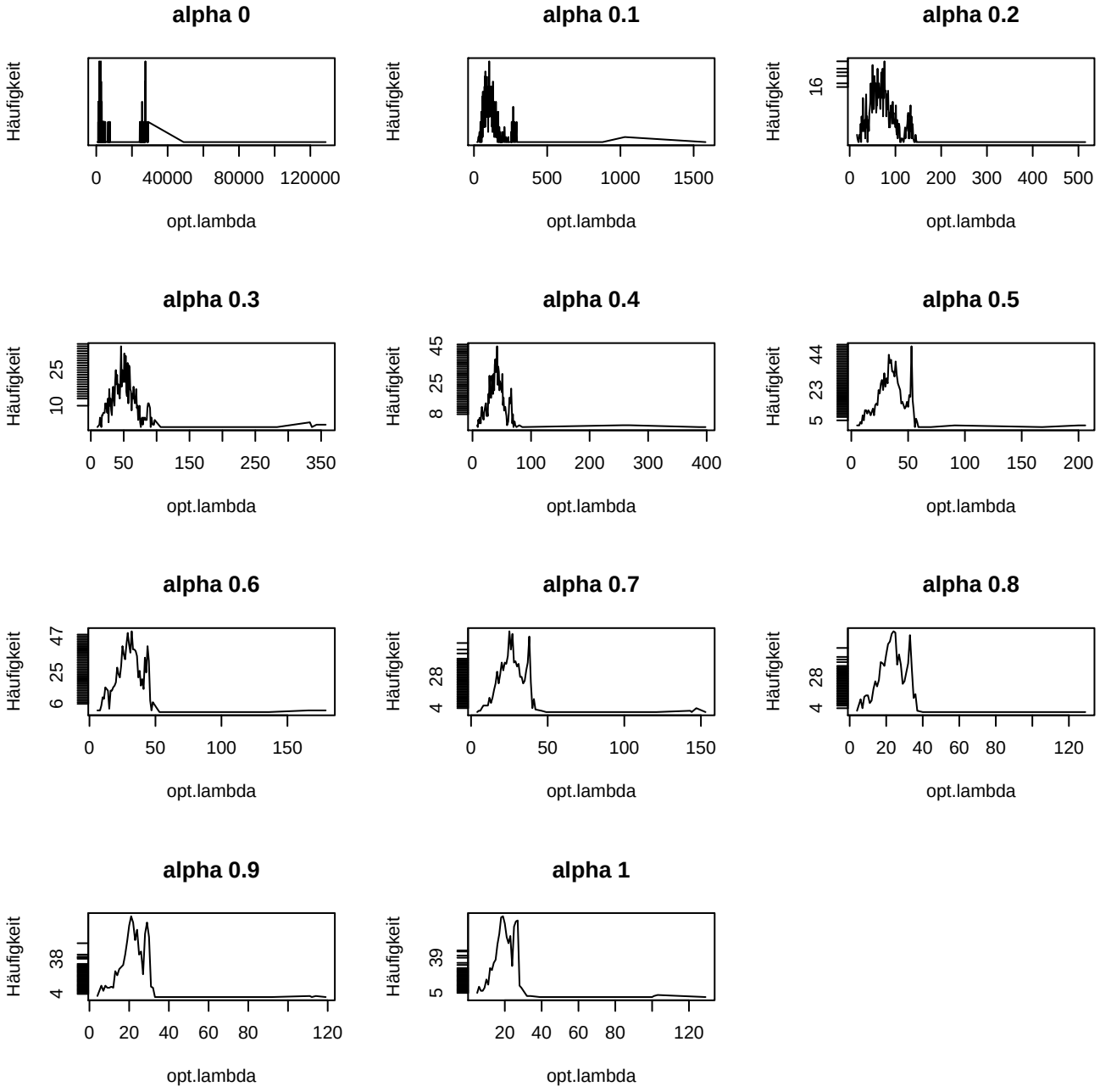


Abbildung D.3.: Häufigkeitsverteilung der optimalen lambda-Werte aus der Kreuzvalidierung mit lower.limits = -Inf.

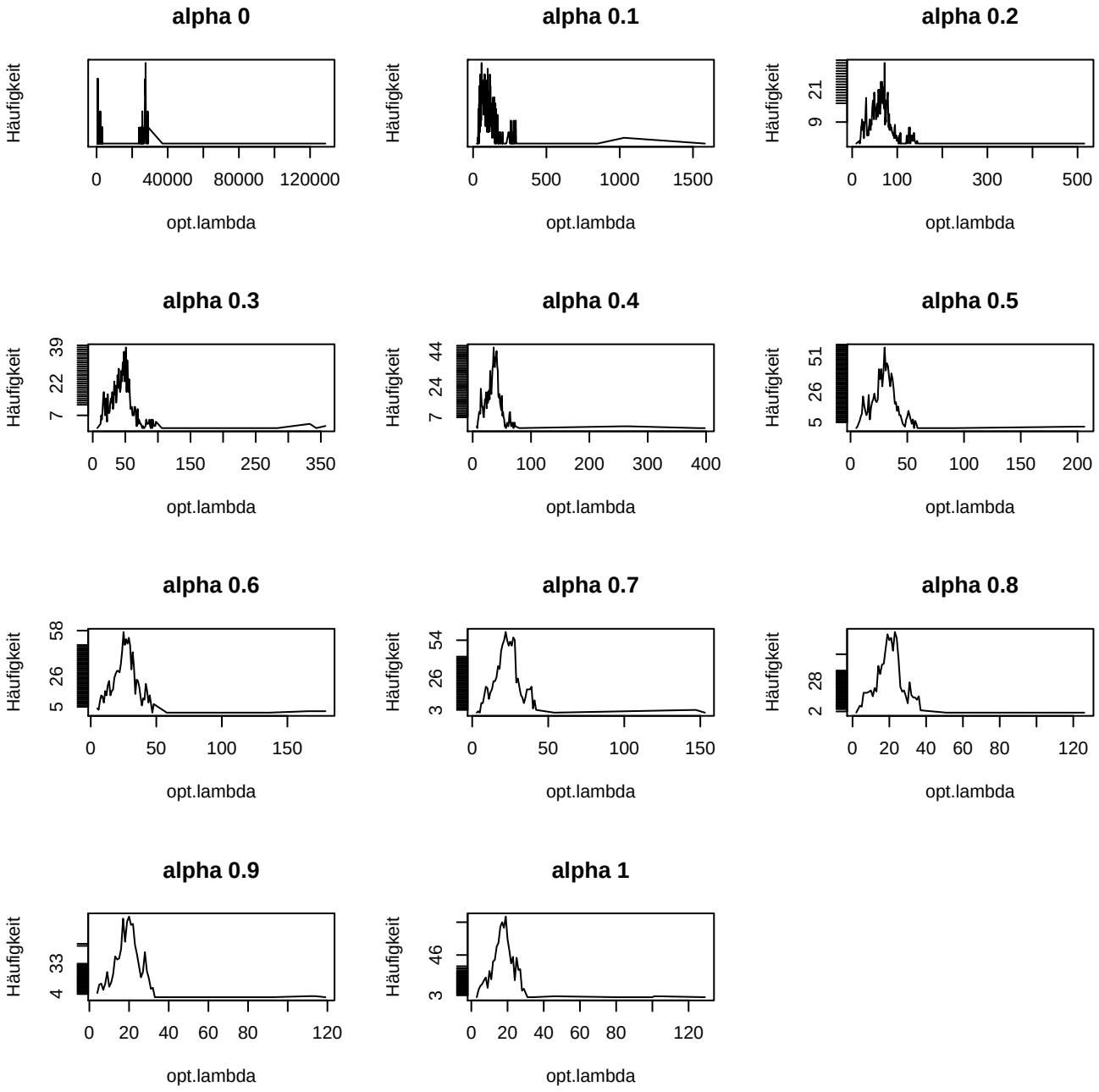


Abbildung D.4.: Häufigkeitsverteilung der optimalen lambda-Werte aus der Kreuzvalidierung mit lower.limits = 0.

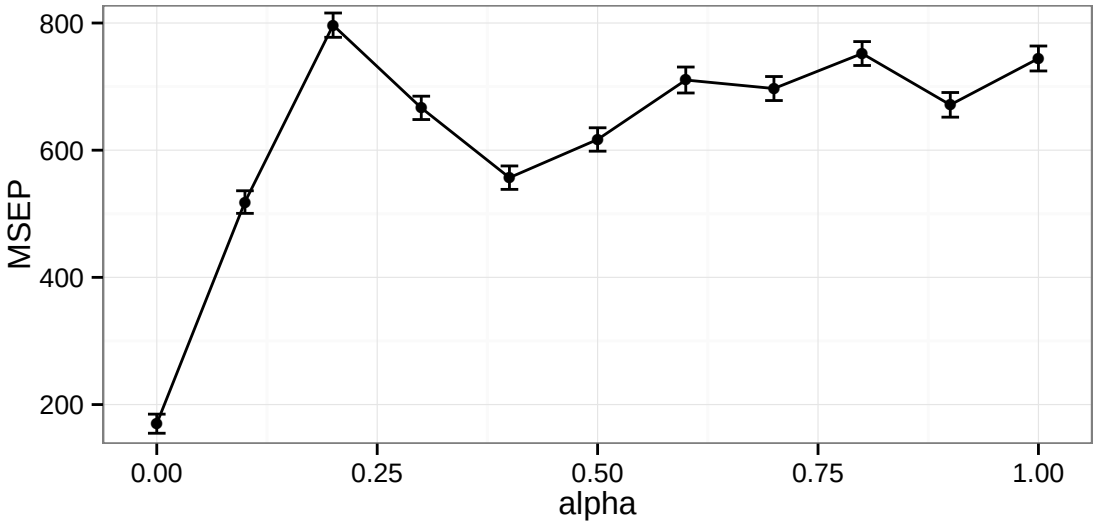


Abbildung D.5.: Plot mit lower.limits = -Inf.

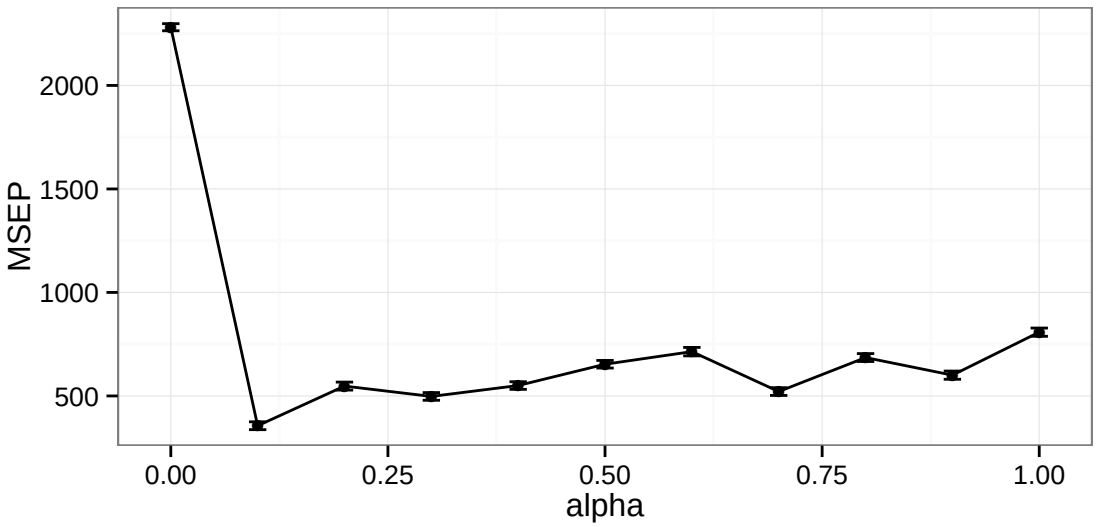


Abbildung D.6.: Plot mit lower.limits = 0.

D.6.2. Hitliste Ridge Regression

Tabelle D.9.: Top 100 Peaks Ridge Regression, lower.limits = -Inf-

Rang	m/z	Regressionskoeffizient	Annotation
1	863.370554	0.17496769	[Rif + K]$^+$, 2. Isotop
2	861.36784	0.174963402	[Rif + K]$^+$
3	862.371694	0.174908585	[Rif + K]$^+$, 1. Isotop
4	864.371634	0.174900294	[Rif + K]$^+$, 3. Isotop
5	823.411514	0.172467722	[Rif + H]$^+$
6	1043.44584	0.16977619	NA
7	824.414744	0.169697071	[Rif + H]$^+$, 1. Isotop
8	845.394644	0.169388419	[Rif + Na]$^+$
9	846.396964	0.169328599	[Rif + Na]$^+$, 1. Isotop
10	792.388694	0.167589985	[Rif - H_2 - MeO + H]$^+$, 1. Isotop
11	825.417584	0.167243383	[Rif + H]$^+$, 2. Isotop
12	789.370414	0.167039783	[Rif - H_2 - MeOH]$^+$
13	791.385994	0.166307319	[Rif - H_2 - MeO + H]$^+$
14	822.40074	0.163714363	[Rif - H_2 + H]$^+$, 1. Isotop
15	899.322884	0.163242902	NA
16	287.152664	0.162005172	NA
17	821.396124	0.161790582	[Rif - H_2 + H]$^+$
18	1044.446264	0.16013386	NA
19	877.360994	0.15747979	NA aus Rif
20	900.327474	0.151878422	NA
21	287.487654	0.150982116	NA
22	1041.432194	0.141537452	NA
23	1045.454184	0.141163315	NA
24	847.403084	0.137119365	[Rif + Na]$^+$, 2. Isotop
25	274.498674	0.135820269	NA
26	790.375574	0.134828628	[Rif - H_2 - MeOH]$^+$, 1. Isotop
27	926.509334	0.132900732	NA aus Rif
28	860.355774	0.131806433	[Rif - H_2 + K]$^+$, 1. Isotop
29	843.377864	0.127950971	[Rif - H_2 + Na]$^+$
30	859.352894	0.125074226	[Rif - H_2 + K]$^+$
31	927.512884	0.123921377	NA aus Rif
32	1041.438234	0.12343076	NA
33	865.370214	0.123040615	NA

Tabelle D.9.: Top 100 Peaks Ridge Regression, lower.limit = -Inf (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
34	865.375844	0.121576934	$[Rif + K]^+$, 3. Isotop (?)
35	925.498724	0.119613747	NA
36	859.349244	0.112073413	$[Rif - H_2 + K]^+$
37	924.4954	0.105194833	NA
38	604.383644	0.102672229	NA
39	793.390914	0.100830285	NA
40	862.367334	0.100771838	$[Rif + K]^+$, 1. Isotop
41	1203.494694	0.100616044	NA
42	847.398364	0.100262387	$[Rif + Na]^+$, 2. Isotop
43	750.462924	0.095811526	$[EryF / EryAEO + H]^+$
44	576.374544	0.095146463	$[EryA - Cladinose + H]^+$
45	522.342994	0.095086575	$[Fragment EryA + H]^+$
46	828.663644	0.095033991	NA
47	682.423924	0.094988089	NA
48	533.279624	0.094963887	NA
49	512.330964	0.094935453	NA
50	968.492984	0.094926954	NA
51	969.495144	0.094922381	NA
52	734.467014	0.094921019	$[EryA + H]^+$
53	559.367864	0.094835566	$[AEryA - Cladinose + H]^+$, 1. Isotop
54	577.377824	0.094820829	$[EryA - Cladinose + H]^+$, 1. Isotop
55	735.469544	0.09478841	$[EryA + H]^+$, 1. Isotop
56	540.35354	0.094714634	$[Fragment EryA + H]^+$
57	158.117334	0.09440523	$[Desosamin + H]^+$ aus Ery
58	720.452124	0.09435927	$[EryC / NdeMeEryA + H]^+$
59	1171.629644	0.094301143	NA
60	578.380924	0.094225031	$[EryA - Cladinose + H]^+$, 2. Isotop
61	192.138344	0.093957084	NA aus Ery
62	721.456024	0.093617243	NA
63	558.363764	0.093598775	$[AEryA - Cladinose + H]^+$
64	718.473754	0.093288989	$[EryB + H]^+$
65	736.474744	0.09314025	$[EryA + H]^+$, 2. Isotop
66	245.178724	0.092996191	NA aus Ery
67	244.844314	0.092901022	NA
68	195.792194	0.092899566	NA
69	748.448164	0.092651652	$[EryE + H]^+$

Tabelle D.9.: Top 100 Peaks Ridge Regression, lower.limit = -Inf (Fortsetzung)

Rang	m/z	Regressionskoeffizient	Annotation
70	772.424684	0.091958268	$[EryA + K]^+$
71	773.427984	0.091855893	$[EryA + K]^+$, 1. Isotop
72	737.47714	0.091323748	$[EryA + H]^+$, 3. Isotop
73	891.373644	0.091001447	NA
74	286.480324	0.090102163	NA
75	716.458264	0.089713166	NA aus Ery
76	925.490324	0.089527243	NA
77	365.157544	0.085494046	NA
78	774.429384	0.083079551	NA
79	837.566874	0.083026114	NA
80	924.49914	0.082969774	NA
81	166.266244	0.080688224	NA
82	455.01074	0.067715869	NA
83	499.219154	0.067620895	NA
84	589.408084	0.067609053	NA
85	763.560944	0.067582535	NA
86	677.458984	0.067576295	NA
87	844.390964	0.067558712	NA
88	321.131354	0.067494325	NA
89	892.380624	0.067490826	NA
90	897.306434	0.067482021	NA
91	901.38114	0.067471423	NA
92	303.721944	0.067441553	NA
93	995.512124	0.067433537	NA
94	997.510314	0.067427812	NA
95	1042.429444	0.06741221	NA
96	1169.601874	0.067405734	NA
97	1213.637334	0.067391504	NA
98	1214.63094	0.067384869	NA
99	286.814714	0.06738054	NA
100	1215.647794	0.067378107	NA

D.6.3. Hitliste Ridge Regression lower.limit=0

Tabelle D.10.: Top 100 Peaks Ridge Regression lower.limit = 0.

Rang	m/z	Regressionskoeffizient	Annotation
1	789.37041	2.68E-35	[Rif - H_2 - MeOH]$^+$
2	861.3678	2.66E-35	[Rif + K]$^+$
3	863.37055	2.65E-35	[Rif + K]$^+$, 2. Isotop
4	862.37169	2.65E-35	[Rif + K]$^+$, 1. Isotop
5	864.37163	2.64E-35	[Rif + K]$^+$, 3. Isotop
6	846.39696	2.58E-35	[Rif + Na]$^+$, 1. Isotop
7	822.4007	2.58E-35	[Rif - H_2 + H]$^+$, 1. Isotop
8	845.39464	2.58E-35	[Rif + Na]$^+$
9	821.39612	2.55E-35	[Rif - H_2 + H]$^+$
10	823.41151	2.52E-35	[Rif + H]$^+$
11	287.15266	2.48E-35	NA
12	824.41474	2.47E-35	[Rif + H]$^+$, 1. Isotop
13	792.38869	2.44E-35	[Rif - H_2 - MeO + H]$^+$, 1. Isotop
14	791.38599	2.44E-35	[Rif - H_2 - MeO + H]$^+$
15	825.41758	2.43E-35	[Rif + H]$^+$, 2. Isotop
16	899.32288	2.36E-35	NA
17	877.36099	2.27E-35	NA aus Rif
18	1043.4458	2.26E-35	NA
19	287.48765	2.15E-35	NA
20	860.35577	2.14E-35	[Rif - H_2 + K]$^+$, 1. Isotop
21	1041.43219	2.10E-35	NA
22	790.37557	2.09E-35	[Rif - H_2 - MeOH]$^+$, 1. Isotop
23	900.32747	2.07E-35	NA
24	1044.44626	2.06E-35	NA
25	859.35289	2.05E-35	[Rif - H_2 + K]$^+$
26	843.37786	2.03E-35	[Rif - H_2 + Na]$^+$
27	926.50933	2.01E-35	NA aus Rif
28	925.49872	1.90E-35	NA
29	927.51288	1.88E-35	NA aus Rif
30	865.37584	1.87E-35	[Rif + K]$^+$, 3. Isotop (?)
31	924.495	1.75E-35	NA
32	750.46292	1.69E-35	[EryF / EryAEO + H]$^+$
33	158.11733	1.69E-35	[Desosamin + H]$^+$ aus Ery

Tabelle D.10.: Top 100 Peaks Ridge Regression lower.limit = 0 (Fortsetzung).

Rang	m/z	Regressionskoeffizient	annotation
34	734.46701	1.68E-35	[EryA + H]$^+$
35	735.46954	1.68E-35	[EryA + H]$^+$, 1. Isotop
36	192.13834	1.68E-35	NA aus Ery
37	522.34299	1.68E-35	[Fragment EryA + H]$^+$
38	859.34924	1.68E-35	NA
39	559.36786	1.68E-35	[AEryA - Cladinose + H]$^+$, 1. Isotop
40	576.37454	1.68E-35	[EryA - Cladinose + H]$^+$
41	540.3535	1.68E-35	[Fragment EryA + H]$^+$
42	720.45212	1.68E-35	[EryC / NdeMeEryA + H]$^+$
43	577.37782	1.68E-35	[EryA - Cladinose + H]$^+$, 1. Isotop
44	578.38092	1.68E-35	[EryA - Cladinose + H]$^+$, 2. Isotop
45	245.17872	1.67E-35	NA aus Ery
46	244.84431	1.67E-35	NA
47	721.45602	1.67E-35	NA
48	558.36376	1.67E-35	[AEryA - Cladinose + H]$^+$
49	718.47375	1.67E-35	[EryB + H]$^+$
50	1041.43823	1.67E-35	NA
51	748.44816	1.67E-35	[EryE + H]$^+$
52	736.47474	1.66E-35	[EryA + H]$^+$, 2. Isotop
53	1045.45418	1.66E-35	NA
54	865.37021	1.65E-35	NA
55	1171.62964	1.64E-35	NA
56	737.4771	1.64E-35	[EryA + H]$^+$, 3. Isotop
57	847.40308	1.63E-35	NA
58	274.49867	1.62E-35	NA
59	195.79219	1.61E-35	NA
60	716.45826	1.60E-35	NA aus Ery
61	891.37364	1.56E-35	NA
62	286.48032	1.55E-35	NA
63	847.39836	1.52E-35	[Rif + Na]$^+$, 2. Isotop
64	773.42798	1.43E-35	[EryA + K]$^+$, 1. Isotop
65	772.42468	1.43E-35	[EryA + K]$^+$
66	365.15754	1.31E-35	NA
67	924.4991	1.30E-35	NA
68	774.42938	1.23E-35	NA
69	837.56687	1.23E-35	NA

Tabelle D.10.: Top 100 Peaks Ridge Regression lower.limit = 0 (Fortsetzung).

Rang	m/z	Regressionskoeffizient	annotation
70	793.39091	1.23E-35	NA
71	1203.49469	1.23E-35	NA
72	862.36733	1.23E-35	NA
73	166.26624	1.23E-35	NA
74	604.38364	1.22E-35	NA
75	166.25176	1.14E-35	NA
76	166.25316	1.14E-35	NA
77	763.56094	1.14E-35	NA
78	844.39096	1.14E-35	NA
79	892.38062	1.14E-35	NA
80	901.3811	1.14E-35	NA
81	1169.60187	1.14E-35	NA
82	1217.65763	1.14E-35	NA
83	1218.659	1.14E-35	NA
84	1515.28516	1.14E-35	NA
85	273.82648	1.14E-35	NA
86	286.81471	1.14E-35	NA
87	303.72194	1.14E-35	NA
88	321.13135	1.14E-35	NA
89	499.21915	1.14E-35	NA
90	589.40808	1.14E-35	NA
91	677.45898	1.14E-35	NA
92	897.30643	1.14E-35	NA
93	928.51697	1.14E-35	NA
94	995.51212	1.14E-35	NA
95	997.51031	1.14E-35	NA
96	1042.42944	1.14E-35	NA
97	1157.60229	1.14E-35	NA
98	1213.63733	1.14E-35	NA
99	1214.6309	1.14E-35	NA
100	1215.64779	1.14E-35	NA

D.6.4. Hitliste Lasso

Tabelle D.11.: Hitliste Lasso.

Rang	m/z	Regressionskoeffizient	Annotation
1	789.370411	4.889583575	[Rif - H_2 - MeOH]$^+$
2	861.36781	2.527419825	[Rif + K]$^+$

D.6.5. Hitliste Elastic Net

Tabelle D.12.: Hitliste Elastic Net.

Rang	m/z	Regressionskoeffizient	annotation
1	789.370412	1.498069946	[Rif - H_2 - MeOH]$^+$
2	822.40072	1.458236497	[Rif - H_2 + H]$^+$, 1. Isotop
3	821.396122	1.428897382	[Rif - H_2 + H]$^+$
4	365.157542	1.029233338	NA
5	861.36782	1.026108315	[Rif + K]$^+$
6	863.370552	1.017413607	[Rif + K]$^+$, 2. Isotop
7	864.371632	1.015752116	[Rif + K]$^+$, 3. Isotop
8	862.371692	1.010874777	[Rif + K]$^+$, 1. Isotop
9	1043.44582	0.963087337	NA
10	899.322882	0.956119427	NA
11	823.411512	0.881091132	[Rif + H]$^+$
12	772.424682	0.83698194	[EryA + K]$^+$
13	773.427982	0.828142899	[EryA + K]$^+$, 1. Isotop
14	860.355772	0.816591813	[Rif - H_2 + K]$^+$, 1. Isotop
15	925.498722	0.792356527	NA
16	824.414742	0.760011981	[Rif + H]$^+$, 1. Isotop
17	846.396962	0.734540702	[Rif + Na]$^+$, 1. Isotop
18	845.394642	0.730336356	[Rif + Na]$^+$
19	1044.446262	0.727070275	NA
20	859.352892	0.658621336	[Rif - H_2 + K]$^+$
21	825.417582	0.656098091	[Rif + H]$^+$, 2. Isotop
22	792.388692	0.653419515	[Rif - H_2 - MeO + H]$^+$, 1. Isotop
23	774.429382	0.607528527	NA

Tabelle D.12.: Hitliste Elastic Net (Fortsetzung).

Rang	m/z	Regressionskoeffizient	annotation
24	837.566872	0.606983922	NA
25	791.385992	0.602623753	[Rif - H_2 - MeO + H]$^+$
26	877.360992	0.571565918	NA aus Rif
27	1045.454182	0.543591037	NA
28	396.259822	0.445239915	NA
29	287.152662	0.414644301	NA
30	750.462922	0.394249028	[EryF / EryAEO + H]$^+$
31	158.117332	0.371140319	[Desosamin + H]$^+$ aus Ery
32	175.11872	0.344521335	NA
33	576.374542	0.331254953	[EryA - Cladinose + H]$^+$
34	522.342992	0.318213733	[Fragment EryA + H]$^+$
35	192.138342	0.317331089	NA aus Ery
36	577.377822	0.307047867	[EryA - Cladinose + H]$^+$, 1. Isotop
37	559.367862	0.298558645	[AEryA - Cladinose + H]$^+$, 1. Isotop
38	540.35352	0.288531664	[Fragment EryA + H]$^+$
39	734.467012	0.273264127	[EryA + H]$^+$
40	720.452122	0.26989468	[EryC / NdeMeEryA + H]$^+$
41	735.469542	0.26485775	[EryA + H]$^+$, 1. Isotop
42	790.375572	0.264735211	[Rif - H_2 - MeOH]$^+$, 1. Isotop
43	578.380922	0.25344551	[EryA - Cladinose + H]$^+$, 2. Isotop
44	865.375842	0.217256915	[Rif + K]$^+$, 3. Isotop (?)
45	245.178722	0.215999031	NA aus Ery
46	926.509332	0.214460262	NA aus Rif
47	244.844312	0.213959382	NA
48	721.456022	0.213860584	NA
49	558.363762	0.203353074	[AEryA - Cladinose + H]$^+$
50	736.474742	0.20088457	[EryA + H]$^+$, 2. Isotop
51	737.47712	0.188800259	[EryA + H]$^+$, 3. Isotop
52	718.473752	0.188093438	[EryB + H]$^+$
53	900.327472	0.185458386	NA
54	748.448162	0.145653608	[EryE + H]$^+$
55	287.487652	0.135410635	NA
56	924.49912	0.126688524	NA
57	828.663642	0.052586845	NA
58	512.330962	0.0520094	NA
59	969.495142	0.052006048	NA

Tabelle D.12.: Hitliste Elastic Net (Fortsetzung).

Rang	m/z	Regressionskoeffizient	annotation
60	533.279622	0.05179464	NA
61	968.492982	0.051781666	NA
62	682.423922	0.051537491	NA
63	924.4952	0.049391709	NA
64	843.377862	0.023394941	[Rif - H_2 + Na]$^+$

E. Korrelationstabelle: Experimente im Laborbuch, 3LC

Tabelle E.1.: Korrelationstabelle: Experimente im Laborbuch, 3LC. Die Nummer des Laborjournals und die entsprechende Seitenzahl sind in Klammern angegeben.

Arbeitsschritt	Acora: Proof of Concept	Multivariate Methoden zur Datenanalyse	Acora mit S. Ampullosporum
Extrakterstellung	K. Michels (ca. Juni 2008)	K. Michels (ca. Juni 2008)	HAM087 (LB04/S.79)
FT-ICR-MS Messung	HAM066_POC2 (LB02/S.92)	HAM066_POC2 (LB02/S.92)	MAA027 (LB04/S.164)
Bioaktivitätsassay	HAM066 (LB02/S.112)	HAM066 (LB02/S.112)	HAM102 (LB05/S.13)
AcorA, XCMS	HAM066 (LB02/S.112)	HAM066(LB05/S.13)	HAM114 (LB05/S.109)
Messung Antibiotika	HAM061 (LB02/S.83)	HAM061 (LB02/S.83)	
Auswertung Antibiotika		HAM066 (LB03/S.82)	
Strukturaufklärung Peptaibole			HAM129 (LB06/S. 78), HAM129IT (LB06/S. 87)
Isolierung Verbindung **61** (AmpA)			HAM115 (LB05/S.141)
Testung Verbindung **61** (AmpA)			HAM121 (LB05/S.11)

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Des Weiteren versichere ich, dass ich diese Arbeit an keiner anderen Institution eingereicht habe.

München, 10.08.2018 Mark Haid

Publikationsverzeichnis

Publikationen

F. D. Henkel, A. Friedl, **M. Haid**, D. Thomas, T. Bouchery, P. Haimerl, M. de los Reyes Jimenez, F. Alessandrini, C. B. Schmidt-Weber, N. L. Harris, J. Adamski, J. Esser-von-Bieren: House dust mite drives pro-inflammatory eicosanoid reprogramming and macrophage effector functions. *Allergy*, 2019, 74(6), 1090-1101

M. Haid, C. Muschet, S. Wahl, W. Römisch-Margl, C. Prehn, G. Möller, J. Adamski: Long-Term Stability of Human Plasma Metabolites during Storage at -80 °C. *Journal of Proteome Research*, 2018, 17 (1), 203-211

S. Molnos, S. Wahl. **M. Haid**, E. M. W. Eekhoff, R.Pool, A. Floegel, J. Deelen, D. Much, C. Prehn, M. Breier, H. H. Draisma, Ni. van Leeuwen, A. M. C. Simonis-Bik, A.Jonsson, G. Willemsen, W. Bernigau, R. Wang-Sattler, K. Suhre, A. Peters, B. Thorand, C. Herder, W. Rathmann, M. Roden, C. Gieger, Mark H. H. Kramer, D. van Heemst, H. K. Pedersen, V. Gudmundsdottir, M. B. Schulze, T. Pischon, E. J. C. de Geus, H.Boeing, D. I. Boomsma, A. G. Ziegler, P. E. Slagboom, S.Hummel, M. Beekman, H. Grallert, S. Brunak, M. I. McCarthy, R. Gupta, E. R. Pearson, J. Adamski, L.M. 't Hart: Metabolite ratios as potential biomarkers for type 2 diabetes: a DIRECT study. *Diabetologia*, 2018, 61 (1), 117-129

J. Tokarz, **M. Haid**, A. Cecil, C. Prehn, A. Artati, G. Möller, J. Adamski: Endocrinology meets metabolomics: achievements, pitfalls, and challenges. *Trends in Endocrinology & Metabolism*, 2017, 28 (10), 705-721

A. Otto, A. Laub, **M. Haid**, A. Porzel, J. Schmidt, L. Wessjohann, N. Arnold: Tulasporins A-D, 19-residue peptaibols from the mycoparasitic fungus Sepedonium tulasneanum. *Natural Product Communications*, 2016, 11 (12), 1821-1824

I. Schäffler, K. E. Steiner, **M. Haid**, S. S. Van Berkel, G. Gerlach, S. D. Johnson, L. Wessjohann, S. Dötterl: Diacetin, a reliable cue and private communication channel in a specialized pollination system. *Scientific Reports*, 2015, 5, Article number: 12779

D. N. Quang, J. Schmidt, A. Porzel, L.Wessjohann, **M. Haid**, N. Arnold: Ampullosine, a new isoquinoline alkaloid from Sepedonium ampullosporum (Ascomycetes). *Natural Product Communications*, 2010, 5 (6), 869-872

A. Kühlberg, **M. Haid**, S. Metzger: Characterization of O-phosphohydroxyproline in rat alpha-crystallin A. *Journal of Biological Chemistry*, 2010, 285 (41), 31484-31490

Vorträge

M. Haid, C. Prehn, S. Becker, V. Raverdy, D. Merkel, J. Dojahn, F. Pattou, J. Adamksi: The Lipidyzer Platform reveals Lipidomics Signatures of T2D Patients with Roux-En-Y Gastric Bypass Surgery: an IMI DIRECT Study. *7th European Lipidomics Meeting*, Leipzig, 2018

M. Haid, C. Prehn, S. Becker, V. Raverdy, D. Merkel, J. Dojahn, F. Pattou, J. Adamksi: Comprehensive Lipidomics Analysis of Plasma Lipid Changes upon Bariatric Surgery using the Lipidyzer Platform: an IMI DIRECT Study. *14th Annual Conference of the Metabolomics Society*, Seattle, 2018

M. Haid, C. Prehn, S. Becker, V. Raverdy, D. Merkel, J. Dojahn, F. Pattou, J. Adamksi: The Lipidyzer Platform in biomedical research, *Analytica Fachmesse (Sciex Seminar)*, München, 2018

M. Haid, C. Prehn, S. Becker, V. Raverdy, D. Merkel, J. Dojahn, F. Pattou, J. Adamksi: Der Lipidyzer: Ein neues Tool zur Targeted-Lipidomics Analyse. Ergebnisse in einer klinischen T2D Studie. *7. Berliner LC-MS/MS Symposium*, Berlin, 2017

M. Haid: Metabolomics in Type 2 Diabetes Research. *1st Sanofi Biomarker Symposium*, Frankfurt, 2015

M. Haid: Effect of long-term storage on metabolite concentrations. *Biocrates User Meeting: 1st Workshop on Targeted Metabolomics*, München, 2015

M. Haid, K. Michels, A. Gohr, L. Wessjohann: Reverse Metabolomics statt Bioguided Fractionation. *38. Doktorandenworkshop: Naturstoffe: Chemie, Biologie und Ökologie*. Halle (Saale), 2009

M. Haid, K. Michels, A. Gohr, L. Wessjohann: Reverse Metabolomics: A new method for direct identification of bioactive compounds in complex mixtures (Postervortrag). *DECHEMA-Naturstofftagung*, Irsee, 2009

Poster

M. Haid, M. Rudisch, C. Muschet, J. Adamski: A target metabolomics assay for absolute quantification of omega-3 and omega-6 oxylipins in human plasma. *12th Annual Conference of the Metabolomics Society*, Dublin, 2016

M. Haid, C. Muschet, C. Prehn, W. Römisch-Margl, J. Adamksi: Long-term stability of plasma metabolites stored at -80°C, *11th Annual Conference of the Metabolomics Society*, San Francisco, 2015

C. Prehn, **M. Haid**, S. Sabrautzki, B.Rathkolb, B. Lorenz-Depiereux, H. Fuchs, E. Wolf, T.-M. Strom, M. Hrabe de Angelis, J. Adamski: Metabolite Alterations in ENU mutagenesis derived ASGR1^{MHDABAP005} mice as a model for human ideopathic hyperphosphatasemia. *9th Annual Conference of the Metabolomics Society*, Glasgow, 2013

C. Prehn, **M. Haid**, S. Häußler, G. Möller, J. Adamski, H. Sauerwein: Lipid mobilization in early-lactating dairy cows: steroid metabolism. *2nd Congress on Steroid Research*, Chicago, 2013

K. Michels, **M. Haid**, A. Gohr, L. Wessjohann: Reverse Metabolomics: AcorA a new method for direct identification of bioactive compounds in complex mixtures, *43. Jahrestagung der Deutschen Gesellschaft für Massenspektrometrie*, Halle (Saale), 2010

A. Kühlberg, **M. Haid**, S. Metzger: Phosphohydroxyprolin, eine proteinogene Aminosäure. *38. Jahrestagung der Deutschen Gesellschaft für Massenspektrometrie*, Rostock, 2005

Mark Haid

Curriculum Vitae

Persönliche Angaben

Name	Mark Haid
Geburtsdatum	20.04.1973
Geburtsort	Moers
Familienstand	Ledig
Staatsangehörigkeit	Deutsch

Berufspraxis

2018–heute **Leiter Lipidomics Plattform**, *Helmholtz-Zentrum Für Gesundheit und Umwelt, Molekulare Endokrinologie & Metabolismus*, München.

2012–2018 **Wissenschaftl. Mitarbeiter, Assay-Entwicklung und Projektkoordination**, *Helmholtz-Zentrum Für Gesundheit und Umwelt, Genomanalyse Zentrum*, München.

2007–2012 **Wissenschaftl. Mitarbeiter (Doktorand)**, *Leibniz-Institut für Pflanzenbiochemie, Abt. Natur-und Wirkstoffchemie*, Halle/Saale.
Thema: Identifzierung eigenschaftsrelevanter Metabolitencluster

2005–2006 **Wissenschaftl. Mitarbeiter (Diplomand)**, *Biologisch-Medizinisches Forschungszentrum*, Düsseldorf.
Thema: Massenspektrometrische Charakterisierung von Hydroxyprolin / Phosphohydroxyprolin und Identifizierung von Phosphohydroxyprolin in Kristallinen

Ausbildung

2007–2012 **Leibniz Institut für Pflanzenbiochemie**, *Wissenschaftl. Mitarbeiter (Doktorand)*, Halle/Saale.

1996-2006 **Heinrich-Heine-Universität**, *Studium der Biologie (Diplom)*, Düsseldorf.

1994–1995 **Diakonie der ev. Kirchengemeinde**, *Zivildienst, Altenpflege*, Moers.

1984–1994 **Gymnasium in den Filder Benden**, *Hochschulreife*, Moers.