Aus dem Institut für Medizintechnologie
der Universität Heidelberg und der Hochschule Mannheim
(Geschäftsführender Direktor: Prof. Dr. med. Norbert Gretz)

# Automatische Datenanalyse von massenspektrometrischen Signaturen zur Klassifikation von Krebszellen und Bestimmungen von Wirkstoffwirkungen

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Theresia Salonikios

aus
Freiburg im Breisgau
2017

Dekan: Prof. Dr. med. Sergij Goerdt

Referent: Prof. Dr. rer. nat. Carsten Hopf

# INHALTSVERZEICHNIS

Seite

# ABKÜRZUNGSVERZEICHNIS

| | |
|---|---|
| ALL | Akute lymphatische Leukämie |
| AML | Akute myeloische Leukämie |
| ANN | *Artificial Neural Network* |
| ANOVA | *Analysis of Variance* |
| BLogReg | Bayes'sche Logistische Regression |
| CLL | Chronisch lymphatische Leukämie |
| CML | Chronisch myeloische Leukämie |
| CMM | *Confusion Matrix Maximum* |
| Da | Dalton |
| *ER* | *Estrogen Receptor* |
| GA | Genetische Algorithmen |
| GIST | Gastrointestinale Stromatumoren |
| HCA | Hierarchische Clusteranalyse |
| HDAC | Histon-Deacetylasen |
| HER2 | *Human Epidermal Growth Factor Receptor 2* |
| KNN | K-Nächste-Nachbarn |
| LDA | Lineare Diskriminanzanalyse |
| m | Mittelwert |
| MALDI | Matrix-assistierte-Laser-Desorption/Ionisation |
| mRMR | *minimum Redundancy Maximum Relevance* |
| MS | Massenspektrometrie |
| MSE | *Mean Squared Error* |
| PAM | *Partitioning Around Medoids* |
| *PCR* | *Principal Component Analysis* |
| PgR | Progesteron-Rezeptor |
| SNR | *Signal to Noise Ratio* |
| SOP | *Standard Operating Procedure* |
| TIC | *Total Ion Current* |
| TOF | *Time of Flight* |

# 1 EINLEITUNG

## 1.1 Personalisierte Medizin in der Krebsforschung

Krebs ist eine der häufigsten Todesursachen weltweit. Aufgrund der steigenden Lebenserwartung der Menschen hat auch die Zahl der Neuerkrankungen zugenommen. Laut einer Schätzung des Zentrums für Krebsregisterdaten sind im Jahr 2012 in Deutschland 225.890 Männer und 252.060 Frauen an Krebs erkrankt. Krebs entsteht durch Veränderungen in der Erbinformation, infolgedessen sich die betroffenen Zellen unkontrolliert teilen. Im weiteren Verlauf können sie das umliegende Gewebe infiltrieren und in die Blutgefäße gelangen. Über die Blut- bzw. Lymphbahnen können sie sich im Körper ausbreiten und an entfernten Stellen Metastasen bilden. Zur Bekämpfung dieser Krankheit steht seit einigen Jahren die Entwicklung personalisierter Therapieansätze (engl. *Targeted Therapies*) im Fokus der klinischen Forschung. Durch diese sollen dem Patienten unnötige Nebenwirkungen erspart bleiben. Des Weiteren kann ein Zeitverlust vermieden werden, wenn gleich zu Beginn eine geeignete Therapie eingesetzt wird. Zudem ist die Wahrscheinlichkeit des Auftretens von Resistenzen geringer, wenn bei Therapiebeginn die Tumormasse kleiner ist. Durch eine individuell optimale Dosierung der Medikamente kann die Lebensqualität bei gleicher Erfolgsquote verbessert werden oder zu einer höheren Erfolgsquote bei gleichen Nebenwirkungen führen (Gerber, 2008; Ross et al., 2004; Wu et al., 2006).

Die Basis für eine personalisierte Medizin ist das Verständnis grundlegender Krankheitsmechanismen auf molekularer Ebene. Humane Proben von Tumoren müssen schnell und zuverlässig klassifiziert und dem entsprechenden Tumorsubtyp zugeordnet werden, um dem Patienten eine solche personalisierte Therapie zu ermöglichen.

Bei den zielgerichteten Krebstherapien werden molekular-gerichtete Medikamente verwendet, welche sich möglichst nur gegen Tumorzellen richten. Eine Gruppe dieser Medikamente besteht aus den Signaltransduktionshemmern, welche die Wachstumssignale der Krebszellen unterdrücken. Ansatzpunkte hierfür sind die Rezeptoren auf der Zelloberfläche und Signalübertragungsketten im Zellinneren. Dazu gehören zum Beispiel die sogenannten Tyrosinkinasehemmer. Als Kinasen werden bestimmte Enzyme bezeichnet, welche an der Signalübertragung beteiligt sind. Eine andere Gruppe besteht aus den Angiogenesehemmer, welche die Bildung neuer Blutgefäßen hemmen, sodass der Tumor nicht mehr mit Sauerstoff und Nährstoffen versorgt werden kann. Proteasom-Hemmer dienen dazu, die Müllentsorgung der Zellen zu blockieren. Dadurch ersticken die Zellen an

ihrem eigenen Abfall und es kommt zum programmierten Zelltod (Apoptose). Das sogenannte PARP-Enzym hilft den Zellen dabei, geschädigtes Erbgut (DNA) zu reparieren. Da in Krebszellen DNA-Reparaturmechanismen häufig gestört sind, kann durch die Einnahme von PARP-Hemmern der Schaden so groß werden, dass die Krebszellen schließlich absterben. Mit Hilfe von Immuntherapien soll die körpereigene Abwehr zur Bekämpfung der Krebszellen genutzt werden (Finley, 2003). Viele zielgerichtete Wirkstoffe sind noch nicht allgemein oder nur für bestimmte Krebsarten verfügbar und müssen noch in klinischen Studien geprüft werden (Dubreuil et al., 2009; Lancet et al., 2011).

Im Folgenden werden die Krebsformen, welche im Rahmen dieses Dissertationsprojektes eine Rolle spielen, im Kontext einer zielgerichteten Therapie kurz erläutert.

### 1.1.1 Brustkrebs

Die häufigste Krebserkrankung bei Frauen ist Brustkrebs. Im Jahr 2012 sind 69.550 Frauen daran erkrankt. Mammakarzinome entwickeln sich im Drüsengewebe oder in den Milchgängen der weiblichen Brust. Das Wachstum der Brustkrebstumore ist häufig von den Geschlechtshormonen abhängig. Wenn im Tumorgewebe Östrogen-Rezeptoren (engl. *Estrogen Receptor*, ER) oder Progesteron-Rezeptoren (PgR) nachgewiesen werden können, kann der Patientin eine Antihormontherapie empfohlen werden. Diese soll die Produktion der Geschlechtshormone unterdrücken oder deren Wirkung blockieren. Bewährte Medikamente hierfür sind Tamoxifen (Paik et al., 2004), Aromatasehemmer (Ellis et al., 2012) und sogenannte GnRH-Analoga (Torrisi et al., 2007).

Ein weiterer Wachstumfaktor-Rezeptor ist der sogenannte HER2-Rezeptor (engl. *Human Epidermal Growth Factor Receptor 2*, HER2). Eine große Zahl dieser Rezeptoren treibt den Tumor besonders zum Wachstum an. Hier haben sich in den letzten Jahren große Fortschritte bei der medikamentösen Behandlung ergeben. Ein Beispiel hierfür ist der Wirkstoff Trastuzumab (Handelsname Herceptin®), welcher zielgerichtet die Vermehrung der Brustkrebszellen hemmen kann (Kaufman et al., 2009; Marty et al., 2005; Moja et al., 2012).

Um eine gute Therapieempfehlung geben zu können, wurde der Brustkrebs in verschiedene tumorbiologische Untergruppen (Subtypen) eingeteilt (Tabelle 1).

**Tabelle 1: Immunhistologische Subtypen und therapeutische Empfehlungen (Goldhirsch et al., 2011)**

| Molekular-biologische Subtypen | Klinisch-pathologische Definition | Behandlung |
|---|---|---|
| Luminal A | **„Luminal A“**<br>ER und/oder PgR positiv<br>HER2 negativ | Hormontherapie |
| Luminal B | **„Luminal B (HER2 negativ)“**<br>ER und/oder PgR positiv<br>HER2 negativ | Hormontherapie<br>(+ Chemotherapie) |
| | **„Luminal B (HER2 positiv)“**<br>ER und/oder PgR positiv<br>HER2 positiv | Hormontherapie<br>+ Chemotherapie<br>+ Anti-HER2 |
| HER2 überexprimiert | **„HER2 positiv“**<br>ER und PgR negativ<br>HER2 überexprimiert oder erhöht | Chemotherapie<br>+ Anti-HER2 |
| Basal Like | **„Triple negativ“**<br>ER und PgR negativ<br>HER2 negativ | Chemotherapie |

### 1.1.2 Leukämie

Leukämie, auch Blutkrebs genannt, ist ein Sammelbegriff für eine Gruppe von Erkrankungen des blutbildenden Systems, bei welchen die Reifung der Blutzellen gestört ist. Die Blutstammzellen werden alle im Knochenmark gebildet, wobei zunächst zwei unterschiedliche Arten von Tochterzellen entstehen: die myeloischen und lymphatischen Vorläuferzellen. Über mehrere Zwischenstufen entwickeln sich daraus dann die roten und die weißen Blutkörperchen und die Blutplättchen. Die roten Blutkörperchen (Erythrozyten) sind für den Sauerstofftransport verantwortlich, die weißen Blutkörperchen (Leukozyten) spielen eine Rolle bei der Immunabwehr und die Blutplättchen (Thrombozyten) helfen bei der Blutgerinnung. Bei der Leukämie werden vermehrt unreife Vorstufen der weißen Blutkörperchen gebildet, welche sich nicht mehr weiter entwickeln. Es wird zwischen myeloischen und lymphatischen Leukämien unterschieden, bei welchen die Zellen vermehrt von myeloischen bzw. von lymphatischen Vorläuferzellen abstammen. Des Weiteren unterscheidet man zwischen einer akuten Leukämie, welche sich innerhalb weniger Wochen entwickelt und schwere Krankheitssymptome zeigt, und einer chronischen Leukämie, welche langsamer fortschreitet und erst nach einem längeren Zeitraum die ersten Symptome erscheinen. Es werden die folgenden Formen der Leukämie unterschieden:

- Akute lymphatische Leukämie (ALL)
- Akute myeloische Leukämie (AML)
- Chronisch myeloische Leukämie (CML)
- Chronisch lymphatische Leukämie (CLL)

Für die Diagnose wird eine Blutuntersuchung durchgeführt und eine Probe des Knochenmarks entnommen. Bei der akuten Leukämie besteht die Behandlung aus einer Chemotherapie und Medikamenten gegen die Nebenwirkungen. Bei der chronischen myeloischen Leukämie hingegen weisen die Patienten meistens eine bestimmte Chromosomveränderung auf („Philadelphia-Chromosom“) und können mit zielgerichteten Medikamenten, den sogenannten Tyrosinkinasehemmern wie zum Beispiel Imatinib (Handelsname Glivec®) behandelt werden (Druker et al., 2006).

### 1.1.3 Gastrointesinale Stromatumoren

Gastrointestinale Stromatumoren (GIST) sind bösartige Bindegewebstumoren, welche hauptsächlich im Verdauungstrakt entstehen. Diese Krebsform tritt verhältnismäßig selten auf. In Deutschland gibt es ca. 800 - 1200 Neuerkrankungen im Jahr, wobei Männer etwas häufiger betroffen sind als Frauen. GIST wird oft zufällig bei Routineuntersuchungen entdeckt, da sie unspezifische Beschwerden verursachen. In einem frühen Stadium kann der Tumor operiert werden, wenn er noch recht klein und noch nicht in das umliegende Gewebe eingewachsen ist. Bei fortgeschrittener Erkrankung wird auch hier Imatinib gegeben. Die Ursache von GIST ist eine Genmutation des KIT-Rezeptors oder des PDGF-Rezeptors (engl. *Platelet Derivated Growth Factor*, PDGF). Sind keine Mutationen des KIT-Gens oder des PDGF-Gens nachweisbar, spricht man vom „Wild-Typ“. In manchen Fällen liegt jedoch eine Resistenz gegen den selektiven Wirkstoff Imatinib vor, oder nach längerer Behandlung wirkt das Medikament nicht mehr (sekundäre Resistenz). Dann kann zum Beispiel auf den Tyrosinkinasehemmer Sunitinib (Handelsname Sutent®) ausgewichen werden (Nannini et al., 2013; Rammohan et al., 2013).

## 1.2 Massenspektrometrie als Analysemethode

Instrumentelle Analysen, welche eine exakte Diagnostik (die Basis für eine personalisierte Therapie) und eine zuverlässige Prävention gewährleisten, können nur durch analytische Hochleistungsgeräte bewerkstelligt werden. Ein vielversprechendes analytisches Verfahren, welches bereits erfolgreich für die Klassifizierung von Zellpopulationen eingesetzt wurde, ist die Massenspektrometrie (MS).

Mit Hilfe eines Massenspektrometers können die *m/z*-Werte von Atomen oder Molekülen bestimmt werden. Es besteht aus einer Ionenquelle, einem Analysator und einem Detektor. Die Ionenquelle dient zur Erzeugung von Ionen. Dabei wird die zu untersuchende Substanz in die Gasphase überführt und ionisiert. Die „Matrix-assistierte-Laser-Desorption/Ionisation“ (MALDI) (Karas et al., 1987) stellt hierbei ein besonders sanftes Ionisierungsverfahren dar,

bei welchem die Probe mit einer organisch-chemischen Matrix, welche Licht mit der Wellenlänge des verwendeten Lasers absorbiert, kokristallisiert. Die geladenen Moleküle werden dann durch ein elektromagnetisches Feld beschleunigt und dem Analysator zugeführt, der sie entsprechend ihres Masse-zu-Ladung-Verhältnisses (*m/z*) auftrennt. Dies muss immer in einem Hochvakuum stattfinden, da bei zu hohem Druck die Ionen mit den Luftmolekülen zusammenstoßen und von ihrem Weg abkommen könnten. Der Detektor fängt die Ionen zur Registrierung auf. Oft werden sogenannte Flugzeitmassenspektrometer (engl. *Time of Flight*, TOF) verwendet (Guilhaus et al., 1997; Weickhardt et al., 1996). Durch die unterschiedlichen Massen ergeben sich unterschiedliche Laufzeiten, welche als elektronisches Signal erfasst werden. Das resultierende Massenspektrum stellt die Intensität bzw. Häufigkeit der Ionen (dimensionslos) in Abhängigkeit ihres *m/z*-Wertes dar, welcher in der Biomedizin in Dalton [Da] angegeben wird (Gross, 2012).

Der Vorteil einer linearen Flugbahn (Linearmodus) besteht darin, dass ein Zerfall von Ionen nach der Beschleunigung keine Auswirkungen auf das Massenspektrum ausübt, da sich die Geschwindigkeit der Bruchstücke nachdem Zerfall nicht wesentlich ändert. Sie erreichen zum gleichen Zeitpunkt den Detektor, an dem auch das ursprüngliche Ion angekommen wäre. Die Ionen besitzen jedoch nach Verlassen der Ionenquelle weder die gleiche Startzeit noch die gleiche kinetische Energie, wodurch die Massenauflösung (vor allem im höheren Massenbereich) verschlechtert wird.

Diese Unterschiede können durch einen Reflektor-TOF-Analysator (Mamyrin, 1994) kompensiert werden. Im Reflektor-TOF-Analysator dient der Reflektor als Ionenspiegel, welcher Ionen mit unterschiedlichen kinetischen Energien zeitlich fokussiert. Ein einfacher Reflektor besteht aus einem hemmenden elektrischen Feld, welches sich hinter der feldfreien Driftregion gegenüber der Ionenquelle befindet. Die Ionen dringen in den Reflektor ein, wo sie nach und nach ihre kinetische Energie verlieren. Dabei dringen Ionen mit höherer kinetischer Energie tiefer in das Bremsfeld ein und halten sich länger im Reflektor auf als energieärmere Ionen. Anschließend werden sie in entgegengesetzter Richtung aus dem Reflektor ausgeworfen. Dadurch wird eine Flugzeitkorrektur erreicht, welche das Auflösungsvermögen des TOF-Analysators gegenüber einer linearen Laufbahn erheblich verbessert (Cornish und Cotter, 1997). Jedoch geht aber auch aufgrund von Ionenverlust der Vorteil der hohen Empfindlichkeit der TOF-Massenspektrometer etwas verloren, da vor dem Reflektron zerfallende Ionen nicht detektiert werden können.

Die MALDI Massenspektrometrie gestattet eine simultane Analyse einer Vielzahl von Biomolekülen wie Lipide, Peptide oder Proteine. Somit können zum Beispiel viele abundante Proteine einer Zellpopulation gleichzeitig erfasst werden. Unter standardisierten Bedingungen entstehen zellspezifische Signaturen, ein reproduzierbares Muster, genannt

*Fingerprints*. Diese können dazu verwendet werden, um verschiedene Zelllinien voneinander zu unterscheiden. Die MALDI-TOF-MS basierte Klassifizierung findet seit ihrer Einführung durch Anhalt und Fenselau (Anhalt und Fenselau, 1975) mittlerweile breite Anwendung in der Qualitätskontrolle von Lebensmitteln (Barreiro et al., 2012; Herrero et al., 2012), der Umwelttechnologie (Koubek et al., 2012) und der klinischen Pathologie zur Identifizierung von Mikroorganismen (Croxatto et al., 2012; van Veen et al., 2010).

## 1.3 Klassifikation von Tumortypen und Tumorsubtypen

Die Herausforderung der modernen Krebsdiagnostik besteht in der schnellen und zuverlässigen Klassifizierung humaner Tumorproben. Dabei spielt die Identifikation des jeweiligen Tumorsubtypes eine zentrale Rolle, um dem Patienten eine personalisierte Behandlung zu ermöglichen. Bisher wurde die Entscheidung über Diagnose und Therapie vorwiegend durch die klassische immunhistopathologische Beurteilung des Biopsiegewebes durch den behandelnden Arzt getroffen. Gewebeaufbereitung, Antikörper- oder PCR-basierte Analytik nimmt dabei sehr viel Zeit in Anspruch (Gerber, 2008; Ross et al., 2004; Wu et al., 2006). Im Gegensatz zu histopathologischen Analyseverfahren erhofft man sich durch MS-basierte Methoden eine bedeutend schnellere und vor allem vom Erfahrungswert des Arztes unabhängige Klassifizierung. Es wurde bereits gezeigt, dass durch die MALDI-TOF-MS malignes Gewebe von gesundem unterschieden werden kann und sogar unterschiedliche Subtypen klassifiziert werden können (Gámez-Pozo et al., 2009; Liao et al., 2010). Bei der Messung von Massenspektren entstehen jedoch Datensätze im Umfang von mehreren Gigabytes, deren Bearbeitung und Auswertung meist einen verhältnismäßig großen Aufwand darstellt und nur durch Hilfe moderner Signalverarbeitung ermöglicht wird. Oftmals führt erst ein zeitintensives Austesten verschiedener Einflussgrößen oder gar ein mühsames Durchsuchen des zu analysierenden Datensatzes „per Hand“ zu den erwünschten Informationen. Die Erforschung und Entwicklung systematischer Arbeitsprozesse zur automatischen Klassifikation humaner Karzinome wäre daher hilfreich, um in Zukunft medizinisch relevante Merkmale schneller und zuverlässiger als bisher extrahieren zu können. Dazu bedarf es zunächst jedoch einer Methodenetablierung anhand eindeutig charakterisierter Zelllinien eines definierten Tumorsubtypes.

Aufgrund der biologischen Komplexität von Säugerzellen war es jedoch lange nicht möglich, Klassifikationen anhand von massenspektrometrischen Protein-*Fingerprints* durchzuführen. Zhang et al. (Zhang et al., 2006) zeigten erstmals, dass mit Hilfe von *Fingerprints* auch eine Unterscheidung verschiedener Säugerzellen durchgeführt werden kann. Die Unterscheidung wurde hierbei visuell anhand einiger herausragender Peaks durchgeführt. Es folgten weitere

visuelle Betrachtungen herausragender Peaks von *Fingerprints* zur Proteinanalyse von Hühner-Makrophagen (Kannan et al., 2007), zur Untersuchung von Peptidhormonen in Pankreas-Zelllinien (Buchanan et al., 2007) und zur Identifikation von apoptotischen Zellen (Dong et al., 2011).

Um die Zell-Klassifikation mit Hilfe statistischer Methoden durchführen zu können, bedarf es zunächst einer geeigneten Vorverarbeitung der *Fingerprints*. Wird die biologische Probe durch verschiedene Personen, an verschiedenen Messtagen oder mit unterschiedlichen Parametereinstellungen vermessen, kann dies zu einer Variabilität der spektralen Information führen. Auch das Messgerät selbst kann zum Beispiel durch Detektorungenauigkeiten Schwankungen verursachen. Einfluss auf die Präzision der Daten haben auch die Messdauer und die spektrale Auflösung. Abbildung 1 zeigt die Arbeits- schritte, welche standardmäßig für die Aufbereitung der Massenspektren verwendet werden (Hilario et al., 2006). Ein Softwarepaket, welches oft für die Vorverarbeitung und Analyse verwendet wird, ist *ClinProTools*, welches von der Firma Bruker Daltonik zum Messgerät mitgeliefert wird (Ketterlinus et al., 2005). Des Weiteren bietet MATLAB (The MathWorks, Inc.), eine kommerzielle Software zur Lösung mathematischer Probleme und zur graphischen Darstellung der Ergebnisse, die *Bioinformatics Toolbox* an, in welcher speziell für Massenspektren Funktionen zur Vorverarbeitung bereitgestellt werden (Henson und Cetto, 2005).

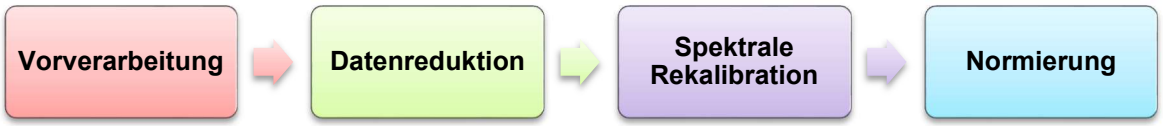


**Abbildung 1: Standard-Verarbeitung von Massenspektren**

Für eine optimale Zell-Klassifikation ist es neben der Beseitigung den eben erwähnten Schwankungen, welche bei einer MS-Messung entstehen können, sehr wichtig, dass zunächst die enorme Datendimension der Massenspektren reduziert wird. Die relevante Information ist in einem nur sehr geringen Teil der *Fingerprints* enthalten. Schon ein paar wenige Signale reichen erfahrungsgemäß zur Unterscheidung mehrerer Zelllinien aus. Durch Entfernen der unwichtigen Information kann das Klassifikationsergebnis um ein Vielfaches verbessert werden.

Ein weiterer wichtiger Punkt ist die spektrale Rekalibration der *Fingerprints*. Zum einen müssen die Peaks dem genauen *m/z*-Wert zugeordnet werden, um sie später identifizieren zu können. Zum anderen ist es für eine Klassifikation (und andere statistische Analysen) zwingend notwendig, dass alle Massenspektren eine gemeinsame Massenachse besitzen.

Der Klassifikator erwartet, dass die Spektren in Form einer Matrix vorliegen, bei welcher die Intensitäten der Spektren in den Zeilen und die *m/z*-Werte in den Spalten stehen. Ist dieser Sachverhalt nicht gegeben, kann keine Klassifikation durchgeführt werden.

Für die Entwicklung einer optimalen Methode zur Zell-Klassifikation bedarf es also zunächst einer Untersuchung der Standard-Verarbeitungsschritte der massenspektrometrischen *Fingerprints*, da diese die Grundlage für nachfolgende Analyse darstellen.

### 1.3.1 Vorverarbeitung

Nach der Messung liegen die Spektren im firmenspezifischen Datenformat vor. Sie können dann mit der mitgelieferten Software eingelesen, bearbeitet und analysiert werden. Ist jedoch eine Auswertung der Spektren mit einem externen Analyse-Tool notwendig, müssen die Spektren in das entsprechende Datenformat konvertiert werden. Dies erfordert einen hohen Zeitaufwand, wenn die einzelnen Spektren zum Beispiel in Excel-Tabellen übertragen werden müssen (Hanrieder et al., 2011). Für eine schnelle Klassifikation wäre es daher hilfreich, ein Datenformat zur definieren, welches eine einfache Handhabung der Spektren ermöglicht und für die Klassifikation wichtige Informationen wie Tumortyp und Subtyp bereitstellt.

Häufig weisen spektrale Daten ein chemisches Hintergrundrauschen auf. Dieses erfordert eine Basislinienkorrektur, um die Grundlinie auf null zu verschieben. Die Software *ClinProTools* der Firma Bruker Daltonik zum Beispiel bietet zur Korrektur von Massenspektren zwei Möglichkeiten an: *Convex Hull Baseline* und *Top Hat Baseline*. Letztere wurde schon oft zur Vorverarbeitung von *Fingerprints* verwendet (Munteanu et al., 2012; Ouedraogo et al., 2010; Schwamb et al., 2013). Des Weiteren bietet die *Bioinformatics Toolbox* von MATLAB die Funktion *msbackadj* zur Basislinienkorrektur an, welche ebenfalls schon zur Vorverarbeitung von *Fingerprints* eingesetzt wurde (Povey et al., 2014). Die Basislinienkorrektur als standardmäßiger Schritt in der Vorverarbeitung von Massenspektren ist eine bereits etablierte Methode und bedarf keiner weiteren Optimierung.

Rauschen kann jedoch nicht nur chemischen, sondern auch instrumentellen Ursprungs sein. Hierbei kann durch einen Filter zur Glättung der Spektren das Signal-Rausch-Verhältnis (engl. *Signal to Noise Ratio*, SNR) noch weiter verbessert werden. Dafür eignet sich der Savitzky-Golay-Filter (Savitzky und Golay, 1964) welcher im Wesentlichen eine polynomielle Regression um den zu glättenden Datenpunkt durchführt. Der Hauptvorteil dieser Methode ist die Erhaltung wichtiger, spektraler Merkmale wie lokale Maxima / Minima und Impulsbreite. Dieser Algorithmus wird von der Software *ClinProTools* zur Verfügung gestellt und wurde schon sehr oft zur Rauschunterdrückung von *Fingerprints* eingesetzt (Munteanu et al., 2012; Ouedraogo et al., 2010; Schwamb et al., 2013). MATLAB bietet die Funktionen

*sgolayfilt* (allgemein) aus der *Signal Processing Toolbox* und *mssgolay* (speziell für Massenspektren) aus der *Bioinformatics Toolbox* zur Signalglättung an. Letztere wurde bereits zur Signalglättung von *Fingerprints* eingesetzt (Povey et al., 2014). Die Verwendung des Savitzky Golay Filters zur Signalglättung von Massenspektren ist ebenfalls eine bereits etablierte Methode, welche keiner Optimierung mehr bedarf. Jedoch werden bei dieser Methode keine Datenpunkte gelöscht. Bereiche, welche nur aus Rauschen bestehen, sind nach wie vor in den Spektren enthalten und könnten weitere Analysen verfälschen. Aus diesem Grund wäre die Entwicklung einer Methode, welche diese Bereiche nachträglich noch aus den Spektren eliminiert, von Vorteil.

### 1.3.2 Datenreduktion

Eine optimale Zell-Klassifikation setzt eine Datenreduktion voraus, bei welcher die zur Unterscheidung relevanten Peaks erhalten bleiben. Die Herausforderung besteht hierbei darin, auch diejenigen relevanten *m/z*-Werte zu finden, welche nur kleine Unterschiede aufweisen. Des Weiteren könnte bei der systematischen Entwicklung einer optimalen Klassifikationsmethode eine große Anzahl von Signalwerten (dies ist vor allem bei Massenspektren der Fall) zu Problemen bei der Verwendung von rechenintensiven Algorithmen führen, sodass schnell die Grenze der Rechenkapazität des Computers erreicht wird.

Das derzeitige Standard-Verfahren zur Reduktion der Datenpunkte in einem Massenspektrum ist das sogenannte *Peak Picking* Verfahren, welches schon wiederholt in der Literatur diskutiert wurde. Dabei werden die Peaks anhand bestimmter Merkmale wie Peakhöhe, Peakbreite oder dem Signal-Rauchsch-Verhältnis ausgewählt. Kempka et al. (Kempka et al., 2004) entwickelten eine Methode, welche auf der Annahme basiert, dass die Peak-Form einer Gauß-Verteilung entspricht. Du et al. (Du et al., 2006) schlagen einen Ansatz vor, bei welcher kontinuierliche Wavelet-Transformation durchgeführt wird. Yang et al. (Yang et al., 2009) führten eine Studie durch, in welcher sie mehrere *Peak Picking* Verfahren miteinander verglichen. Das *Peak Picking* Verfahren wird im Softwarepaket *ClinProTools* standardmäßig zur Datenreduktion eingesetzt. Die Peak-Analyse basiert dort auf den Mittelwertspektren der Klassen oder dem Mittelwertspektrum des gesamten Datensatzes. Es wurde schon oft zur Datenreduktion von *Fingerprints* eingesetzt (Munteanu et al., 2012; Ouedraogo et al., 2010; Schwamb et al., 2013).

Das *Peak Picking* Verfahren gibt jedoch keine Auskunft über die statistische Relevanz der Peaks, das heißt, wie gut sich die gefunden Peaks für Zell-Klassifikation eigenen. Auch trifft es keine Aussage darüber, welche und wie viele der detektierten Peaks für eine optimale Zell-Klassifikation benötigt werden. Des Weiteren können eventuell wichtige, sehr kleine

Peaks übersehen werden. Werden die Peaks im Mittelwertspektrum des Datensatzes gesucht, kann es passieren, dass manche Peaks durch die Mittelung verloren gehen. Daher bedarf es einer Untersuchung, ob sich andere, statistische Merkmalsextraktionsmethoden besser für die Zell-Klassifikation eignen.

Ein weiteres Verfahren, welches schon oft zur Datenreduktion von Fingerprints eingesetzt wurde (Chiu et al., 2015; Feng et al., 2010; Sakai et al., 2015), ist die Hauptkomponentenanalyse (engl. *Principal Component Analysis*, PCA). Sie ist statistisches Verfahren, welches zur Strukturierung und Vereinfachung hochdimensionaler Merkmalsräume eingesetzt wird (Pearson, 1901). Das Ziel der PCA ist es, die Dimension oder das Rauschen eines Datensatzes zu reduzieren, ohne dabei die in den Daten enthaltene Information zu verlieren. Datensätze mit sehr vielen Variablen zeigen oft ähnliches Verhalten und besitzen redundante Information. Der mehrdimensionale Datenraum wird bei der PCA so gedreht, dass die Richtung der größten Varianz die erste Koordinatenachse des projizierten Datenraums wird. Die resultierenden neuen Achsen (Hauptkomponenten) stehen orthogonal aufeinander und sind nach absteigender Varianz sortiert. In den meisten Fällen werden nur die ersten Hauptkomponenten betrachtet, da sie den Großteil der Varianz besitzen. Die restlichen Hauptkomponenten bestehen hauptsächlich aus Rauschen. Ein großer Nachteil dieser Methode ist, dass nur schwer Rückschlüsse auf die zur Unterscheidung wichtigen *m/z*-Werte geschlossen werden können, da eine Hauptkomponente aus der Linearkombination *aller* ursprünglichen *m/z*-Werte besteht. Für eine optimale Zell-Klassifikation würde sich daher besser eine Methode eignen, bei welcher jedem *m/z*-Wert eindeutig ein Gewicht zugeordnet werden kann.

Oft wird die PCA auch gleichzeitig als Klassifikationsmethode eingesetzt, indem die Hauptkomponenten als Punktwolken in einem Schaubild dargestellt werden. Die Beurteilung des Erfolgs der Klassifikation erfolgt dann durch die visuelle Betrachtung des Schaubilds. Für eine sichere Klassifikation wäre jedoch eine objektive Kennzahl für die Bewertung besser geeignet.

Beim *Resampling* wird die Anzahl der Datenpunkte in den Massenspektren um einen bestimmten Faktor verringert. In der *Bioinformatics Toolbox* von MATLAB ist dazu die Funktion *msresample* enthalten, welche schon für die Datenreduktion von *Fingerprints* eingesetzt wurde (Povey et al., 2014). Dieses Verfahren ist jedoch mit einem Informationsverlust verbunden und gibt keine Auskunft über die Signifikanz der Peaks. Daher eignet sich nicht für die Entwicklung einer optimalen Klassifikationsmethode.

Auswahl der Peaks

### 1.3.3 Auswahl der Peaks

Mit Hilfe einer Varianzanalyse (engl. *Analysis of Variance*, ANOVA) kann geprüft werden, ob die Unterschiede der Peaks (Mittelwerte) zwischen den Spektren der verschiedenen Zelllinien signifikant oder ob sie so klein sind, dass sie wahrscheinlich durch zufälliges Rauschen entstanden sind (Hartung und Elpelt, 2007). Ist dies der Fall, kann mit Hilfe eines sogenannten *Post-Hoc*-Tests festgestellt werden, welche dieser Peaks sich unterscheiden (Hochberg und Tamhane, 2009). Häufig wird bei der ANOVA ein Signifikanzniveau (Alpha-Wert) von 5% verwendet. Dieses sagt aus, dass der Unterschied zwischen den Zelllinien zu 95% signifikant ist. Der sogenannte p-Wert (Prüfgröße) muss also kleiner oder gleich dem Signifikanzniveau sein.

In ihrer einfachsten Form stellt die ANOVA eine Alternative zum T-Test (Student, 1908) dar. Der Nachteil des T-Tests ist, dass dieser immer nur zwei Stichproben / Zelllinien miteinander vergleichen kann. Mit Hilfe der ANOVA kann er jedoch zur Untersuchung von mehr als nur zwei Stichproben / Zelllinien erweitert werden.

Marvin-Guy et al. (Marvin-Guy et al., 2008) führten eine einfache ANOVA zur Bestimmung der Signifikanz der Unterschiede von Epithelzellen eines Darmkarzinoms durch. Als *Post-Hoc*-Test wurde hier der *Fisher-LSD* (engl. *Least Significant Difference*) verwendet. Dabei handelt es sich um einen paarweise T-Test zwischen allen Gruppen / Zelllinien, wobei jedoch immer die gesamte Varianz aller Zelllinien und nicht nur die Varianz der jeweils zwei beteiligten Zelllinien verwendet wird. Es gilt der gleiche Alpha-Wert für alle Vergleiche, wodurch die Signifikanz erschwert wird. Trotzdem könnten hier zu viele Signifikanzen detektiert werden.

Harnrieder et al. (Hanrieder et al., 2011) führten eine einfache ANOVA zur Unterscheidung von Gliazellen durch. Im Anschluss daran wurde der *Tukey-HSD*-Test (engl. *Honestly Significant Difference*) verwendet. Er ist ähnlich dem *Fisher-LSD*-Test, hier wird jedoch ein höheres Signifikanzniveau festgelegt.

Ein weiteres Verfahren der ANOVA ist der parameterfreie Wilcoxon-Vorzeichen-Rang-Test (Mann und Whitney, 1947; Wilcoxon, 1945). Dieser beruht auf den Differenzen der Wertepaare (Intensitäten des gleichen *m/z*-Wertes zweier Zelllinien), wobei das Vorzeichen der Differenz berücksichtigt wird. Er wurde von Portevin et al. (Portevin et al., 2015) zur Unterscheidung von Immunzellen eingesetzt. Im Anschluss daran wurde eine Bonferroni-Korrektur (Bonferroni, 1936) durchgeführt, welche zu viele falsch-positiv Ergebnisse vermeiden soll. Bei diesem wird der Alpha-Wert durch die Anzahl der durchgeführten Experimente geteilt.

In allen drei Fällen wurden alle Peaks für die Zell-Klassifikation verwendet, welche nach dem *Post-Hoc*-Test einen bestimmten p-Wert erreichten, welche also mit einer bestimmten Wahrscheinlichkeit für die Unterschiede der Zelllinien verantwortlich sind. Für eine sichere Zell-Klassifikastion wäre jedoch eine Bewertung der Peaks anhand von Gewichten und nicht basierend auf Wahrscheinlichkeiten besser geeignet, um grundsätzlich das Auftreten von falsch-positiv Ergebnissen zu umgehen. Des Weiteren könnte das zuvor durchgeführte *Peak-Picking* eventuell schon im Vorfeld zu einem Informationsverlust geführt haben. Es existieren zahleiche Methoden zur Merkmalsextraktion, welche eine Beurteilung der Peaks anhand von Gewichten ermöglichen und noch nicht alle auf *Fingerprints* angewandt wurden. Eine systematische Untersuchung dieser Methoden für die Optimierung der Zelllinien-Klassifikation wurde bisher jedoch noch nicht durchgeführt.

Eine Möglichkeit zur Bestimmung einer geeigneten Peak-Anzahl und Peak-Kombination für eine gute Klassifikation, ist die Anwendung von genetischen Algorithmen (engl *Genetic Algorithms*, GA). Diese sind heuristische Optimierungsverfahren und beruhen auf Methoden und Erkenntnissen der biologischen Genetik. Der grundlegende Steuerungsmechanismus dabei ist: Mutation, Selektion und Rekombination. Sie werden bei sehr großen und komplexen Datenmengen angewendet, wenn es sonst zu keinen brauchbaren Ergebnissen kommen oder die Berechnung zu viel Zeit beanspruchen würde. Dabei werden die Parameter einer Gleichung oder eines anderen strukturierten Lösungsansatzes optimiert (Goldberg und Holland, 1988; Holland, 1975). Schwamb et al. (Schwamb et al., 2013) benutzen diesen Algorithmus, um die beste Peak Kombination zur Identifikation von Signaturen als Indikator für Zellstress und Apoptose anhand von *Fingerprints* zu erhalten. Allerdings kann es vorkommen, dass genetische Algorithmen nicht die optimale Lösung finden, wenn die eingestellten Startparameter zu weit von der besten Endlösung entfernt sind. Des Weiteren kann dieser Algorithmus nicht angewandt werden, wenn zuvor kein *Peak Picking* stattgefunden hat, da er aufgrund der enorm großen Anzahl an Datenpunkten in einem Massenspektrum wahrscheinlich nicht in annehmbarer Zeit durchführbar ist. Daher bedarf es zum einen der Entwicklung einer Methode, welche auch in großen Datenmengen Peak-Kombinationen selektiert, und zum anderen der Untersuchung der Leistung und Rechenzeit in Hinblick auf eine schnelle Zelllinien-Klassifikation.

### 1.3.4 Spektrale Rekalibration

Durch Fehler bei der Kalibrierung oder durch Unschärfen während der Spektrengenerierung kann es zu systematischen Verschiebungen der Massenachse kommen. Dies kann zur Folge haben, dass ein Protein in verschiedenen Spektren nicht exakt die gleichen *m/z*-Werte aufweist. Zudem kann die Auflösung eines Peaks bei verschiedenen Experimenten (in Abhängigkeit zum *m/z*-Wert) variieren. Aus diesen Gründen ist es notwendig, eine

Rekalibration durchzuführen, so dass am Ende alle Spektren eine einzige gemeinsame Massenachse besitzen. Die *Fingerprints* müssen, wie bereits erwähnt, in Form einer Matrix für die nachfolgenden Analysen zur Verfügung stehen. Die Software *ClinProTools* rekalibriert die Spektren standardmäßig mit Hilfe herausragender Peaks im Anschluss an das *Peak Picking* Verfahren (Munteanu et al., 2012; Ouedraogo et al., 2012; Schwamb et al., 2013). Für die systematische Untersuchung anderer Peak-Auswahl-Methoden muss die Ausrichtung der Massenache jedoch auf den unreduzierten Daten erfolgen. Gobom et al. (Gobom et al., 2002) entwickelten zwar eine Methode zur Rekalibrierung, welche ein Polynom höheren Grades auf mehrere im gesamten Massenbereich verteilte *m/z*-Werte legt, doch werden hierfür einige Referenz-Peaks benötigt, welche wiederum ein gewisses Fachwissen voraussetzen. Daher bedarf es der Entwicklung einer Methode zur Berechnung einer gemeinsamen Massenachse und zur Ausrichtung der Spektren, welche sich einerseits schnell und einfach durchführen lässt, und andererseits unabhängig von bestimmten Kalibrations-Peaks sind, so dass sie sich auf alle Zelllinien und biomedizinischen Fragestellungen anwenden lässt.

### 1.3.5 Normierung

Mit einer Normierung werden die unterschiedlichen Massenspektren für die nachfolgende Analyse vergleichbar gemacht. Im Softwarepaket *ClinProTools* werden die Spektren einer TIC-Normierung unterzogen. Als Totalionenstrom (engl. *Total Ion Current*, TIC) wird die Summe der Ströme bezeichnet, welche von den Ionen aller *m/z*-Werte im Spektrum erzeugt wird. Dies ist die Standardmethode, welche häufig für die Normierung von *Fingerprints* eingesetzt wird (Munteanu et al., 2012; Ouedraogo et al., 2012; Schwamb et al., 2013). Die *Bioinformatics Toolbox* von MATLAB stellt die Funktion *msnorm* bereit, welche die Fläche unter der Kurve eines Spektrums auf den Gruppenmedian normiert. Auch sie wurde bereits für die Normierung von *Fingerprints* verwendet (Povey et al., 2014). Diese beiden Normierungsmethoden sind bereits etabliert, weshalb hier keine Optimierung mehr nötig ist.

### 1.3.6 Klassifikation

Die Klassifikation ist eine Vorhersagemethode des maschinellen Lernens, bei der Objekte anhand ihrer Merkmale verschiedenen Klassen zugeordnet werden. Der Ausgabewert ist ein diskreter Wert. In der Medizin wird die Klassifikation dazu benutzt, um eine Diagnose zu erstellen. Die Merkmale sind in diesem Fall die Symptome einer Krankheit, oder bei Verwendung der Massenspektrometrie bestimmte Signale, welche charakteristisch für den Krebszelltyp sind.

Bei der überwachten Klassifikation werden die Trainingsdaten zusammen mit ihren zugeordneten Ausgabewerten dem Klassifikator übergeben. Diese Art der Klassifikation

findet ihre Anwendung zum Beispiel bei Existenz einer Datenbank, in welcher Referenzmessungen enthalten sind und eine unbekannte Probe einer Krebsart oder Mutation zugeordnet werden sollen (Avila et al., 2016). Üblicherweise werden bei der überwachten Klassifikation die Daten in Trainings- und Testdaten aufgeteilt, wobei mit Hilfe der Trainingsdaten ein Modell erstellt wird, mit welchem die Testdaten im Anschluss klassifiziert werden.

Ein überwachtes Klassifikationsverfahren, welches schon für die Analyse von Zelllinien anhand von *Fingerprints* eingesetzt wurde (Schwamb et al., 2013), ist der k-Nächste-Nachbarn Algorithmus (eng. *k-Nearest Neighbours*, KNN). Hier wird für jeden Datenpunkt der Abstand zu seinen k nächsten Nachbarn berechnet. Als Distanzmaß wird hierbei oft die Euklidische Distanz verwendet. Die Entscheidung fällt dann durch einen Mehrheitsbeschluss (Altman, 1992). Der Nachteil von überwachten Klassifikationsverfahren ist, dass relativ zur Anzahl der Merkmale eine sehr große Anzahl an Proben benötigt wird, um ein „auswendig lernen“ des Klassifikators zu verhindern. Dies ist vor allem bei Massenspektren besonders kritisch, da diese aus enorm vielen Datenpunkten bestehen. Hier zeigt sich noch einmal die Notwendigkeit einer effektiven Datenreduktion der MS-*Fingerprints* für eine sichere Zelllinien-Klassifikation, welche allein durch das *Peak Picking* Verfahren nicht sichergestellt ist.

Bei der unüberwachten Klassifikation werden die Trainingsdaten ohne die zugehörigen Ausgabewerten dem Klassifikator übergeben. Die einzige Information, welche dem Klassifikator mitgegeben werden kann, ist die Anzahl der Klassen, in welche er den Trainingsdatensatz unterteilen soll. Für die unüberwachte Klassifizierung werden typischerweise sogenannte Clustering-Verfahren eingesetzt. Diese partitionieren die Datenmenge in mehrere Cluster, so dass sich ähnliche Daten im gleichen Cluster befinden. Für den klinischen Betrieb kann diese Variante der Klassifizierung von Nutzen sein, um zu prüfen, ob der Pathologe sich eventuell bei der Beurteilung einer Probe geirrt hat oder ob eine Probe von vorne herein falsch zugeordnet wurde.

Ein unüberwachtes Klassifikationsverfahren, welches schon sehr oft zur Klassifikation von Zelllinien anhand von *Fingerprints* verwendet wurde (Karger et al., 2010; Munteanu et al., 2012; Ouedraogo et al., 2012), ist die hierarchische Clusteranalyse (engl. *Hierarchical Cluster Analysis*, HCA). Die HCA ist ein distanzbasiertes Verfahren zur Clusteranalyse, bei dem die Objekte, welche zueinander eine geringe Distanz aufweisen, zu einem Cluster zusammengefasst werden (Rokach und Maimon, 2005). Dabei können verschiedene Distanzmaße verwendet werden. Auch hier wird häufig die Euklidische Distanz verwendet. Das Ergebnis ist eine hierarchische Struktur, welche graphisch durch ein zweidimensionales Diagramm, dem sogenannten Dendrogramm, dargestellt wird. Das Dendrogramm ist eine spezielle Baumstruktur, deren Knoten jeweils ein Cluster darstellen. Die Wurzel repräsentiert

ein einziges Cluster, welches die gesamten Datenobjekte beinhaltet. Ein innerer Knoten entspricht der Vereinigung aller seiner Kinderknoten. Eine Kante zwischen zwei Knoten stellt die Distanz zwischen den beiden verbundenen Clustern dar.

Bei der HCA wird grob zwischen zwei Verfahren unterschieden. Zum einen gibt es die divisiven Clusterverfahren *(top-down)*, in welchen zunächst alle Objekte einem Cluster angehören und dann schrittweise in immer kleinere Cluster aufgeteilt werden, bis jedes Cluster nur noch aus einem Objekt besteht. Zum anderen gibt es die agglomerativen Clusterverfahren *(bottom-up)*, in welchen zunächst jedes Objekt ein Cluster bildet, welche dann schrittweise zu immer größeren Clustern zusammengefasst werden, bis alle Objekte zu einem einzigen Cluster gehören. Des Weiteren können verschiedene Algorithmen, welche die Cluster zusammenfasst bzw. aufteilt, verwendet werden, wie zum Beispiel das *Single-Linkage*-Verfahren, welches den kürzesten Abstand zwischen zwei Punkten wählt. Die HCA bietet ein hohes Maß an Flexibilität, da auch komplexere Distanzmaße verwendet werden können. Ein Vorteil ist, dass sie außer der Distanzfunktion keine eigenen Parameter besitzt. Durch die Baumstruktur ist keine Festlegung auf eine bestimmte Clusterzahl notwendig, da die Cluster-Hierarchie auch Unterstrukturen erlaubt. Die Beurteilung des Erfolgs der Zelllinien-Klassifikation geschieht meist durch die visuelle Betrachtung des Dendrogramms. Dies kann jedoch bei einer großen Anzahl an Zelllinien und Proben sehr schnell unübersichtlich werden.

Eine systematische Untersuchung verschiedener Methoden zur Merkmalsextraktion in Kombination mit verschiedenen Klassifikationsmethoden für eine optimale Zelllinien-Klassifikation anhand von *Fingerprints* wurde bisher noch nicht durchgeführt. Dies wäre jedoch notwendig, um die optimale Methode zur Zelllinien-Klassifikation zu finden, mit welcher schneller als bisher und trotzdem sicher klassifiziert werden kann. Vor allem sind die bisherigen Methoden aus der Signalverarbeitung für *Fingerprints* noch nicht auf Geschwindigkeit optimiert worden, was in der Praxis zu Problemen führen könnte.

### 1.3.7 Objektive Bewertungskennzahlen

Zur Bewertung eines Klassifikationsmodells wird ein objektives Kriterium benötigt. Im Falle einer überwachten Klassifikation, bei welcher die Klassenzugehörigkeit im Vorfeld bekannt ist, wird das Ergebnis durch die Klassifikationsrate beurteilt. Sie gibt das Verhältnis der richtig klassifizierten Objekte zur Gesamtzahl der Objekte im Datensatz an. Bei der unüberwachten Klassifikation wird im Anschluss eine Clusteranalyse durchgeführt (Jain und Dubes, 1988). Dabei geht es darum zu beurteilen, wie gut das Clustering funktioniert hat. Ein Clusteralgorithmus teilt die Daten in eine meist vorgegebene Anzahl von Gruppen, ohne zu wissen, zu welcher Klasse sie tatsächlich gehören.

Ist die Klassenzugehörigkeit in der Realität jedoch bekannt, kann der verwendete Algorithmus mit Hilfe von sogenannten *External Indicies* evaluiert werden. Zwei bekannte Kennzahlen aus der Literatur, mit welchen eine externe Clusteranalyse durchgeführt werden kann, ist die Reinheit (engl *purity*) und die Entropie (Deepa et al., 2012; Rendón et al., 2011; Sripada und Rao, 2011). Die Reinheit bestimmt für ein Cluster, inwiefern nur Objekte einer Klasse enthalten sind. Die Entropie berücksichtigt die Verteilung der Klassen in einem einzelnen Cluster. Die Reinheit bzw. die Entropie des Clusterergebnisses ist dann das gewichtete Mittel der Reinheitswerte bzw. der Entropien aller Cluster. Kremer et al. (Kremer et al., 2011) stellen eine Übersicht über weitere Bewertungskennzahlen dar. Der Nachteil dieser *External Indices* ist, dass sie im Vergleich zur Klassifikationsrate keine exakte Aussage über den Erfolg der Klassifikation liefern.

Eine externe Clusteranalyse mit Hilfe von Kennzahlen zur Beurteilung des Klassifikationsergebnisses bei der Unterscheidung von Zelllinien wurde bisher noch nicht durchgeführt. Karger et al. (Karger et al., 2010) zum Beispiel verwendeten eine HCA zur Klassifikation von 66 stabilen Zelllinien. Die Auswertung erfolgte jedoch durch eine visuelle Betrachtung des Dendrogramms.

Für die systematische Untersuchung verschiedener Methoden zur Merkmalsextraktion und Klassifikation (unüberwacht), bei welcher eine Vielzahl an Kombinationen getestet und miteinander verglichen werden sollen, muss daher zunächst eine Kennzahl entwickelt werden, welche sowohl eine möglichst exakte Aussage über die Qualität der Ergebnisse liefert (sichere Klassifikation), als auch die Rechenzeit und den Speicherbedarf optimiert (schnelle Klassifikation).

## 1.4 Bestimmungen von Wirkstoffwirkungen

### 1.4.1 Pharmakodynamik

Die Pharmakodynamik beschäftigt sich mit der Wirkung von Arzneimitteln auf den Körper (Hollinger, 2007). Dabei zeigt das Wirkprofil, welche Effekte auftreten und welche Organe oder biologische Funktionen beeinflusst werden. Spezifische Wirkstoffe sind weitgehend von der molekularen Struktur abhängig und entfalten ihre Wirksamkeit an einem bestimmten Ort. Sie wirken auf körpereigene Strukturen, wie zum Beispiel Rezeptoren oder Enzyme, und es reichen meist niedrige Konzentrationen, um eine Wirkung hervorzurufen. Es wird von einem sogenannten „Schlüssel-Schloss-Prinzip“ gesprochen. Andere Wirkstoffe hingegen wirken unspezifisch. Sie sind eher in hohen Konzentrationen wirksam und verteilen sich im gesamten Organismus. Moderne molekular-gerichtete Therapeutika weisen häufig eine

verbesserte Wirksamkeit bei günstigerem Nebenwirkungsprofil auf als konventionelle Chemotherapeutika. Sie stehen daher im Fokus der klinischen Forschung und der Entwicklung einer individualisierten Medizin. Biomarker, oder allgemein molekulare Signaturen, sind dabei von besonderer Bedeutung (Ebert et al., 2012). Prädiktive Biomarker ermöglichen es, das Ansprechen eines individuellen Patienten auf spezielle Wirkstoffe vorherzusagen. Dabei wird nach Zielstrukturen für Arzneistoffe gesucht, welche vorliegen müssen, wenn die Therapie mit einem bestimmten Therapeutikum sinnvoll sein soll. Pharmakodynamische Biomarker erlauben es, die Wirksamkeit der molekular-gerichteten Therapeutika anzuzeigen. Es können Variationen erkannt werden, um gegebenenfalls die Wirkstoffdosis individuell anzupassen. Massenspektrometrische Methoden haben bereits wesentlich zur Entdeckung molekularer *Targets* bzw. Therapeutika beigetragen (Bantscheff et al., 2011; Dawson et al., 2011; Kruse et al., 2008; Ramsden et al., 2011).

### 1.4.2 Konzentrations-Wirkungs-Beziehung

Eine zentrale Bedeutung in der Pharmakodynamik besitzt die Dosis-Wirkungs-Beziehung bzw. die Konzentrations-Wirkungs-Beziehung, da durch diese Aussagen über die Wirksamkeit und Sicherheit von Arzneimitteln getroffen werden können. Sie beschreibt graphisch den Zusammenhang zwischen der verabreichten Dosis / Konzentration eines Medikaments und seiner Wirkung. Es kann abgelesen werden, welche Konzentrationen eine Wirkung verursachen, wie stark die Effekte in Abhängigkeit der Konzentration sind und ab wann toxische Effekte auftreten. Die meisten Wirkstoffe weisen allerdings keine lineare Beziehung auf. Eine doppelte Konzentration zum Beispiel verursacht nicht zwingend einen doppelt so großen Effekt. Im Idealfall ergibt sich eine sigmoide Kurve. Anhand der Steigung der Kurve können die Auswirkungen der Konzentrationsänderungen auf die Wirkung des Medikaments erkannt werden. Als mittlere effektive Konzentration (EC50) wird die Konzentration bezeichnet, bei welcher ein halbmaximaler Effekt beobachtet wird. Dies entspricht dem Wendepunkt der Kurve. Der sogenannte *Fold Change* gibt das Verhältnis der Wirkung zwischen der minimalen und maximalen Wirkstoffkonzentration an.

Um eine Konzentrations-Wirkungs-Kurve zu erstellen, kann zum Beispiel die Statistiksoftware *GraphPad Prism* (Motulsky und Christopoulos, 2004) verwendet werden. Sie besitzt ein umfassendes Repertoire zum Thema nichtlineare Regression und Kurvenanpassung. Im Rahmen der MS-basierten *Fingerprint*-Analyse wurde sie bereits eingesetzt, um die Zellaktivität mit Imatinib behandelter Zellen anhand von Histon-Deacetylasen (HDAC) Inhibitoren zu messen (Munteanu et al., 2014). Des Weiteren wurden toxische Effekte in Zell-basierten ökotoxischen Testsystemen untersucht (Kober et al., 2015). Dabei wurden jeweils einzelne Konzentrations-Wirkungs-Kurven erstellt und deren EC50-

Werte berechnet. Eine Methode zur automatischen Auswertung von MS-*Fingerprints* für die Wirkstoff-Profilierung wurde bisher jedoch noch nicht entwickelt.

# 2 ZIELSETZUNG

Moderne molekular-gerichtete Therapeutika weisen häufig eine verbesserte Wirksamkeit bei günstigerem Nebenwirkungsprofil auf als konventionelle Chemotherapeutika. Sie stehen daher im Fokus der klinischen Forschung und der Entwicklung einer individualisierten Medizin. Biomarker oder allgemein molekulare Signaturen, welche die Wirksamkeit molekular-gerichteter Therapeutika anzeigen oder vorhersagen, sind dabei von besonderer Bedeutung. Die Grundlage dafür besteht in der schnellen und zuverlässigen Klassifizierung humaner Tumorproben. Massenspektrometrische Methoden haben bereits wesentlich zur Entdeckung molekularer Therapeutika beigetragen. Diese generieren jedoch Datensätze im Umfang von mehreren Gigabytes, deren Bearbeitung und Auswertung nur mit Hilfe moderner Signalverarbeitung möglich ist.

Dieses Dissertationsprojekt verfolgte daher folgende Ziele:

1. Überführung der MS-Daten in ein MATLAB-kompatibles Datenformat sowie die Optimierung der Vorverarbeitung von massenspektrometrischen *Fingerprints*
2. Entwicklung einer geeigneten Kennzahl zur Bewertung der systematischen Untersuchung verschiedener Merkmalsextraktions- und Klassifikationsmethoden
3. Entwicklung und Optimierung einer Methodik für die Merkmalsextraktion zur Identifizierung geeigneter Biomarker-Kandidaten
4. Entwicklung eines auf spektralen, molekularen *Fingerprints* basierenden Verfahrens zur schnellen Klassifikation von Tumorsubtypen anhand von der in 2 entwickelten Kennzahl
5. Entwicklung einer Methode zur automatisierten Erstellung von Konzentrations-Wirkungskurven zur Bewertung der zellulären Wirkung von zielgerichteten Medikamenten

Es sollte eine Datenstruktur erstellt werden, welches einen schnellen und einfachen Zugriff auf die MS-Daten ermöglicht. Bereits bestehende Methoden zur Vorverarbeitung massenspektrometrischer *Fingerprints* sollten optimiert und erweitert werden. Zur systematischen Untersuchung verschiedener Merkmalsextraktions- und Klassifikationsmethoden sollte eine Kennzahl entwickelt werden, welche genaue Ergebnisse liefert und eine Durchführung in annehmbarer Zeit ermöglicht. Anhand geeigneter Datensätze, (Brustkrebs-, Leukämie- und GIST-Zelllinien) sollte ein Verfahren zur schnellen Klassifikation von Tumortypen und -subtypen entwickelt werden. Des Weiteren sollte eine Methode entwickelt werden, welche automatisch in MS-Datensätzen nach Massenbereichen sucht,

welche eine Reaktion auf zielgerichtete Wirkstoffe zeigen. Die Algorithmen sollten in MATLAB implementiert werden.

# 3 MATERIAL UND METHODEN

## 3.1 Zelllinien

Folgende Tabellen geben eine Übersicht der Zelllinien, welche in diesem Dissertationsprojekt zur informationstechnischen Methodenentwicklung verwendet wurden. Die Zellkultur sowie die MS-Messungen wurden im Institut für Instrumentelle Analytik und Bioanalytik an der Hochschule Mannheim von Jan-Hinrich Rabe, Dr. Carolina v. Reitzenstein und Dr. Bogdan Munteanu durchgeführt.

**Tabelle 2: Brustkrebs-Zelllinien und Tumor-Subklassifizierung (Zellkultur und MS-Messungen: Jan-Hinrich Rabe)**

| Zelllinie | Tumor-Subklassifizierung |
|---|---|
| Cal51 | Triple negativ |
| EFM192A | Luminal B (HER2 positiv) |
| MCF7 | Luminal A |
| MDA-MB-453 | keine eindeutige Klassifizierung |
| MDA-MB-468 | Triple negativ |
| SKBr3 | HER2 positiv |

**Tabelle 3: GIST-Zelllinien, Mutation und Imatinib-Wirkung (Consolino et al., 2016) (Zellkultur und MS-Messungen: Jan-Hinrich Rabe)**

| Zelllinie | Mutation | Imatinib Wirkung |
|---|---|---|
| 430 | Kit Exon 11 und Exon 13 V654A | resistent |
| 882 | Kit Exon 13 K642E | responsiv |
| T1 | Kit Exon 11 57bp Deletion | sensitiv |

**Tabelle 4: Leukämie-Zelllinien und Form der Erkrankung (Zellkultur: Dr. Carolina von Reitzenstein, MS-Messungen : Dr. Bogdan Munteanu)**

| Zelllinie | Form der Erkrankung |
|---|---|
| K562 | Chronische myeloische Leukämie (CML) |
| MV4-11 | Akute myeloische Leukämie (AML) |

### 3.1.1 Wirkstoffe

Folgende Tabelle gibt eine Übersicht über die Wirkstoffe, welche zur Entwicklung eines Verfahrens zur Bewertung der zellulären Wirkung durch Dr. Bogdan Munteanu verwendet wurden. In dieser Arbeit wurden hierzu die informationstechnischen Methoden entwickelt.

**Tabelle 5: Wirkstoffe**

| Wirkstoff | Handelsname | Gruppe der | Einsatz |
|---|---|---|---|
| Axitinib | Inlyta® | Tyrosinkinase-Inhibitoren | Nierenzellkarzinome |
| Bortezomib | Velcade® | Proteasom-Inhibitoren | multiple Myelome, Mantelzell-Lymphom |
| Bosutinib | Bosulif® | Kinase-Inhibitoren (BCR-ABL-Kinase) | CML |
| Chloroquin | Resochin® | Malariamittel | Malaria, Rheuma |
| Dasatinib | Sprycel® | Tyrosinkinase-Inhibitoren | ALL, CML |
| Ethanol | | | Zellgift |
| Imatinib | Glivec® | Tyrosinkinase-Inhibitoren | CML, GIST |
| Obatoclax | noch nicht zugelassen | Bcl-2-Inhibitoren | solide Tumore, multiple Myelome, Leukämie, Lymphon |
| Tamoxifen | Nolvadex® | Antiöstrogene | Brustkrebs |
| Paclitaxel (Taxol) | Taxol® | Taxane | Verschiedene Krebsarten (z.B. Brustkrebs, Pankreas) |

### 3.1.2 MS-Messungen

Die MS-Messungen der Zelllinien wurden mit einem *Autoflex Speed* MALDI-TOF/TOF Massenspektrometer (Bruker Daltonics, Bremen) von Mitarbeitern des Instituts für Instrumentelle Analytik und Bioanalytik, Jan-Hinrich Rabe und Dr. Bogdan Munteanu an der Hochschule Mannheim durchgeführt. Probenpräparation, Matrix und MALDI-Grund-einstellungen und Messung wurden mit einer an der Hochschule Mannheim entwickelten *Biotyping*-Methode (Munteanu et al., 2012) durchgeführt. Zur Rekalibrierung wurde die Software *flexAnalysis* (Bruker Daltronics) verwendet.

## 3.2 MATLAB

Alle die in diesem Dissertationsprojekt entwickelten Methoden und Algorithmen wurden in MATLAB R2016a (The MathWorks, Inc.) erstellt.

Zur Vorverarbeitung und Klassifikation der massenspektrometrischen *Fingerprints* wurden die MATLAB Funktionen aus den folgenden Tabellen verwendet. Diese stellen bereits einen verwendbaren Algorithmus für eine Aufgabenstellung dar oder wurden für eine Teilberechnung in einem der in diesem Dissertationsprojekt entwickelten Methoden und Algorithmen und verwendet.

**Tabelle 6: MATLAB Funktionen zur Vorverarbeitung der Massenspektren**

| Funktion | Toolbox | Anwendung |
|---|---|---|
| msbackadj | Bioinformatics Toolbox | Basislinienkorrektur |
| sgolayfilt | Signal Processing Toolbox | Savitzky Golay Filter |
| interp1 | Mathematics Toolbox | Lineare Interpolation |
| xcorr | Signal Processing Toolbox | Berechnung der Korrelationsfunktion |
| mapstd | Neural Network Toolbox | Z-Score Normierung |
| hist | Graphics Toolbox | Erstellung eines Histogamm |
| mspeaks | Bioinformatics Toolbox | Peak Picking |

**Tabelle 7: MATLAB Funktionen zur Klassifikation der Massenspektren aus der *Statistics and Machine Learning Toolbox***

| Funktion | Anwendung |
|---|---|
| pca | Hauptkomponentenanalyse |
| pdist | Hierarchische Clusteranalyse (Berechnung Distanz) |
| linkage | Hierarchische Clusteranalyse (Verknüpfung) |
| cluster | Hierarchische Clusteranalyse (Cluster-Bildung) |
| kmeans | k-Means Klassifikation |
| kmedoids | k-Medoids Klassifikation |
| fitcdiscr | Lineare Diskriminanzanalyse |
| fitcknn | k-Nächste-Nachbarn Klassifikation |
| predict | Überwachte Klassifikation |

## 3.3 Methoden zur Merkmalsextraktion

Für die systematische Untersuchung verschiedener Merkmalsextraktionsmethoden zur Datenreduktion von massenspektrometrischen *Fingerprints* wurden neben einer selbst implementierten Variante des Fisher-Testes noch weitere MATLAB Funktionen aus dem Softwarepaket der Arizona State University[1] verwendet:

***Fisher-Score***

Der Fisher-Test (Fisher, 1936) stellt eine einfache Möglichkeit zur Bewertung der Trennbarkeit von Klassen anhand ihrer Merkmale dar. Er ist für ein Merkmal und mehrere Klassen wie folgt definiert:

$$Fisher\ Score = \frac{\sum_{k=1}^{K} n_k * (m_k - M)^2}{\sum_{k=1}^{K} n_k * v_k}$$

Dabei ist:

*M*: Mittelwert eines Merkmals über den gesamten Datensatz

$m_k$: Mittelwert der Klasse *k*

$v_k$: Varianz der Klasse *k*

*K*: Anzahl Klassen

$n_k$: Anzahl Proben der Klasse *k*

***Bayes'sche Logistische Regression***

Das Bayes'sche logistische Regressionsverfahren (BLogReg) basiert auf dem Algorithmus von Shevade & Keerthi (Gilbert et al., 1988). Es dient zur Berechnung des Mindestabstandes von konvexen Mengen, so dass sich diese gerade noch berühren. Hierbei wird nicht nur eine Regressionsfunktion geliefert, sondern auch einen Konfidenzkorridor. Cawley und Talbot (Cawley und Talbot, 2006) verwenden die Bayes'sche Logistische Regression zur Klassifikation von Leukämie und Darmkrebs.

***Chi-Quadrat-Test***

Der Chi-Quadrat-Test wurde zu ersten Mal von Karl Pearson beschrieben (Pearson, 1992) und stellt eine der bekanntesten Wahrscheinlichkeitsverteilungen dar. Er wird für kategorische Variablen verwendet und beschreibt die Signifikanz der Unterschiede zwischen den beobachteten und den erwarteten Häufigkeiten. Liu und Setiono (Liu und Setiono, 1995) zeigen, dass der Chi-Quadrat-Test eine effektive Methode zur Merkmalsextraktion darstellt.

[1] Arizona State University (2015): Feature Selection Algorithms.
Online: http://featureselection.asu.edu/software.php, Stand: 21.12.2015

***Informationsgewinn***

Ein Entscheidungsbaum dient zur Klassifikation von Objekten (Cover und Thomas, 2006). Er besteht aus einer speziellen Datenstruktur, welche als Baumdiagramm grafisch dargestellt werden kann. Dabei entsprechen die inneren Knoten des Baumes den Merkmale einer Klasse und die Kanten den Werten der Merkmale. Zur Klassifikation eines Objektes werden vom Wurzelknoten aus (engl. *top down*) die Merkmalsknoten anhand ihrer Werte getestet. Der Informationsgewinn (engl. *Information Gain*) gibt den Beitrag eines Merkmals zur Entscheidungsfindung wieder. Das Merkmal mit dem höchsten Informationsgewinn wird zum Wurzelknoten. Dieses ist am besten geeignet um Objekte zu klassifizieren.

***Kruskal Wallis-Test***

Der Kruskal-Wallis-Test (Kruskal und Wallis, 1952), auch H-Test genannt, ist ein nichtparametrisches Verfahren zum Vergleich der Mittelwerte mehrerer Stichproben. Er stellt eine Erweiterung des Wilcoxon's Test dar und kann auf mehr als zwei Stichproben angewendet werden. Eine Normalverteilung wird hier nicht vorausgesetzt, jedoch müssen die Daten ordinalskaliert, die Strichproben unabhängig und die Merkmale stetig sein. Wei (Wei, 1981) hat diesen Test auf seine Robustheit untersucht.

***mRMR***

Das mRMR-Verfahren (engl, *minimum Redundancy Maximum Relevance*, mRMR) ist eine Methode zur Merkmalsextraktion, bei welcher die Redundanz minimiert und gleichzeitig die Relevanz eines Merkmals maximiert wird (Peng et al., 2005). Die Redundanz ist hierbei ein Maß dafür, wie stark die Merkmale voneinander abhängen. Die Relevanz gibt die mittlere Transinformation (gegenseitige Information) zwischen einem einzelnen Merkmal und der Klassenzugehörigkeit an.

***Relief-F***

Der Relief-F Algorithmus entwickelte sich aus dem ursprünglichen Relief-Algorithmus (Kira und Rendell, 1992). Dieser konnte zur Lösung des Zwei-Klassen-Problems verwendet werden. Sein Vorteil ist, dass er auch auf nichtlineare Problemstellungen angewendet werden kann. Er bewertet ein Merkmal anhand der Manhattan Distanz zu seinen nächsten Nachbarn im Merkmalsraum (Liu und Motoda, 2008).

***T-Test***

Der T-Test (Student, 1908) ist ein parametrisches Verfahren zur systematischen Unterscheidung der Mittelwerte zweier Gruppen. Er hilft bei der Entscheidung, ob ein gefundener Unterschied rein zufällig oder wirklich bedeutsam ist. Die Stichproben müssen voneinander unabhängig sein, und in beiden Proben wird eine Normalverteilung des untersuchten Merkmals vorausgesetzt. Die Formel lautet:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Dabei ist:

*m*: Mittelwert Klasse 1/2

*s*: Standardabweichung Klasse 1/2

*n*: Anzahl Proben Klasse ½

## 3.4 Methoden zur Klassifikation

Für die systematische Untersuchung verschiedener Klassifikationsmethoden mit dem Ziel der Entwicklung einer schnellen Tumorsubtypen-Klassifikation wurden folgende Verfahren verwendet:

***Lineare Diskriminanzanalyse***

Ein Verfahren aus der multivariaten Statistik ist die lineare Diskriminanzanalyse (engl. *Linear Discriminant Analysis*, LDA). Sie ist eine Verallgemeinerung der Fisher's Diskriminanz-analyse und dient zur Unterscheidung von zwei oder mehreren Gruppen, wobei sie diese auf zur Unterscheidung relevante Merkmale untersucht (McLachlan, 2004).

***k-Nächste-Nachbarn Algorithmus***

Beim k-Nächste-Nachbarn Algorithmus (eng. *k-Nearest Neighbours*, KNN) wird für jeden Datenpunkt der Abstand zu seinen k nächsten Nachbarn berechnet. Der Abstand wird mit Hilfe eines Distanzmaßes berechnet. Die Entscheidung fällt dann durch einen Mehrheitsbeschluss (Altman, 1992). In diesem Dissertationsprojekt dabei die nächsten drei Nachbarn gewählt. Als Distanzmaß wurde die Euklidische Distanz verwendet.

***Hierarchische Clusteranalyse***

Die hierarchische Clusteranalyse (engl. *Hierarchical Cluster Analysis*, HCA) ist ein distanzbasiertes Verfahren zur Clusteranalyse, bei dem die Objekte, welche zueinander eine geringe Distanz aufweisen, zu einem Cluster zusammengefasst werden (Rokach und Maimon, 2005). Das Ergebnis ist eine hierarchische Struktur, welche grafisch durch ein zweidimensionales Diagramm, dem sogenannten Dendrogramm, dargestellt wird. Als Distanzmaß wurde hier die *ward*-Methode eingesetzt (Ward Jr, 1963). Das Zusammenfassen der Cluster erfolgte durch das *Single-Linkage*-Verfahren.

***k-Means***

Der *k-Means* Algorithmus (Lloyd, 1982) ist neben der HCA eines des am häufigsten verwendeten Clusterverfahren. Es werden k Startpunkte gewählt, von denen aus die Datenpunkte so zusammengefasst werden, dass die Summe der quadratischen Abweichungen der Cluster-Mittelpunkte minimal wird. Daher muss die Anzahl der Cluster im Vorfeld bekannt sein. Des Weiteren sollten die Cluster im Datensatz ungefähr gleich groß sein und auch nicht viele Ausreißer enthalten. Das Ergebnis stark davon ab, an welcher Stelle die Start-Mittelpunkte gesetzt wurden.

***k-Medoids***

Der *k-Medoids* Algorithmus ist dem *k-Means* Clusterverfahren sehr ähnlich. Beide teilen die Datenmenge in Gruppen und versuchen den Abstand der Punkte in einer Gruppe zu minimieren. Im Gegensatz zum *k-Means* Verfahren werden hier jedoch Datenpunkte als Cluster-Zentren gewählt. Ein häufig verwendeter Algorithmus hierfür ist der *Partitioning Around Medoids (PAM)* (Theodoridis et al., 2010).

## 3.5 Erstellung von Konzentrations-Wirkungskurven

Eine Konzentrations-Wirkungskurve (Motulsky und Christopoulos, 2004) stellt die Wirkung eines Stoffes in Abhängigkeit von deren Konzentration dar. Dabei entspricht die x-Achse der Konzentration des Stoffes, während die y-Achse die entsprechende Reaktion anzeigt. Es ist üblich, die Reaktion gegen den Logarithmus der Konzentration aufzutragen. Im Idealfall ergibt sich eine sigmoide Kurve. Diese lässt sich durch die minimale Reaktion *(Bottom)*, die maximale Reaktion *(Top)*, die Steigung *(Hill Slope)* und den EC50-Wert (mittlere Effektivdosis) definieren. Folgende Formel wurde für die Erstellung von Konzentrations-Wirkungskurven verwendet:

$$Y = Bottom + \frac{(Top - Bottom)}{1 + 10^{(LogEC_{50} - X) * HillSlope}}$$

## 3.6 Kennzahlen

Zur Bewertung verschiedener Klassifikations- und Regressionsergebnisse wurden die folgenden Kennzahlen verwendet.

***Reinheit und Entropie***

Zum Vergleich der Qualität und Rechenzeit der in diesem Dissertationsprojekt entwickelten Kennzahl zur Bewertung einer unüberwachten Klassifikation wurden die Reinheit (engl. *purity*) und Entropie (engl. *entropy*) verwendet. Diese werden nach folgenden Formeln berechnet:

$$p_{ij} = \frac{n_{ij}}{n_j}$$

$$entropy_j = -\sum_{i=1}^{L} p_{ij} log_2 p_{ij} \qquad purity_j = \max p_{ij}$$

$$entropy = \sum_{i=1}^{K} \frac{n_i}{N} entropy_j \qquad purity = \sum_{i=1}^{K} \frac{n_i}{N} purity_j$$

Dabei ist:

$p_{ij}$: Wahrscheinlichkeit, ein Objekt aus Cluster *j* zur Klasse *i* gehört

*L*: Anzahl Klassen

*K*: Anzahl Cluster

$n_j$: Anzahl Objekte in Cluster *j*

$n_{ij}$: Anzahl Objekte der Klasse *i* in Cluster *j*

*N*: Gesamtzahl Objekte im Datensatz

***Mittlerer quadratischer Fehler***

Für die Erstellung der Konzentrations-Wirkungs-Kurven wurde als Maßzahl der mittlere quadratische Fehler (engl. *Mean Squared Error*, MSE) verwendet (Allen, 1971). Dieser gibt die Abweichung des geschätzten Wertes zum berechneten Wert an:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Dabei ist:

$y_i$: gemessener Wert

$\hat{y}_i$: vorhergesagter Wert

*n*: Anzahl Werte

***Korrelationskoeffizient***

Bei der Entwicklung der Methode zur automatischen Bestimmung von Wirkstoffeffekten wurde als Ausschlusskriterium einer Konzentrations-Wirkungs-Kurve der Korrelationskoeffizient nach Pearson (Pearson, 1895) verwendet. Dieser gibt den linearen Zusammenhang zwischen zwei oder mehreren Variablen an. Er ist dimensionslos und kann Werte zwischen -1 und +1 annehmen:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2}}$$

Dabei ist:

$x_i$: gemessener Wert

$y_i$: vorhergesagter Wert

$\bar{x}$: Mittelwert der Messwerte

$\bar{y}$: Mittelwert der vorhergesagten Werte

$n$: Anzahl Werte

***„Fold Change“***

Als weitere Kennzahl bei der Entwicklung einer Methode zur automatischen Bestimmung von Wirkstoffeffekten wurde das Verhältnis der Intensität bei keiner Wirkstoffbehandlung zur Intensität bei maximaler Konzentration des Wirkstoffes verwendet. Normalerweise wird hierbei der Logarithmus zur Basis 2 verwendet (Love et al., 2014). Dieses wird als *Fold Change* bezeichnet und lässt sich wie folgt berechnen:

$$Fold\ Change = abs(log_2 \frac{I_{max}}{I_0})$$

Dabei ist:

$I_{max}$: Intensität bei maximaler Wirkstoffkonzentration

$I_0$: Intensität bei keiner Wirkstoffbehandlung

Da die Bestimmung für steigende und fallende Kurven getrennt erfolgte, wurde der Betrag des *Fold Change* verwendet.

# 4 ERGEBNISSE

Die Ergebnisse dieser Arbeit gliedern sich in folgende Teile:

i. Vor der Entwicklung geeigneter Analysemethoden widmet sich der erste Teil zunächst der optimierten und schnellen Vorverarbeitung von Massenspektren.
ii. Der zweite Teil beinhaltet die Entwicklung geeigneter Kennzahlen zur Bewertung eines optimalen Merkmalsextraktions-Klassifikations-Workflows.
iii. Im Fokus des dritten Teils steht die Etablierung eines Workflows zur optimierten Merkmalsextraktion und Klassifikation von MS-*Fingerprints*. Anhand geeigneter Tumorzelllinien werden verschiedene Möglichkeiten zur Merkmalsextraktion und Klassifikation (überwacht und unüberwacht) miteinander verglichen und mit Hilfe der in Punkt 2 entwickelten Kennzahl bewertet.
iv. Der vierte Teil beschäftigt sich mit der Entwicklung eines Regressionsmodells zur automatisierten Erstellung von Konzentrations-Wirkungskurven und der quantitativen Bewertung der zellulären Wirkung von zielgerichteten Arzneimittel-Wirkungen.
v. Um die hier entwickelten Workflows, Methoden und Algorithmen schnell und einfach anwenden zu können, werden sie zuletzt in einer graphischen Benutzeroberfläche zusammengeführt.

## 4.1 Optimierung der Vorverarbeitung von Massenspektren

Aufgrund von verschiedenen Messeinstellungen, Geräteschwankungen, Experimentatoren oder Messtagen können Varianzen zwischen den Spektren auftreten. Zudem können die Spektren durch verschiedene Arten von Rauschen verfälscht sein. Für nachfolgende statistische Analysen ist es jedoch zwingend notwendig, dass die Massenspektren eine gemeinsame Massenachse besitzen. Standardmäßig wird dies derzeit durch ein *Peak Picking* mit nachfolgender Rekalibration realisiert. Um die Peak-Auswahl jedoch mit Hilfe statistischer Methoden durchführen zu können, bedarf es einer Vorverarbeitung der Spektren auf den unreduzierten Datensätzen. Abbildung 2 zeigt, aus welchen Vorverarbeitungs-schritten der hier verwendete Workflow aufgebaut ist.

Um überhaupt mit den Spektren in MATLAB arbeiten zu können, wurde zunächst eine neue Datenstruktur entwickelt, welche einen einfachen und schnellen Zugriff auf die Daten ermöglicht. Die Basislinienkorrektur wurde mit Hilfe der MATLAB Funktion *msbackadj* aus der *Bioinformatics Toolbox* durchgeführt. Die Glättung der Spektren wurde mit dem Savitzky-

Golay-Filter realisiert. Dazu wurde die MATLAB Funktion *sgolayfilt* aus der *Signalprocessing Toolbox* verwendet. Das normalerweise angewandte *Peak Picking* Verfahren wurde hier nicht standardmäßig in den Vorverarbeitungsprozess miteinbezogen. Dadurch sollte verhindert werden, dass eventuell wichtige Peaks mit kleinen Intensitäten verloren gehen. Stattdessen wurde eine neue Methode entwickelt, welche für die nicht reduzierten Spektren eine gemeinsame Massenachse berechnet. Des Weiteren wurde ein einfaches Verfahren zur Ausrichtung der Spektren realisiert, um die Ähnlichkeit zwischen ihnen zu erhöhen. Die Normierung der Spektren erfolgte entweder durch eine TIC-Normierung (engl. *Total Ion urrent*, TIC) oder einer *Z-Score*-Normierung, je nachdem welches Analyseverfahren im Anschluss angewendet werden sollte. Zuletzt wurde eine neue Methode zum Entfernen unerwünschter Rauschanteile entwickelt.

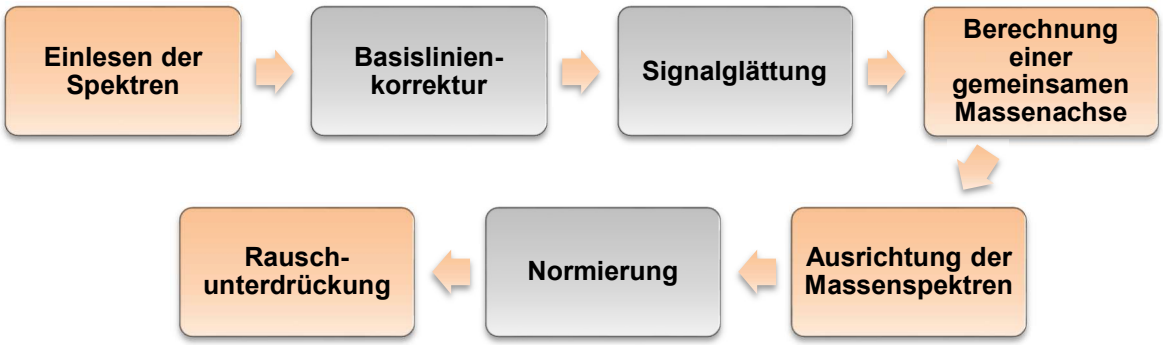


**Abbildung 2: Vorverarbeitung der Massenspektren.** Zunächst wurden die Spektren in eine MATLAB geeignete Datenstruktur übertragen. Um eine eventuell verschobene Grundlinie auf null zu verschieben, wurde im ersten Schritt eine Basislinienkorrektur durchgeführt. Danach wurden die Spektren zur Verbesserung des Signal-Rausch-Verhältnisses einer Signalglättung mit dem Savitzky-Golay-Filter unterzogen. Im nächsten Schritt wurde eine gemeinsame Massenachse für alle Spektren berechnet. Ungenauigkeiten bei der Kalibrierung wurden durch Ausrichten der Massenspektren beseitigt. Um die Spektren miteinander vergleichen zu können, wurden sie einer Normierung unterzogen. Zuletzt wurden die Rauschanteile der Spektren entfernt.

### 4.1.1 Datenstruktur zum Einlesen der Spektren

Nach einer Messung liegen die Spektren unsortiert in einem Datenformat vor, welches vom Gerätehersteller vorgegeben wird. Das *Autoflex Speed* liefert für jedes Massenspektrum eine sogenannte *fid*-Datei, in welcher die Intensitäten abgespeichert sind. Diese können mit der MATLAB Funktion *fread* aus der *Data Import and Export Toolbox* eingelesen werden. Die Parameter (ml1, ml2, ml3, DELAY, DW, TD) zur Berechnung der Massenachse werden in einer Datei namens *acqu* abgelegt. Die *m/z*-Werte wurden in folgender Schleife berechnet:

```
for i = 1 : TD
    time = DELAY;
    time = time + DW * i;
    a = (-ml2-(0.1E7*(-5E5+sqrt(0.25E12-ml1*ml2*ml3+ml1*ml3*time)));
    b = (ml1*ml3)+time)/ml3;
    mz(i) = a / b
end
```

**Abbildung 3: Code-Auszug aus dem MATLAB Skript zur Berechnung der Massenachse**

Um einfach und schnell mit den Massenspektren arbeiten zu können, wurden im Anschluss daran mit Hilfe von objektorientierter Programmierung verschiedene Schablonen erstellt, welche neben den Massenspektren selbst noch weitere Informationen und Eigenschaften speichern können. Diese Schablonen werden dabei in der Informatik als Klassen und die zusätzlichen Informationen als Klassenvariablen bezeichnet. Es können mehrere so-genannte Objekte einer Klasse angelegt werden.

Zunächst wurde eine Klasse für einen Datensatz (Experiment) erstellt, welche neben der Bezeichnung zum Beispiel noch die Information über die Anzahl der untersuchten Erkrankungen oder Subtypen enthält. Des Weiteren wurde eine Klasse für die Erkrankungen / Subtypen erstellt, welche analog die jeweilige Bezeichnung enthält und Auskunft über die Anzahl der zugehörigen Massenspektren gibt. Zuletzt wurde eine Klasse für ein Massenspektrum selbst erstellt, welche über die eigentlichen *m/z*-Werte und Intensitäten verfügt. Die Klassen wurden derart verschachtelt, so dass ein Datensatz-Objekt mehrere Erkrankungen / Subtypen-Objekte enthält, welche wiederum mehrere Massenspektren-Objekte enthalten (Abbildung 4). Diese spezielle Datenstruktur vereinfacht den Umgang mit den Daten erheblich. Es ist jederzeit ersichtlich, zu welchem Datensatz / Experiment und zu welcher Erkrankung / Subtyp ein Spektrum gehört, und es kann unkompliziert auf alle wichtigen Informationen zugegriffen werden (Abbildung 5).

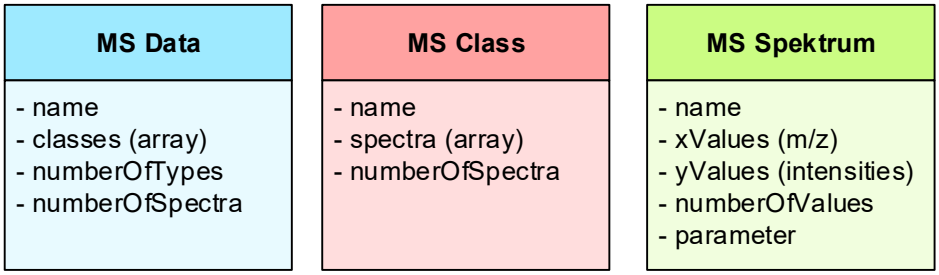


**Abbildung 4: Klassen und Klassenvariablen für die Datenhaltung im Rechner.** Mit Hilfe von objektorientierter Programmierung wurden für die Datensätze, die verschiedenen Zelltypen und Einzelspektren Klassen erstellt, welche die Möglichkeit zur Speicherung aller wichtigen Informationen bereitstellen.

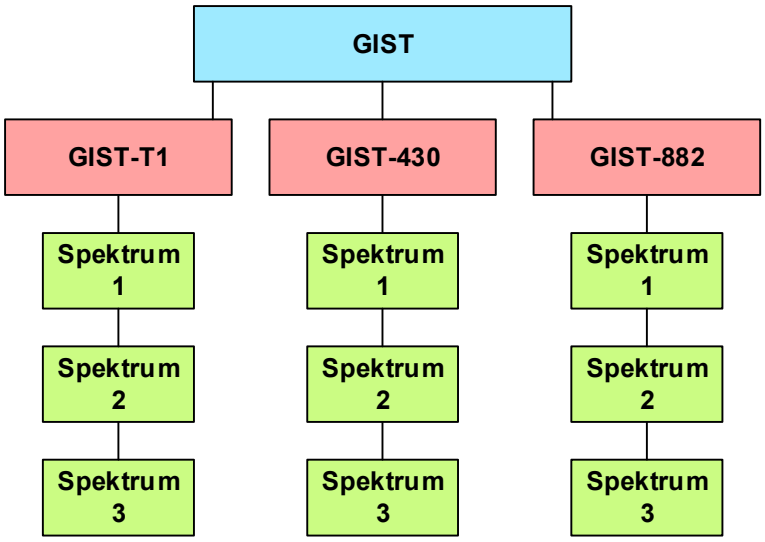


**Abbildung 5: Datenstruktur am Beispiel eines GIST-Zelllinien-Datensatzes.** Der GIST-Zelllinien-Datensatz repräsentiert hier ein einzelnes MS-Experiment. Die erste Hierarchie-Ebene besteht aus dem allgemeinen GIST-Zelllinien-Datensatz. Dieser enthält die drei Zelllinien GIST-T1, GIST-430 und GIST-882. Jede Zelllinie wiederum enthält mehrere Einzelspektren (technische Replikate).

### 4.1.2 Methode zur Berechnung einer gemeinsamen Massenachse

Die Massenachsen der Spektren können in unterschiedlichen Wertebereichen liegen. Ein Grund dafür kann zum Beispiel sein, dass noch keine standardisierte Vorgehensweise (engl. *Standard Operating Procedure*, SOP) im Laborbetrieb für das entsprechende Experiment existiert. Sinnvolle, statistische Analysen können jedoch nur auf einer gemeinsamen Massenbereich durchgeführt werden. Um auf das standardmäßig angewandte *Peak Picking* zu verzichten, muss die Berechnung der gemeinsamen Massenachse auf den unreduzierten Spektren erfolgen. Dazu wurde im ersten Schritt der gemeinsame Massenbereich aus allen Spektren eines Datensatzes extrahiert werden. Es wurde in MATLAB eine Funktion implementiert, welche in allen Spektren den jeweils größten bzw. kleinsten Massenwert bestimmt und alle Datenpunkte, welche sich nicht in diesem Bereich befinden, löscht. Abbildung 6 zeigt den jeweils rechten und linken Rand dreier GIST-Datensätze, welche von diversen Experimentatoren an verschiedenen Tagen mit unterschiedlichen Geräteeinstellungen gemessen wurden. Sie unterscheiden sich deutlich im aufgenommen Massenbereich, der gemeinsame Massenbereich konnte jedoch erfolgreich extrahiert werden.

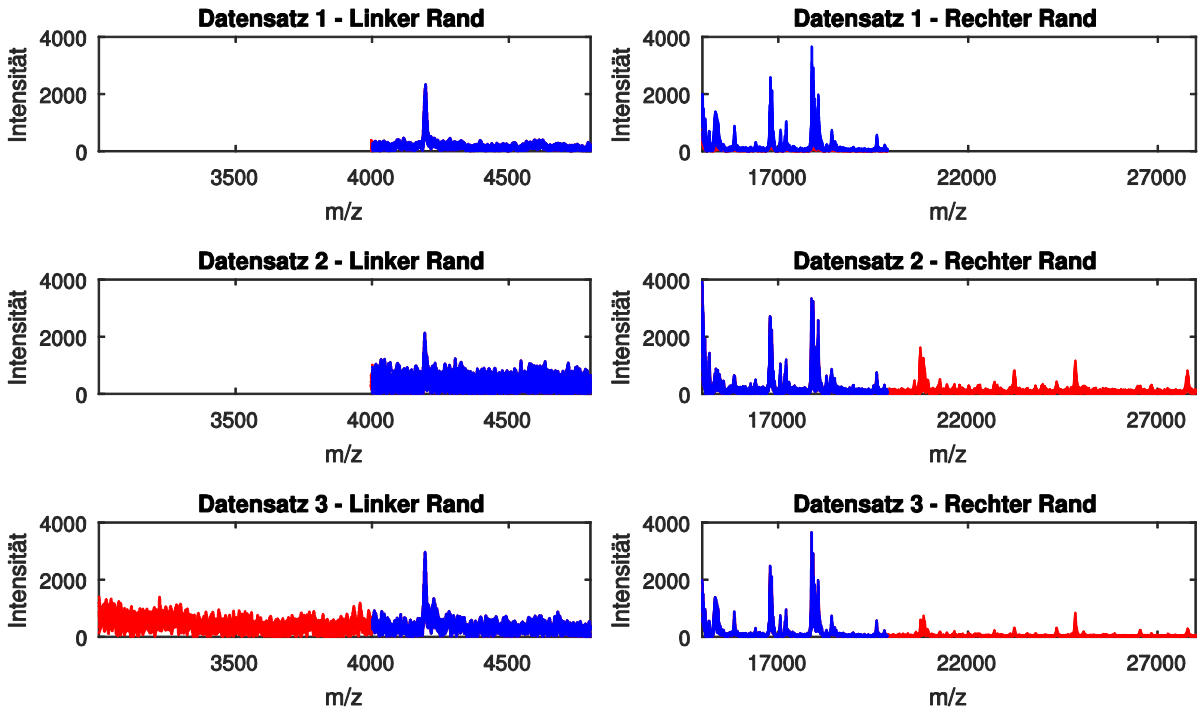


**Abbildung 6: Extraktion des gemeinsamen Massenbereichs aus drei GIST-Zelllinien-Datensätzen.** Die drei Datensätze wurden von verschiedenen Experimentatoren mit unterschiedlichen Geräteeinstellungen gemessen. Der Randbereich (rot) wurde von der weiteren Verarbeitung ausgeschlossen. Der gemeinsame Massenbereich der MS-Fingerprints (blau) liegt hier zwischen 4001,25 Dalton und 19.876,05 Dalton.

Neben dem Wertebereich der Massenachse kann sich auch aus dem oben genannten Gründen deren Auflösung unterscheiden. Des Weiteren können aufgrund von Detektorungenauigkeiten dieselben Proteine in verschiedenen Spektren leicht variierende Massenwerte aufweisen (Abbildung 7). Um alle Spektren auf eine gemeinsame Massenachse zu bringen, wurde zunächst eine Referenzachse gewählt. Damit keine Datenpunkte verfälscht werden, wurde dazu das Spektrum des Datensatzes gewählt, welches die niedrigste Datenauflösung besitzt. Da Massenspektren im Allgemeinen sehr hoch aufgelöst sind, stellt dies eine Datenreduktion ohne Informationsverlust dar.

Im nächsten Schritt wurden alle anderen Spektren mit Hilfe einer linearen Interpolation an dieses Referenzspektrum angepasst. Aufgrund der hohen Auflösung ist eine kubische oder *Spline*-Interpolation nicht notwendig gewesen. Realisiert wurde dies mit Hilfe der MATLAB Funktion *interp1* aus der *Mathematics Toolbox* (Abbildung 8).

Durch diesen Vorverarbeitungsschritt war es nun möglich, auf das normalerweise durchgeführte *Peak Picking* zu verzichten, welches eventuell wichtige Peaks mit kleinen Intensitäten aussortieren könnte. Die Bewertung der Signifikanz der Peaks konnte nun mit Hilfe statistischer Methoden durchgeführt werden.

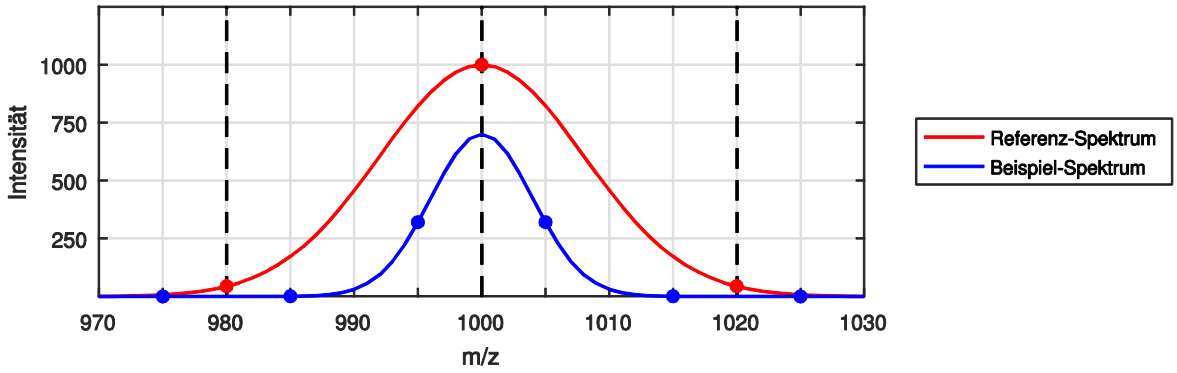


**Abbildung 7: Unterschiedliche Auflösung und Verteilung der *m/z*-Bins.** Anhand dieser sehr stark vereinfachten Abbildung soll die Notwendigkeit einer Interpolation verdeutlicht werden. Insgesamt ist hier ein Massenbereich von 60 Dalton dargestellt. Das Referenz-Spektrum verfügt in diesem über drei, das Beispiel-Spektrum über fünf Datenpunkte. Jedoch besitzen die beiden Spektren nicht einen gemeinsamen *m/z*-Wert, was eine nachfolgende Analyse unmöglich macht.

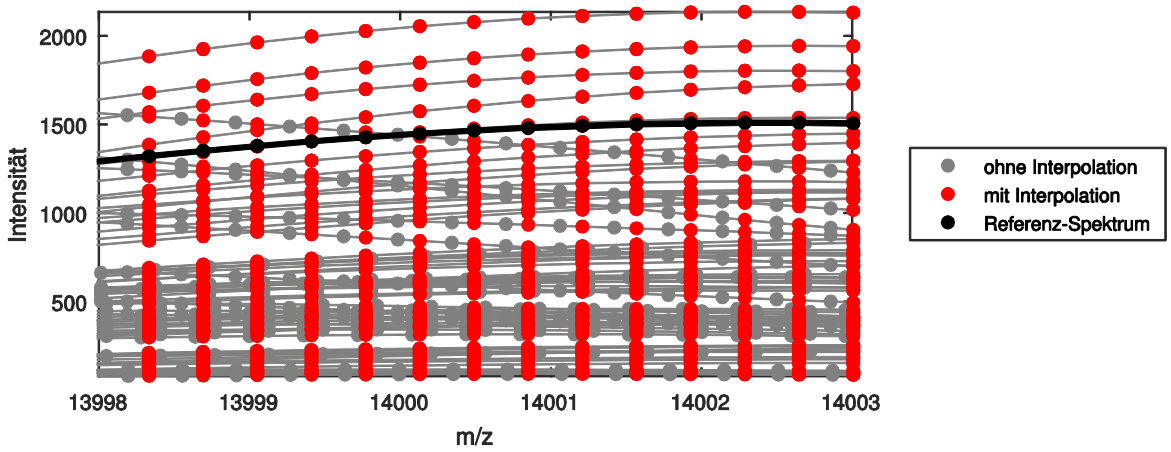


**Abbildung 8: Berechnung der gemeinsamen Massenachse dreier GIST-Zelllinien-Datensätze.** Zur Berechnung einer gemeinsamen Massenachse dreier GIST-Zelllinien-Datensätze wurde das Spektrum mit der geringsten Auflösung als Referenz-Spektrum festgelegt (schwarz). Die Intensitäten der ursprünglichen Massenwerte (grau) wurden mit Hilfe einer linearen Interpolation an die *m/z*-Werte des Referenz-Spektrums angepasst (rot).

### 4.1.3 Methode zum Ausrichten der Massenspektren

Zur Beseitigung von systematischen Verschiebungen der Massenachse, welche zum Beispiel durch Fehler bei der Kalibrierung entstehen können, wurde eine Methode entwickelt, welche Massenspektren derart gegeneinander ausrichtet, so dass sie die größtmögliche Korrelation untereinander besitzen. Als Referenzspektrum wurde hierfür das erste Spektrum des jeweiligen Datensatzes gewählt. Zwischen diesem und allen anderen Spektren wurde nun die Korrelationsfunktion berechnet, wobei die Spektren jeweils 50 Schritte nach rechts

und links linear verschoben wurden. Anschließend wurde das Maximum der Korrelationsfunktion und die Anzahl der Schritte berechnet, um welche die Spektren verschoben werden müssen, um zum Referenzspektrum die größte Ähnlichkeit zu besitzen (Abbildung 9).

Abbildung 10 zeigt am Beispiel eines Peaks das Ergebnis der linearen Ausrichtung. Die Maxima der einzelnen Spektren besitzen jetzt nahezu den gleichen *m/z*-Wert. Im Vergleich zu vorher konnte die Standardabweichung erheblich verringert werden. Ein besonderer Vorteil dieser Methode ist, dass sie, neben der schnellen Durchführung, kein spezielles Vorwissen über Kalibrierungsstandards erfordert.

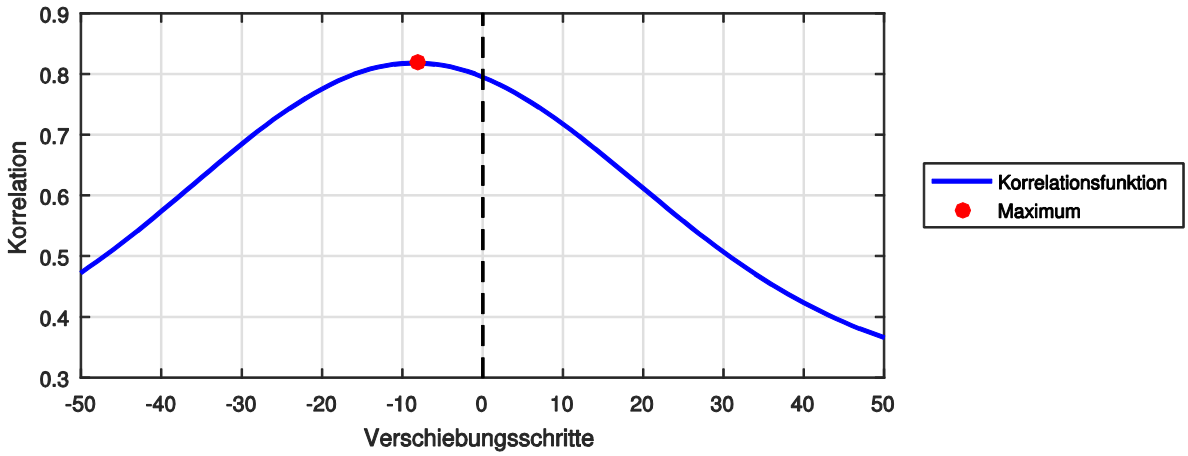


**Abbildung 9: Korrelationsfunktion zwischen einem Spektrum aus einem GIST-Zelllinien-Datensatz und dem Referenz-Spektrum.** Das Spektrum muss gegenüber dem Referenz-Spektrum um acht Schritte linear nach links verschoben werden, damit die Ähnlichkeit zwischen den beiden so groß wie möglich ist.

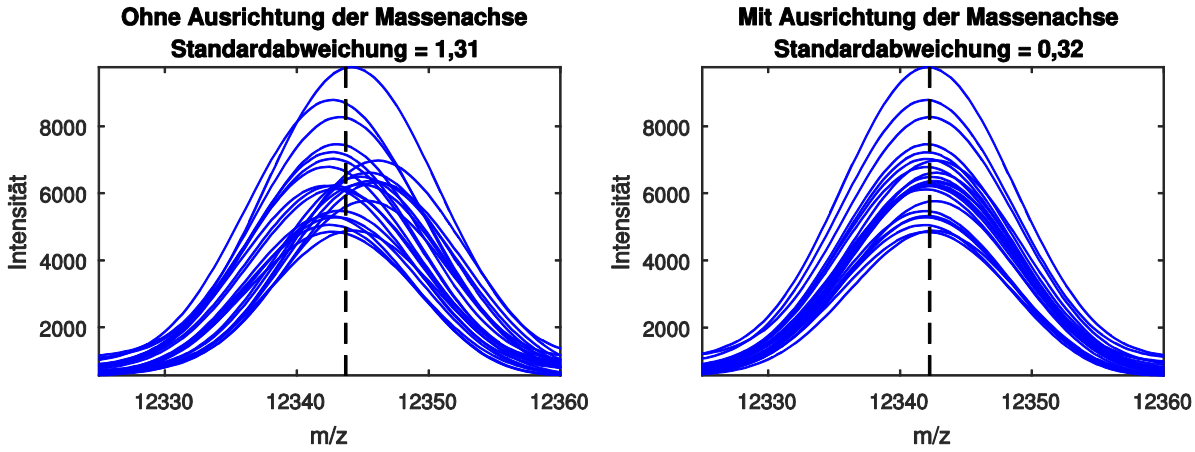


**Abbildung 10: Ergebnis der Ausrichtung der Massenachse.** Die Abbildung zeigt beispielhaft einen Peak eines GIST-Zelllinien-Datensatzes ohne Ausrichtung der Massenachse (links) und mit Ausrichtung der Massenachse (rechts). Die Maxima des Peaks liegen nun deutlicher übereinander und die Standardabweichung des zugehörigen m/z-Wertes konnte von 1,31 auf 0,32 verbessert werden.

### 4.1.4 Methode zu Unterdrückung von Rauschanteilen

Ein „normales“ Linear-Positiv-Spektrum kann aus über 50.000 Datenpunkten bestehen, wobei üblicherweise jedoch nur zwischen 100-150 Signale / Peaks zu sehen sind. Selbst unter Berücksichtigung der Anzahl an Bins, welche zu einem Peak gehören, ist das Verhältnis zwischen dem Signal- und Rauschanteil äußerst gering. Dies lässt zu der Annahme führen, dass die meisten Datenpunkte des Spektrums im Rauschen liegen und somit gelöscht werden können.

Zur Bestimmung des Rauschpegels wurde eine neue und schnelle Methode entwickelt, mit deren Hilfe in wenigen Schritten das Rauschen aus den Spektren entfernt werden kann. Zunächst wurden dazu die Spektren einer Normierung unterzogen, um die Höhe der Peaks auf vergleichbare Werte zu bringen. Die Art der Normierung hängt dabei vom darauffolgenden Analyseverfahren ab. Im nächsten Schritt wurde das Histogramm des gegebenen Datensatzes berechnet. Ein Histogramm gibt die Häufigkeitsverteilung der Intensitäten an, wobei auf die x-Achse die Intensitäten und auf die y-Achse die zugehörige Häufigkeit aufgetragen werden (Abbildung 11). Das Maximum in diesem Diagramm entspricht dann der Intensität, welche am häufigsten im Datensatz enthalten ist. Aufgrund der oben getroffenen Annahme, dass die meisten Datenpunkte eines Spektrums im Rauschen liegen, wurde dieser Intensitätswert als Rauschpegel festgelegt.

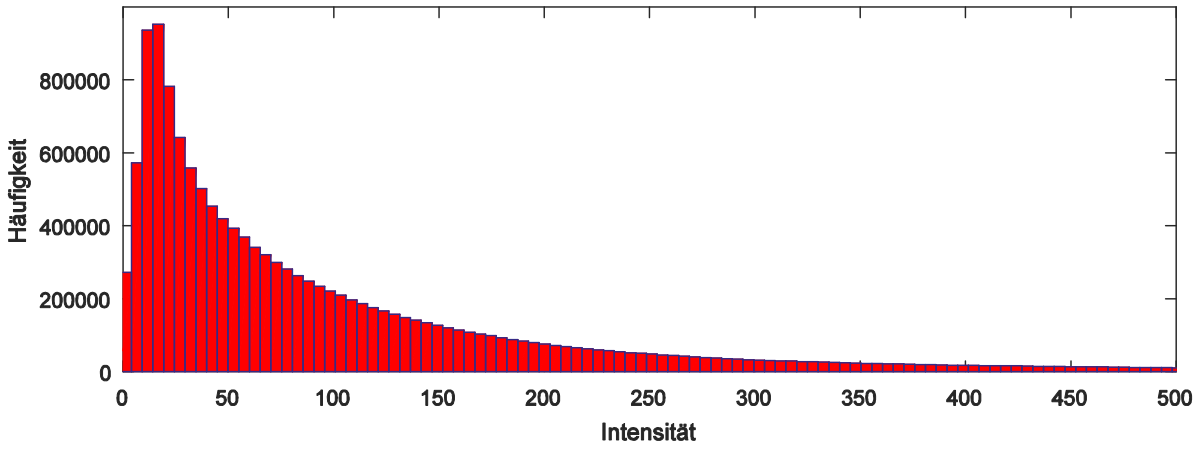


**Abbildung 11: Histogramm eines K562-Zelllinien-Datensatzes.** Die Abbildung zeigt beispielhaft das Histogramm des niedrigen Intensitätsbereiches eines K562-Zelllinien-Datensatzes. Die Intensität, welche am häufigsten in den Spektren enthalten ist, beträgt hier den Wert 21 und kann als Rauschpegel festgelegt werden.

Da das Rauschen im Spektrum jedoch nicht überall gleich ist, sondern in Richtung des höheren Massenbereichs immer mehr abnimmt, muss der Verlauf des Rauschpegels entsprechend angepasst werden. Dazu wurde separat ein Histogramm aus den ersten und

letzten 5000 *m/z*-Werten des Datensatzes erstellt und der jeweilige Rauschpegel bestimmt. Mit Hilfe dieser beiden Werte konnte nun ein linear abfallender Rauschpegel über die gesamte Massenachse berechnet werden (Abbildung 12).

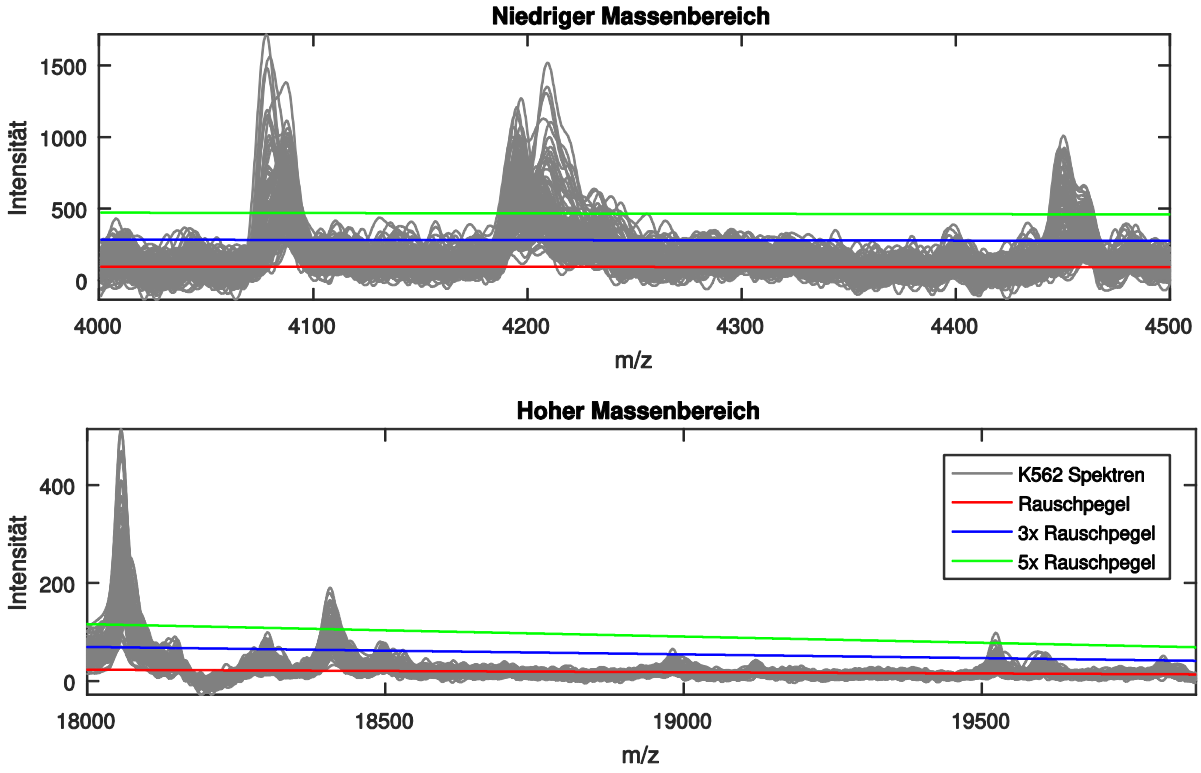


**Abbildung 12: Histogramm basierter Rauschpegel am Beispiel eines K562-Zelllinien-Datensatzes.** Die Abbildung zeigt den linear abfallenden Rauschpegel im niedrigen und hohen Massenbereich. Je nach angestrebtem Signal-Rausch-Verhältnis kann dieser mit einem Faktor multipliziert werden. Alle Datenpunkte, welche sich unter dem Rauschpegel befinden, können gelöscht werden.

Im letzten Schritt wurde zur Bestimmung der Bereiche mit großem Rauschanteil die *Skyline* des Datensatzes berechnet. Diese enthält für jeden *m/z*-Wert die maximale Intensität des Datensatzes. Es wurden alle Datenpunkte in den Einzelspektren gelöscht, für welche die *Skyline* unterhalb des Rauschpegels lag (Abbildung 13).

Durch den Einsatz dieser Methode konnte verhindert werden, dass nachfolgende statistische Analysen verfälscht werden. Gerade in Bereichen mit hohem Rauschanteil ist die Varianz meist sehr gering, was jedoch ein wichtiges Kriterium für die erfolgreiche Unterscheidung verschiedener Objekte darstellt.

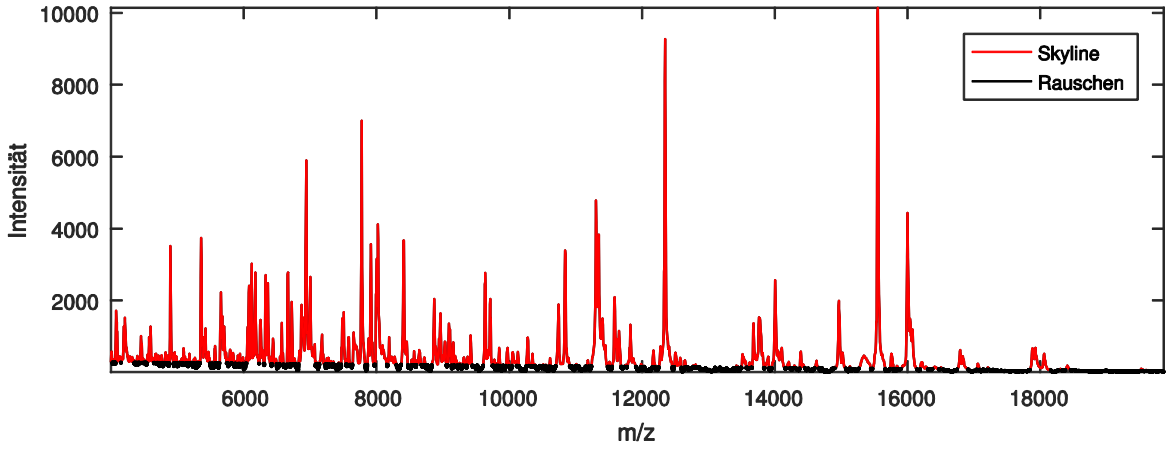


**Abbildung 13: Rauschunterdrückung am Beispiel der *Skyline* eines K562-Zelllinien-Datensatzes.** Mit Hilfe der *Skyline* des Datensatzes (maximale Intensität eines *m/z*-Wertes) werden die verrauschten Bereiche ermittelt, welche in den Einzelspektren gelöscht werden können. Hier wurde der Rauschpegel so gelegt, dass die Signale mindestens ein Signal-Rausch-Verhältnis von fünf besitzen.

## 4.2 Entwicklung einer Kennzahl zur Methodenvalidierung

Zur objektiven und vergleichenden Bewertung eines statistischen Modells wird ein objektives und quantitatives Kriterium benötigt. Dieses muss schon vor der eigentlichen Methodenentwicklung festgelegt werden. Zur Bewertung einer überwachten Klassifikation, bei welcher die Klassenzugehörigkeit im Vorfeld bekannt ist, wurde die Klassifikationsrate verwendet. Diese gibt das Verhältnis zwischen der Anzahl richtig klassifizierter Objekte zur Gesamtzahl der Objekte an. Zur Bewertung einer unüberwachten Klassifikation, bei welcher die Klassenzugehörigkeit zuvor nicht bekannt ist, muss eine interne oder externe Clusteranalyse durchgeführt werden. Bei der internen Clusteranalyse werden keine zusätzlichen Informationen verwendet. Bei der externen Clusteranalyse hingegen werden weitere Informationen zur Bewertung hinzugenommen, welche jedoch nicht schon während der Klassifikation verwendet wurden. Da hier die Klassenzugehörigkeit der Spektren bekannt war, wurde für die unüberwachte Klassifikation eine externe Clusteranalyse durchgeführt. Folgende Abbildung veranschaulicht die Problematik bei der Bewertung einer unüberwachten Klassifikation.

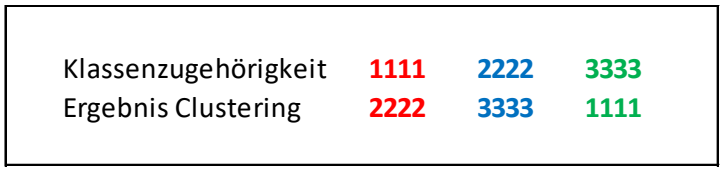


**Abbildung 14: Beispiel Ergebnis Clusteranalyse**. Die Proben einer Datenmenge bestehend aus drei Klassen konnten eindeutig einem Cluster zugewiesen werden. Da die Klassenzugehörigkeit jedoch nicht bekannt war, wurde die Indizierung vom Klassifikator zufällig gewählt. Trotz einer Klassifikationsrate von 100%, entspräche das Ergebnis 0%.

Um eine unüberwachte Klassifikation dennoch mit der exakten Klassifikationsrate bewerten und die Qualität der oben beschriebenen *External Indices* beurteilen zu können, besteht die Möglichkeit, jede Gruppe (der Eingangsdaten) einmal jedem Ergebnis-Cluster zuzuweisen und wiederholt die Klassifikationsrate zu berechnen. Danach werden die Ergebnisse miteinander verglichen. Das beste Ergebnis dieser Kombinationen entspricht dann der exakten Klassifikationsrate. Die Komplexität dieser Berechnung steigt jedoch mit zunehmender Anzahl von Klassen. Für n Klassen gibt es Fakultät(n) Kombinations-möglichkeiten. Mit zunehmender Anzahl von Klassen erfordert dies jedoch eine immer längere Rechenzeit und stößt an die Grenzen des Arbeitsspeichers. Somit musste dieser Ansatz wieder verworfen werden. (Tabelle 8).

**Tabelle 8: Kombinationen der Index-Zuordnung zur Berechnung der exakten Klassifikationsrate**

| Anzahl Klassen | Anzahl Kombinationsmöglichkeiten | Rechenzeit |
|---|---|---|
| 2 | 2 | < 0,1 ms |
| 3 | 6 | < 1 ms |
| 4 | 24 | < 1 ms |
| 5 | 120 | ~ 1 ms |
| 6 | 720 | ~ 7 ms |
| 7 | 5.040 | ~ 50 ms |
| 8 | 40.320 | ~ 400 ms |
| 9 | 362.880 | ~ 4 sec |
| 10 | 3.628.800 | ~ 36 sec |
| 11 | 39.916.800 | ~ 7 min |
| 12 | 479.001.600 | ~ 80 min |
| 13 | 6.227.020.800 | nicht berechnet |
| 14 | 87.178.291.200 | nicht berechnet |
| 15 | 1.307.674.368.000 | nicht berechnet |

Im nächsten Schritt wurden zwei weitere Kennzahlen verwendet, mit welchen allgemein eine externe Clusteranalyse durchgeführt werden kann: Die Reinheit, welche für ein Cluster bestimmt, inwiefern nur Objekte einer Klasse enthalten sind. Und die Entropie, welche die Verteilung der Klassen in einem einzelnen Cluster berücksichtigt. Die Reinheit bzw. die Entropie des Clusterergebnisses ist dann das gewichtete Mittel der Reinheitswerte bzw. der Entropien aller Cluster. Der Vorteil dieser beiden Kennzahlen ist, dass deren Berechnung sehr schnell durchgeführt werden kann, auch bei einer großen Anzahl von Klassen. Der Nachteil ist jedoch, dass sie keine exakte Aussage über den Erfolg der Klassifikation machen können, sondern nur eine Näherung an die wahre Klassifikationsrate geben können.

Aus diesem Grund wurde eine neue Kennzahl entwickelt, die schnell berechnet werden kann, und die möglichst identische Werte im Vergleich zur exakten Klassifikationsrate liefert. Sie basiert auf den Maxima der Konfusionsmatrix der Clusterergebnisse und wurde daher als *Confusion Matrix Maximum (CMM)* bezeichnet. Die Konfusionsmatrix gibt an, wie viele Objekte jeder Klasse jedem Ergebnis-Cluster zugeordnet wurden. Im Folgenden wird der Algorithmus zur Berechnung des CMM anhand eines einfachen Beispiels erklärt: Gegeben ist ein Datensatz, welcher aus vier Klassen und insgesamt aus 46 Objekten besteht. Dieser wird einem unüberwachten Klassifikator übergeben, welcher die Objekte den Clustern A bis D zuordnet.

***Schritt 1: Erstellung der Konfusionsmatrix***

Zu Beginn wird die Konfusionsmatrix des Klassifikationsergebnisses aufgestellt. Sie stellt die Verteilung der Klassenobjekte zu den einzelnen Clustern dar (Abbildung 15).

| | Cluster A | Cluster B | Cluster C | Cluster D |
|---|---|---|---|---|
| Klasse 1 | 2 | 0 | 5 | 10 |
| Klasse 2 | 8 | 0 | 0 | 1 |
| Klasse 3 | 3 | 6 | 0 | 0 |
| Klasse 4 | 1 | 0 | 7 | 3 |

**Abbildung 15: Konfusionsmatrix zur Veranschaulichung des Klassifikationsergebnisses.** In der Konfusionsmatrix kann abgelesen werden, wie viele Objekte jeder Klasse jedem Cluster zugeordnet wurden. Klasse 1 besitzt zum Beispiel 17 Objekte, wovon zwei dem Cluster A, fünf dem Cluster C und zehn dem Cluster D zugeordnet wurden.

***Schritt 2: Bestimmung des Maximums***

Im nächsten Schritt wird das Maximum der Konfusionsmatrix bestimmt (Abbildung 16). Dieses stellt die größte, richtig klassifizierte Gruppe von Objekten der gleichen Klasse dar. Somit kann diese Klasse dem entsprechenden Cluster zugeordnet werden.

| | Cluster A | Cluster B | Cluster C | Cluster D |
|---|---|---|---|---|
| Klasse 1 | 2 | 0 | 5 | 10 |
| Klasse 2 | 8 | 0 | 0 | 1 |
| Klasse 3 | 3 | 6 | 0 | 0 |
| Klasse 4 | 1 | 0 | 7 | 3 |

**Abbildung 16: Bestimmung des Maximums der Konfusionsmatrix.** Das Maximum der Konfusionsmatrix entspricht der größten, richtig klassifizierten Gruppe von Objekten. Hier wurden zehn Objekte der Klasse 1 dem Cluster D zugeordnet

### *Schritt 3: Löschen der zugehörigen Zeile und Spalte*

Sobald eine Klasse mit einem Cluster verbunden wurde, kann die entsprechende Zeile und Spalte gelöscht werden, da es sich bei den anderen darin enthaltenen Objekten um Falsch-Klassifikationen handelt. Das Maximum und die zugehörige Klasse werden abgespeichert (Abbildung 17).

| | Cluster A | Cluster B | Cluster C | Cluster D |
|---|---|---|---|---|
| Klasse 1 | | | | 10 |
| Klasse 2 | 8 | 0 | 0 | |
| Klasse 3 | 3 | 6 | 0 | |
| Klasse 4 | 1 | 0 | 7 | |

**Abbildung 17. Zuordnung einer Klasse zu einem Cluster.** Die Klasse 1 wurde dem Cluster D zugeordnet. Daher können alle Objekte aus Klasse 1, welche den Clustern A bis C zugeordnet waren, gelöscht werden. Des Weiteren können alle Objekte in Cluster D, welche nicht zu Klasse 1 gehören, ebenfalls gelöscht werden.

### *Schritt 4: Solange Schritt 2 und 3, bis alle Klassen einem Cluster zugeordnet sind*

Durch das Löschen einer Zeile und einer Spalte entsteht eine neue Matrix, in welcher wieder das Maximum bestimmt werden kann. Dadurch kann wieder eine Klasse mit einem Cluster verbunden werden. Dies muss solange wiederholt werden, bis alle Klassen einem Cluster zugeordnet sind (Abbildung 18).

| | Cluster A | Cluster B | Cluster C | Cluster D |
|---|---|---|---|---|
| Klasse 1 | | | | 10 |
| Klasse 2 | 8 | | | |
| Klasse 3 | | 6 | | |
| Klasse 4 | | | 7 | |

**Abbildung 18: Ergebnis des CMM Algorithmus.** Durch die wiederholte Ausführung des Algorithmus konnte hier jede Klasse einem Cluster zugeordnet werden, wobei sich in den Clustern die größtmögliche Anzahl an Objekten einer Klasse befinden. Für Klasse 1 konnten zehn Objekte, für Klasse 2 acht Objekte, für Klasse 3 sechs Objekte und für Klasse 4 sieben Objekte richtig klassifiziert werden.

***Schritt 5: Summe aus den Maxima berechnen***

Um zu bestimmen, wie viele Objekte insgesamt richtig klassifiziert wurden, müssen die einzelnen Maxima, also die Anzahl an richtig klassifizierten Objekten jeder Klasse, aufaddiert werden:

$$Summe\ Maxima = max(Klasse\ 1) + \max(Klasse\ 2) + \max(Klasse\ 3) + \max(Klasse\ 4) = 10 + 8 + 6 + 7 = 31$$

***Schritt 6: Division durch die Gesamtzahl an Objekten***

Nun kann die Klassifikationsrate berechnet werden, in dem die Summe der richtig klassifizierten Objekte durch die Gesamtzahl der Objekte geteilt wird:

$$Klassifikationsrate = \frac{Summe\ Maxima}{Anzahl\ Objekte} = \frac{31}{46} = 0{,}67$$

In diesem Beispiel konnten also 67% der Objekte richtig klassifiziert werden.

Zum Vergleich der verschiedenen Kennzahlen für eine externe Clusteranalyse wurde ein Datensatz aus mehreren Krebszelllinien einer hierarchischen Clusteranalyse übergeben. Die Spektren wurden dazu wie zuvor beschrieben vorverarbeitet. Um die Entropie mit den anderen Kennzahlen vergleichbar zu machen, wurde sie normiert, so dass sie nur noch Werte zwischen null (0% richtig klassifiziert) und eins (100% richtig klassifiziert) annehmen kann:

$$Entropie_{norm} = 1 - \frac{Entropie}{ld(N)}$$

Dabei stellt *N* die Gesamtzahl der Objekte dar. Tabelle 9 zeigt die Klassifikationsergebnisse der verschiedenen Kennzahlen. Bis zu einer Anzahl von drei Klassen lieferten diese die gleichen Ergebnisse. Bei der Klassifikation von 11 Klassen fielen die Ergebnisse der Reinheit und der Entropie viel besser aus als das echte Klassifikationsergebnis. Einzig das CMM lieferte das gleiche Ergebnis wie die exakte Klassifikationsrate.

**Tabelle 9: Vergleich von Kennzahlen zur Bewertung einer unüberwachten Klassifikation verschiedener Krebszelllinien**

| **Zelllinie** | **Leukämie** | **GIST** | **Brustkrebs** | **gesamt** |
|---|---|---|---|---|
| **Anzahl Klassen** | **2** | **3** | **6** | **11** |
| Exakte Klassifikationsrate | 1,00 | 1,00 | 0,91 | 0,86 |
| CMM | 1,00 | 1,00 | 0,91 | 0,86 |
| Reinheit | 1,00 | 1,00 | 0,91 | 0,95 |
| Entropie | 1,00 | 1,00 | 0,93 | 0,97 |

Für einen weiteren Vergleich wurden Clusterergebnisse simuliert, welche aus zwei bis zehn Gruppen bestehen (Tabelle 10). Hier unterschieden sich die Ergebnisse der Reinheit und der Entropie auch bei einer kleinen Anzahl von Klassen. Erst ab einer Anzahl von neun Klassen wichen die Ergebnisse des CMM von der exakten Klassifikationsrate ab. Jedoch wurden sie nicht größer als die wahren Klassifikationsergebnisse. Der Grund dafür ist, dass bei der Berechnung des CMM, sobald eine Klasse einem Cluster zugeordnet wurde, die Objekte von anderen Klassen in diesem Cluster nicht weiter berücksichtig werden. Bei der Berechnung der Reinheit und der Entropie existiert ein derartiges Ausschlussverfahren nicht.

**Tabelle 10: Vergleich von Kennzahlen zur Bewertung einer unüberwachten Klassifikation von simulierten Clusterergebnissen**

| **Anzahl Klassen** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|
| Exakte Klassifikationsrate | 0,50 | 0,44 | 0,42 | 0,40 | 0,39 | 0,43 | 0,33 | 0,52 | 0,40 |
| CMM | 0,50 | 0,44 | 0,42 | 0,40 | 0,39 | 0,43 | 0,33 | 0,44 | 0,37 |
| Reinheit | 0,67 | 0,56 | 0,67 | 0,40 | 0,44 | 0,48 | 0,42 | 0,59 | 0,40 |
| Entropie | 0,08 | 0,28 | 0,54 | 0,37 | 0,47 | 0,54 | 0,53 | 0,67 | 0,56 |

Der Hauptvorteil einer Clusteranalyse mit Hilfe des CMM ist die geringe Rechenzeit, da im Gegensatz zur exakten Klassifikationsrate nicht alle möglichen Klassen- und Clusterkombinationen durchlaufen werden müssen (Abbildung 19). Auch der Speicherbedarf stellt hier keinen Engpass dar. Daher eignet sich das CMM sehr gut als Kriterium für eine objektive und vergleichende Bewertung zur systematischen Entwicklung von statistischen Modellen.

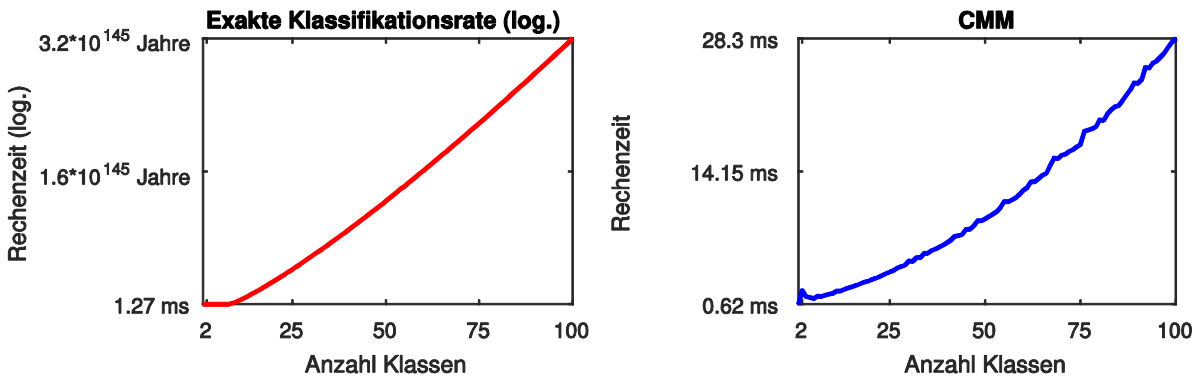


**Abbildung 19: Vergleich der Rechenzeit der exakten Klassifikationsrate und dem *Confusion Matrix Maximum (CMM)*.** Die Abbildung zeigt die Rechenzeit zur Berechnung der exakten Klassifikationsrate (rot) und dem CMM (blau). Für die Clusteranalyse von 100 Klassen benötigt das CMM ca. 21 ms. Die Rechenzeit der exakten Klassifikationsrate wurde ab einer Anzahl von 13 Klassen extrapoliert, da sie in endlicher Zeit nicht mehr berechenbar war.

## 4.3 Methode zur schnellen Klassifikation von Tumorzellen und Tumorzell-Subtypen

Eine schnelle und zuverlässige Klassifizierung humaner Tumorproben und die Identifikation des jeweiligen Tumorsubtypes spielen für die Entscheidung, welche personalisierte Behandlung einem Patienten zugeführt werden soll, eine zentrale Rolle. Aus diesem Grund wäre die Erforschung und Entwicklung systematischer Arbeitsprozesse zur automatischen Klassifikation humaner Karzinome hilfreich, um in Zukunft medizinisch relevante Merkmale schneller und zuverlässiger als bisher extrahieren zu können. Zunächst bedarf es dazu jedoch einer Methodenetablierung anhand eindeutig charakterisierter Zelllinien eines definierten Tumorsubtypes. Es ist bisher allerdings nicht bekannt, ob eine solche Klassifikation nur anhand von Linear-Positiv-Massenspektren möglich ist. Daher wurden in diesem Dissertationsprojekt systematisch verschiedene Methoden zur Merkmalsextraktion und Klassifikation getestet und miteinander verglichen (Abbildung 20).

Anhand eines Models aus massenspektrometrischen Signaturen von Brustkrebs-, GIST- und Leukämie-Zelllinien sollte ein optimaler Workflow entwickelt werden, welcher diese Zelllinien bestmöglich unterscheiden kann. Für jede Zelllinie wurden zwei Experimente / Datensätze angefertigt, welche wiederum aus acht technischen Replikaten bestanden. Die Massenspektren wurden wie zuvor beschrieben vorverarbeitet, wobei sie einer *Z-Score*-Normierung unterzogen wurden. Diese setzt den Mittelwert der Spektren auf null und die Varianz auf eins.

Um eine gute Klassifikation durchzuführen, ist es notwendig, im Vorfeld die hohe Dimension der Massenspektren zu reduzieren. Des Weiteren kann dadurch auch der Speicherbedarf und die Rechenzeit verringert werden, was gerade bei einer systematischen Methoden-entwicklung eine große Rolle spielt.

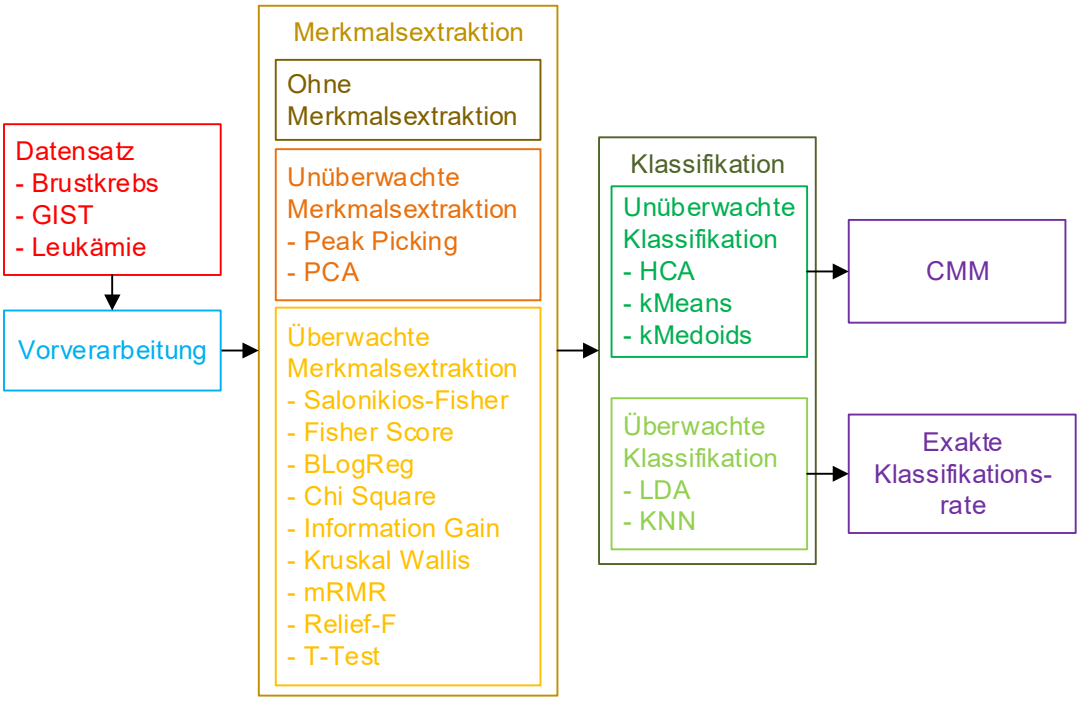


**Abbildung 20: Workflow zur systematischen Entwicklung einer optimierten Klassifikation verschiedener Krebs-Zelllinien.** Der Datensatz, bestehend aus Massenspektren verschiedener Krebszelllinien im linearen Positivionenmodus (Brustkrebs, GIST und Leukämie) wurde ***allen*** hier aufgelisteten Methoden zur Merkmalsextraktion übergeben. Im ersten Schritt wurde für eine Negativkontrolle keine Merkmalsextraktion durchgeführt. Für die unüberwachte Merkmalsextraktion wurden verschiedene *Peak Picking* Verfahren sowie eine Hauptkomponentenanalyse (engl. *Principal Component Analysis*, PCA) angewandt. Die überwachte Merkmalsextraktion erfolgte durch eine Diskriminanzanalyse nach Fisher, das Bayes'sche logistische Regressionsverfahren (BLogReg), einen Chi-Quadrat-Test (engl. *Chi Square*), einer Berechnung des Informationsgewinns (engl. *Information Gain*), einen Kruskal-Wallis-Test, die Minimierung der Redundanz mit gleichzeitiger Maximierung der Relevanz (engl. *minimum Redundancy Maximum Relevance*, mRMR) und einem T-Test. Im Anschluss wurde ***jeder*** dieser reduzierten Datensätze mit ***allen*** hier aufgeführten Methoden klassifiziert. Für die überwachte Klassifikation wurde eine hierarchische Clusteranalyse (engl. *Hierarchical Cluster Analysis*, HCA), der *k-Means* und der *k-Medoids* Algorithmus verwendet. Die überwachte Klassifikation wurde mit einer linearen Diskriminanzanalyse (eng. *Linear Discriminant Analysis*, LDA) und dem k-Nächste-Nachbarn-Algorithmus (engl. *k-Nearest-Neighbors*, KNN) durchgeführt. Die unüberwachten Klassifikationen wurden mit der neu entwickelten Kennzahl *Confusion Matrix Maximum (CMM)* und die überwachten Klassifikationen mit der exakten Klassifikationsrate bewertet.

Zunächst wurde hier das Standard-Verfahren zur Datenreduktion durch ein *Peak Picking* und der Hauptkomponentenanalyse (engl. *Principal Component Analysis*, PCA) untersucht. Das *Peak Picking* wurde dabei mit Hilfe des Softwarepakets *ClinProTools* durchgeführt. Die dort gefundenen Peaks wurden in einer Liste zusammengefasst und anschließend in MATLAB aus den Spektren extrahiert. Die Durchführung der PCA wurde anschließend in MATLAB realisiert (Abbildung 21). Dargestellt sind hier die ersten drei Hauptkomponenten. Die erste Hauptkomponente besitzt hier 31%, die zweite 27% und die dritte 13% der Gesamtvarianz. Die einzelnen Punkte repräsentieren jeweils ein Spektrum, wobei die rötlichen Punkte zu einer Brustkrebszelllinie, die bläulichen Punkte zu einer GIST-Zelllinie und die grünlichen Punkte zu einer Leukämie-Zelllinie gehören. Die Spektren einer Zelllinie aus den beiden Experimenten / Datensätzen wurden jeweils zusammengefasst. Die einzelnen Zelllinien ließen sich nur bedingt in verschiedene Cluster aufteilen, und es ergab sich insbesondere keine Trennung von Brustkrebs, GIST und Leukämie.

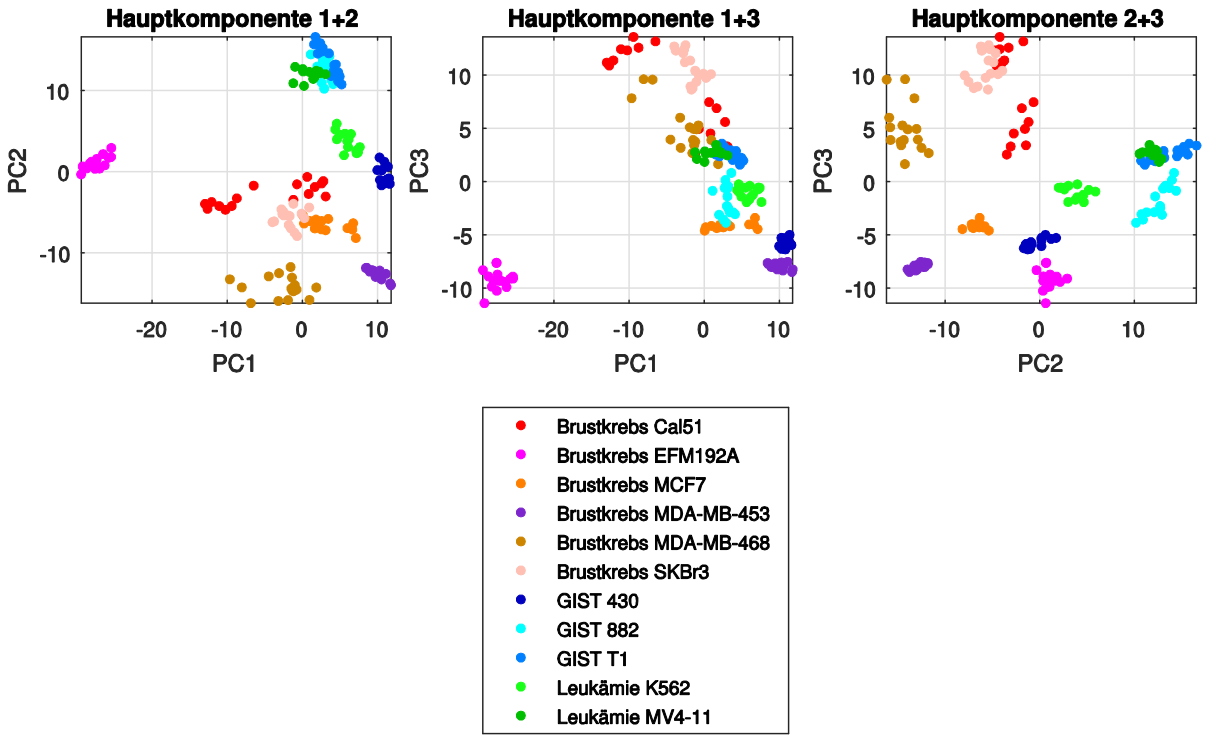


**Abbildung 21: Hauptkomponentenanalyse der verschiedenen Krebs-Zelllinien.** Die Abbildung zeigt das Ergebnis einer Hauptkomponentenanalyse der verschiedenen Krebs-Zelllinien. Dabei besitzt die erste Hauptkomponente 31%, die zweite 27% und die dritte 13% der Gesamtvarianz. Die Spektren der Brustkrebs-Zelllinien sind als rötliche Punkte, die der GIST-Zelllinien als bläuliche Punkte und die der Leukämie-Zelllinien als grünliche Punkte dargestellt. Die einzelnen Subtypen ließen sich nur bedingt in verschiedene Cluster aufteilen, und es ergab sich keine Trennung der einzelnen Tumortypen Brustkrebs, GIST und Leukämie.

### 4.3.1 Untersuchung allgemeiner Merkmalsextraktionsmethoden

Der Nachteil der Hauptkomponentenanalyse besteht darin, dass nach der Transformation sehr schwer Rückschlüsse auf die zur Unterscheidung relevanten *m/z*-Werte gezogen werden können, da jede Hauptkomponente aus einer Linearkombination *aller* ursprünglichen *m/z*-Werte besteht. Beim zweiten Standard-Verfahren, dem *Peak Picking*, bleibt die Information über die *m/z*-Werte zwar enthalten, jedoch können eventuell wichtige Peaks mit kleinen Intensitäten verloren gehen. Ein Grund dafür ist zum Beispiel, dass es zum Beispiel auf dem Mittelwertspektrum des Datensatzes durchgeführt wird (*ClinProTools*), wodurch kleine Peaks durch die Mittelung verschwinden können. Zudem macht das *Peak Picking* Verfahren keine Aussage über die Signifikanz der gefundenen Peaks. Aus diesem Grund wurden verschiedene, allgemeine Methoden zur Merkmalsextraktion untersucht, welche bisher noch nicht alle zur Datenreduktion von Massenspektren etabliert worden sind. Dazu wurden einige Algorithmen aus dem Softwarepaket der Arizona State University verwendet: *Fisher Score*, *BLogReg*, *Chi Square*, *Information Gain*, *Kruskal Wallis*, *mRMR*, *Relief-F* und *T-Test*.

Der Ansatz zur Merkmalsextraktion mit Hilfe der Fisher Diskriminanzanalyse stammt aus der Masterarbeit der Autorin dieser Dissertation (Salonikios, 2012) und ist bereits von Ruh et al. (Ruh et al., 2013) zur unvoreingenommenen Biomarkersuche verwendet worden, um anhand von MS Imaging potentiell krankheitsrelevante Metaboliten aufspüren zu können. Dieser Ansatz wurde in diesem Dissertationsprojekt weiterentwickelt und für die konkreten biomedizinischen Fragestellungen optimiert. Im Folgenden wird diese Methode als *Fisher Score Optimized* bezeichnet. Diese Implementierungen des *Fisher Score* der Arizona State University und des *Fisher Score Optimized* führten zu den gleichen Klassifikations-ergebnissen, unterschieden sich jedoch im Aufbau des Algorithmus. Daher wurden sie im Folgenden zum Vergleich weiterer Leistungsmerkmale als eigenständige Methoden behandelt.

Für die Anwendung dieser Merkmalsextraktionsmethoden war es sehr wichtig, dass die Peak Maxima genau übereinander liegen. Die in diesem Dissertationsprojekt entwickelte Methode zur Ausrichtung der Massenspektren verringerte zwar die Standardabweichung der Peak Maxima erheblich, doch lagen sie noch nicht exakt übereinander. Aus diesem Grund wurde als Grundlage für die Merkmalsextraktion ein Verfahren entwickelt, welches ein einfaches und rudimentäres *Peak Picking* durchführt und die Peak Maxima der einzelnen Spektren einem einzigen *m/z*-Wert zuordnet. Da die Auflösung der Massenspektren im hohen Massenbereich bei einer Messung im Linearmodus nicht mehr so gut ist, besitzen die Peaks dort eine relativ große Breite, weshalb der exakte *m/z*-Wert allein mit informations-technischen Methoden nicht bestimmbar ist. Um sicherzustellen, dass keine Peaks mit

kleinen Intensitäten verloren gehen, wurde statt dem Mittelwertspektrum des Datensatzes die *Skyline* verwendet. Des Weiteren wurde nur ein einziges Kriterium zur Peak Detektion eingesetzt: Es musste sich bei dem untersuchten Datenpunkt um ein lokales Maximum handeln.

Abbildung 22 zeigt den Vergleich dieser Methode mit dem *Peak Picking* Verfahren von ClinProTools und MATLAB (hier wurde die Funktion *mspeaks* aus der *Bioinformatics Toolbox* verwendet). Es wurden wesentlich mehr „Peaks“ gefunden als mit den anderen beiden Methoden. An dieser Stelle konnte jedoch noch keine Aussage darüber getroffen werden, bei welchen dieser lokalen Maxima es sich tatsächlich um einen Peak handelte. Dies sollte die Aufgabe der nachfolgenden statistischen Analyse sein. Da in einem „normalen“ Linear-Positiv-Spektrum üblicherweise nur 100-150 Signale zu sehen sind, war jedoch sichergestellt, dass kein wichtiger Peak verloren gegangen ist.

Im Anschluss daran wurden in den Einzelspektren die korrespondierenden Peak Maxima gesucht (Abbildung 23). Aufgrund der zuvor durchgeführten Ausrichtung der Spektren lagen sie bereits sehr nah beieinander, was die Wahrscheinlichkeit einer unbeabsichtigten Falschdetektion sehr verringerte.

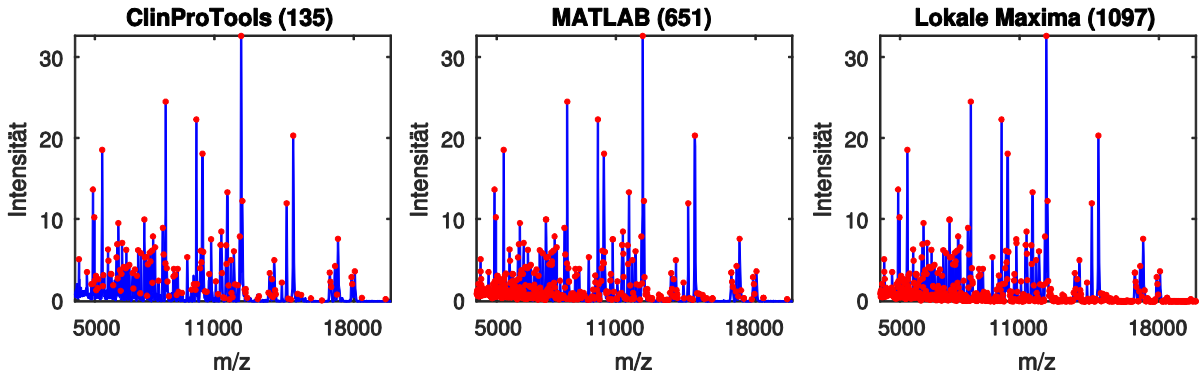


**Abbildung 22: Vergleich verschiedener *Peak Picking* Verfahren als Grundlage für eine statistische Analyse.** Als Basis wurde hier die *Skyline* des Krebszelllinien-Datensatzes verwendet. Mit dem *Peak Picking* Verfahren von ClinProTools wurden 135 Peaks gefunden. Die MATLAB Funktion *msbackadj* aus der *Bioinformatics Toolbox* lieferte 654 Peaks. Ob es sich dabei um eine vollständige Peakliste handelt, ist nicht bekannt. Insgesamt wurden 1097 lokale Maxima detektiert (links). Da hier wirklich *jedes* lokale Maximum ermittelt wurde, ist davon auszugehen, dass kein Peak verloren gegangen ist. Die Signifikanz dieser Maxima kann nun anhand statistischer Methoden bewertet werden.

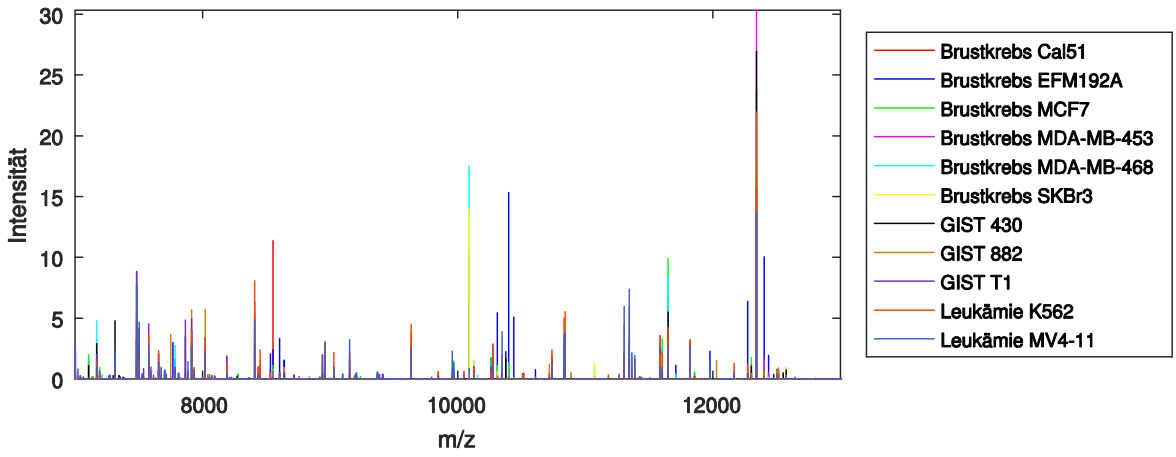


**Abbildung 23: Zentrierung der Peaks anhand der gefundenen lokalen Maxima am Beispiel des Krebszelllinien-Datensatzes (Zoom).** Für eine statistische Untersuchung der Peaks müssen die Peak Maxima exakt übereinander liegen. Als Zentrum der Peaks wurden hier die m/z-Werte der lokalen Maxima der Skyline verwendet. Dargestellt ist hier der Massenbereich 7000 Da – 13.000 Da.

Im nächsten Schritt wurden die Diskriminanten der einzelnen Merkmalsextraktionsmethoden für den Krebszelllinien-Datensatz berechnet (Abbildung 24). Diese geben für jedes Merkmal ein Gewicht an. Die Merkmale bestehen in diesem Fall aus den Peaks (lokalen Maxima) der Massenspektren der Krebszelllinien. Je höher dieses Gewicht ist, desto besser ist dieser Peak dazu geeignet, die verschiedenen Krebszelllinien voneinander zu trennen.

Um zunächst herauszufinden, welche Peaks am besten dazu geeignet sind, die verschiedenen Tumortypen (Brustkrebs, GIST und Leukämie) auseinanderzuhalten, wurden die Spektren der einzelnen Subtypen zusammengefasst, so dass der Datensatz im Prinzip nur noch aus drei Klassen bestand.

Die Ergebnisse aus Abbildung 24 zeigten, dass die Methoden *Fisher Score Optimized*, *Fisher Score*, *Relief-F* und *T-Test* deutlich einige herausragende Massenbereiche aufwiesen. Die Diskriminanten der Methoden *Chi Square* und *Information Gain* besaßen deutlich sehr viele Peaks mit einem hohen Gewicht, die des *Kruskal Wallis* Tests hingegen deutlich sehr wenige. Die Diskriminante des *mRMR* Verfahrens enthielt nur im oberen Massenbereich hohe Gewichte. Die Untersuchung, welche und wie viele der hier gefunden Peaks sich zur Trennung der Zelllinien eignen, ist Bestandteil des nächsten Kapitels.

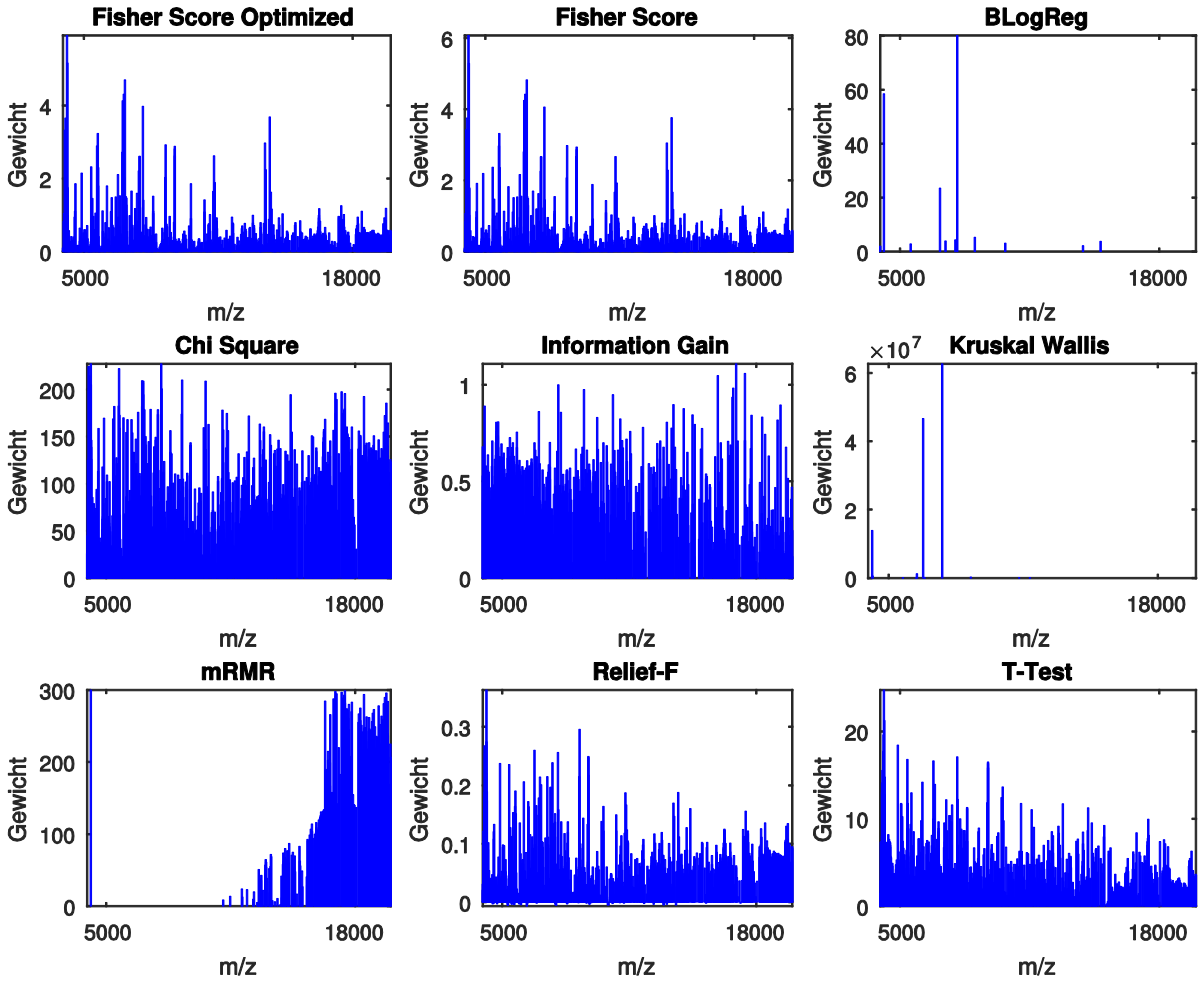


**Abbildung 24: Diskriminanten verschiedener Methoden zur Merkmalsextraktion aus den Krebs-Zelllinien.** Die Abbildung zeigt die Diskriminanten der verschiedenen Merkmalsextraktionsmethoden aus dem Softwarepakte der Arizona State University sowie die Methode *Fisher Score Optimized*. Die Varianten *Fisher Score Optimized*, *Fisher Score*, *Relief-F* und *T-Test* stellen graphisch das beste Ergebnis dar, da hier deutlich einige herausragende Massenbereiche zu erkennen sind.

Im Folgenden wurden zunächst die Rechenzeiten der verschiedenen Merkmalsextraktionsmethoden untersucht, da diese bei einer systematischen Methodenentwicklung eine große Rolle spielen. Dazu wurde für jede Methode die Rechenzeit zur Erstellung der Diskriminante bestimmt (Tabelle 11). Die Berechnung des *Fisher Score Optimized* benötigte mit Abstand die geringste Rechenzeit, gefolgt vom *T-Test*, welcher ungefähr die achtfache Rechenzeit im Vergleich Methode benötigte. Die Rechenzeiten vieler Methoden lagen nicht mehr im Millisekunden-Bereich und könnten daher zu einer nicht mehr annehmbaren Gesamtrechenzeit bei einer systematischen Methodenentwicklung führen.

**Tabelle 11: Rechenzeit zur Berechnung der Diskriminanten verschiedener Methoden zur Merkmalsextraktion**

| Methode | Rechenzeit [sec.] |
|---|---|
| Fisher Score Optimized | 0,05 |
| T-Test | 0,17 |
| BLogReg | 0,58 |
| Fisher Score | 0,66 |
| Information Gain | 1,09 |
| Chi Square | 1,13 |
| Relief-F | 3,13 |
| Kruskal Wallis | 5,12 |
| mRMR | 9,18 |

Die enorme Zeitdifferenz der beiden Berechnungen des *Fisher Scores* ist auf die unterschiedliche Art der Implementierung zurückzuführen. Die Funktion *fsFisher* (Arizona State University) besteht hauptsächlich aus *for*-Schleifen, welche in MATLAB sehr viel Rechenzeit benötigen. Die in diesem Dissertationsprojekt entwickelte MATLAB Funktion hingegen stellt die effizientere Umsetzung dar, da sie weniger *for*-Schleifen, dafür aber einige Matrix-orientierte Befehle nutzt (Abbildung 25). Diese erlauben es, Probleme kompakt zu formulieren und lösen.

```
1     function FD = GetFisher(data)
2
3     numberOfValues = length(data.xValues);
4
5     matrix = GetMatrix(data, 'SingleSpectra');
6     M = mean(matrix);
7
8     m_class = zeros(data.numberOfTypes, numberOfValues);
9     v_class = zeros(data.numberOfTypes, numberOfValues);
10    n_class = zeros(data.numberOfTypes, 1);
11
12    for i = 1 : data.numberOfTypes
13        matrix = GetMatrix(data.types{i}, 'Type');
14        m_class(i,:) = mean(matrix);
15        v_class(i,:) = var(matrix);
16        n_class(i) = data.types{i}.numberOfSpectra;
17    end
18
19    FD = zeros(1, numberOfValues);
20
21    for i = 1 : data.types{1}.spectra{1}.numberOfValues
22        FD(i) = sum(n_class.*(m_class(:,i)-M(i)).^2) /
23                 sum(n_class.*v_class(:,i));
24    end
```

**Abbildung 25: MATLAB Codes zur verbesserten Berechnung der Fisher Diskriminante.** Gegenüber früheren Implementierungen (Duda et al., 2001) wurden hier einige Berechnungen durch Matrix-orientierte Befehle umgesetzt, was zu einer erheblichen Verkürzung der Rechenzeit führte.

Die Implementierung des *T-Tests* im Softwarepaket der Arizona State University erfolgte bereits durch einige Matrix-orientierte Befehle, so dass hier in Bezug auf die Rechenzeit kein Optimierungsbedarf mehr bestand.

### 4.3.2 Methodik zur Bestimmung der optimalen Anzahl an Merkmalen

Durch Berechnung der Diskriminanten können den einzelnen Peaks Gewichte zugeordnet werden, welche ein Maß dafür darstellen, wie gut sich dieser Peak für die Unterscheidung verschiedene Zelllinien eignet. Sie geben jedoch keine Auskunft darüber, wie viele und welche Kombinationen dieser Peaks für eine optimale Klassifikation benötigt werden. Daher wurde hier ein neues Verfahren herausgearbeitet, welches die optimale Peak-Anzahl zur Unterscheidung der Krebszelllinien ermittelt.

Der erste Schritt dieses Verfahrens besteht darin, die Peaks nach ihren Gewichten zu sortieren. Da die Diskriminanten auf Basis der Tumortypen (Brustkrebs, GIST und Leukämie) erstellt worden sind, stehen die Peaks für deren Unterscheidung ganz oben in der Peak-Liste. Um nun herausfinden, wie viele dieser Peaks benötigt werden, um auch die Tumorsubtypen auseinanderzuhalten, werden die Peaks nacheinander einem Klassifikator übergeben, welcher eine Klassifikation auf Grundlage der elf verschiedenen Krebszelllinien durchführt. Begonnen wird dabei zunächst nur mit dem ersten Peak aus der Liste (dieser besitzt das höchste Gewicht). Bei jedem Durchlauf wird die Peak-Anzahl sukzessive um eins erhöht, bis die Liste abgearbeitet ist.

Im Fall einer unüberwachten Klassifikation (Abbildung 26) wurde jedes der Ergebnisse mit dem CMM bewertet. Zum Schluss entstand dadurch eine Kurve, welche für jede Peak-Anzahl das entsprechende Klassifikationsergebnis anzeigte. Das Maximum dieser Kurve entsprach dann der optimalen Anzahl an Peaks für die Klassifikation. Diese Peaks sollten dann im Idealfall sowohl die Tumortypen als auch die Tumorsubtypen trennen können.

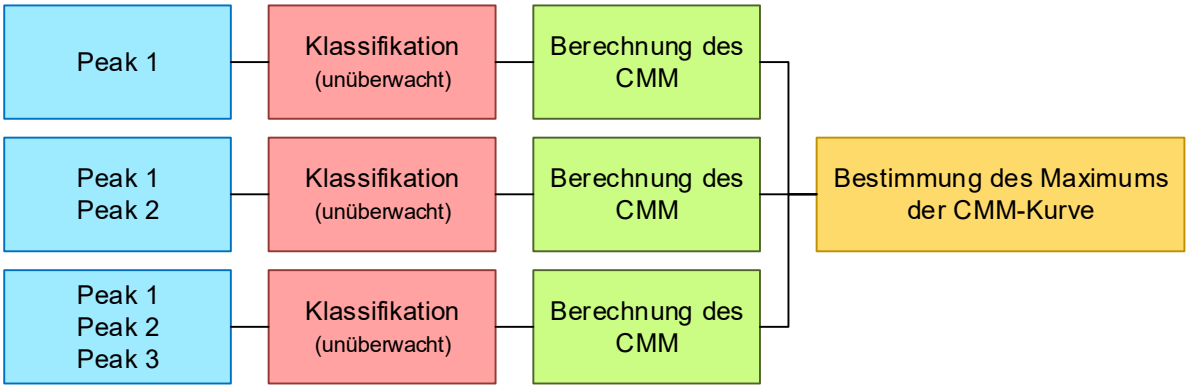


**Abbildung 26: Vereinfachte Darstellung des Prinzips zur Bestimmung der optimalen Peak-Anzahl für die unüberwachte Klassifikation.** Durch sukzessive Erhöhung der Peak-Anzahl (aus einer sortierten Peak-Liste), welche einem unüberwachten Klassifikator übergeben werden, kann eine CMM-Kurve erstellt werden, welche die Anzahl an Peaks in Abhängigkeit darstellt. Der maximale Wert dieser Kurve entspricht dann der optimalen Peak-Anzahl für die Klassifikation.

Bei der überwachten Klassifikation ist es üblich, den Datensatz zunächst in Trainingsdaten und Testdaten aufzuteilen. Dies wurde hier durch einen Zufallsgenerator realisiert, welcher dem Trainings- und Testdatensatz ungefähr die gleiche Anzahl an Spektren pro Zelllinie zuordnete. In der Trainingsphase (Abbildung 27) wurden dann nach und nach alle Peaks der sortierten Peak-Liste aus dem Trainingsdatensatz extrahiert, mit welchen für jeden Durchlauf ein Klassifikationsmodell erstellt wurde. In der Testphase (Abbildung 28) wurden dieselben Peaks aus dem Testdatensatz extrahiert und dem entsprechenden Klassifikationsmodell übergeben. Die einzelnen Ergebnisse konnten hier mit der exakten Klassifikationsrate bewertet werden, da es sich um eine überwachte Klassifikation handelte, bei welcher die Klassenzugehörigkeit im Vorfeld bekannt war. Die Bestimmung der optimalen Peak-Zahl erfolgte dann analog durch Bestimmung des Maximums der einzelnen Klassifikationsraten.

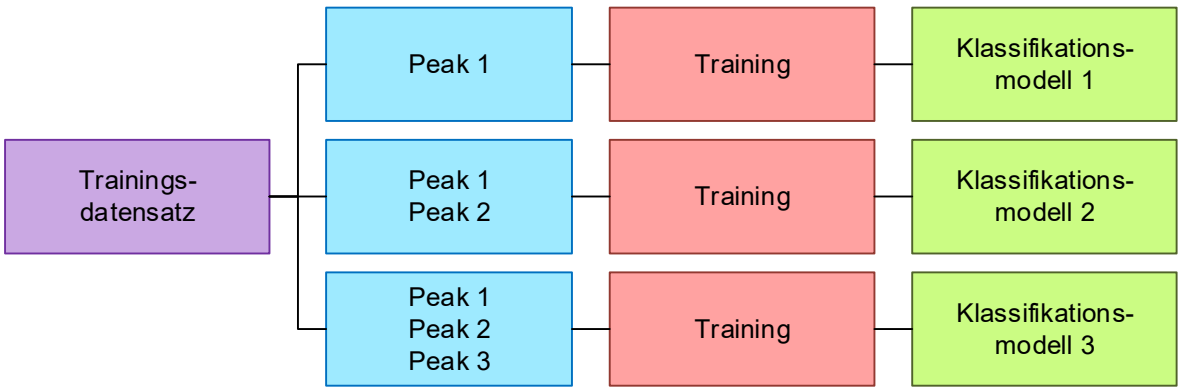


**Abbildung 27: Vereinfachte Darstellung des Prinzips zur Bestimmung der optimalen Peak-Anzahl für die überwachte Klassifikation (Trainingsphase).** Nach und nach werden alle Peaks aus dem Trainingsdatensatz extrahiert und dem Klassifikator übergeben. Somit entsteht für jeden Durchlauf ein eigenes Klassifikationsmodell.

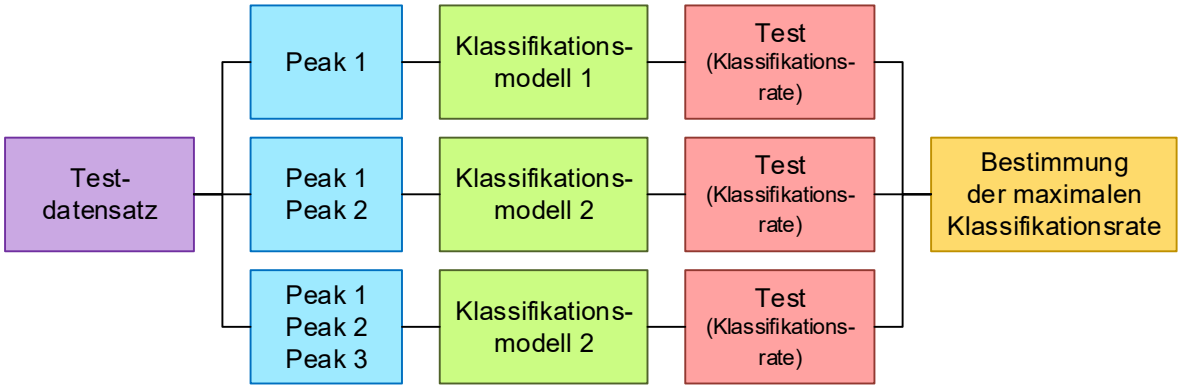


**Abbildung 28: Vereinfachte Darstellung des Prinzips zur Bestimmung der optimalen Peak-Anzahl für die überwachte Klassifikation (Testphase).** Nach und nach werden alle Peaks aus dem Testdatensatz extrahiert und dem in der Trainingsphase erstellten zugehörigen Klassifikationsmodell übergeben. Die Bewertung der einzelnen Ergebnisse erfolgt durch die Berechnung der exakten Klassifikationsrate. Dir optimale Peak-Anzahl entspricht dann der maximalen Klassifikationsrate.

Abbildung 29 zeigt beispielhaft die Bestimmung der optimalen Peak-Anzahl für die Trennung der Krebszelllinien durch eine hierarchische Clusteranalyse. Zuvor wurden die Peaks mit Hilfe des *Fisher Score Optimized* gewichtet. Mit 110 Peaks konnten hier sowohl die Tumortypen als auch die Tumorsubtypen auseinandergehalten werden. Des Weiteren wird hier verdeutlicht, dass mit größerer Peak-Anzahl die Klassifikationsergebnisse wieder schlechter wurden. Die Peaks, welche ganz unten in der Peak-Liste standen, wurden zur Trennung nicht benötigt und stellten daher wohl keine Tumor-relevanten Peaks dar. Die hier entwickelte Methodik konnte also erfolgreich zur selektiven Datenreduktion mit gleichzeitiger Optimierung der Klassifikation eingesetzt werden.

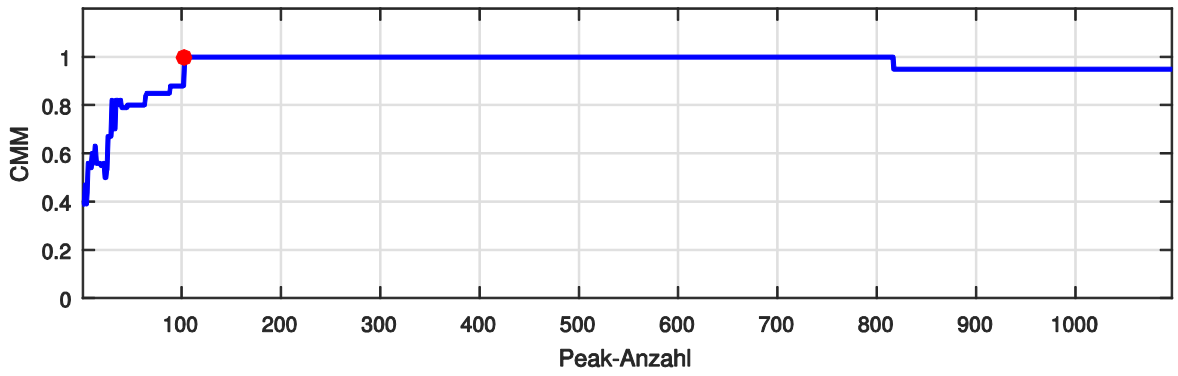


**Abbildung 29: Beispiel zur Bestimmung der optimalen Peak-Anzahl für die Klassifikation der Krebszelllinien.** Die Abbildung zeigt die CMM-Kurve einer hierarchischen Clusteranalyse basierend auf der neuen Methodik zur Bestimmung der optimalen Peak-Anzahl für die Klassifikation. Zuvor wurden die Peaks mit Hilfe des *Fisher Score Optimized* gewichtet. Das Maximum der CMM-Kurve hier konnte einer Peak-Anzahl von 110 zugeordnet werden. Mit größerer Peak-Anzahl wurden die Klassifikationsergebnisse wieder schlechter.

### 4.3.3 Validierung der Klassifikation durch Anwendung der neuen Kennzahl

Um herauszufinden, welche Methode zur Merkmalsextraktion zusammen mit welchem Klassifikationsverfahren die besten Ergebnisse bei der Unterscheidung der Krebszelllinien erzielt, wurde jede der zuvor untersuchten Merkmalsextraktionsmethoden einmal mit jedem der in Abbildung 20 dargestellten Klassifikationsverfahren kombiniert. Als Vergleichsgrößen wurden zusätzlich Klassifikationen ohne Merkmalsextraktion, PCA und *Peak Picking* durchgeführt. Durch Verwendung der neu entwickelten Kennzahl zur Bewertung einer unüberwachten Klassifikation benötigte dieser systematische Test, welcher insgesamt aus ungefähr 50.000 Durchläufen bestand, nur ca. 20 min. an Rechenzeit.

Für die unüberwachte Klassifikation wurden die Klassifikationsverfahren HCA, *k-Means* und *k-Medoids* verwendet. Die HCA wurde mit der euklidischen Distanz und der Verknüpfungsmethode *ward* (Ward Jr, 1963) durchgeführt. Die Klassifikatoren *k-Means* und *k-Medoids* besitzen bei jedem Durchlauf zufällige Startparameter. Deshalb wurden sie jeweils zehn Mal durchgeführt und die Ergebnisse gemittelt. Die überwachte Klassifikation wurde mit Hilfe einer linearen Diskriminanzanalyse (engl. *Linear Discriminant Analysis*, LDA) und dem k-Nächste-Nachbarn Algorithmus (KNN) durchgeführt. Die Daten wurden hier ebenfalls mit einem Zufallsgenerator in einen Trainings- und Testdatensatz aufgeteilt.

Tabelle 12 zeigt eine Übersicht der Klassifikationsergebnisse. Wie bereits dargestellt, konnte die Trennung der elf Krebszelllinien durch eine PCA oder das *Peak Picking* Verfahren nicht optimiert werden. Dies zeigt aber jedoch noch einmal deutlich, dass für die Selektion der Peaks statistische Verfahren notwendig sind. Mit Hilfe der allgemeinen Methoden zur Merkmalsextraktion konnten die Klassifikationsergebnisse in den meisten Fällen sehr deutlich gegenüber den nicht reduzierten Daten verbessert werden.

Die unüberwachten Klassifikationsmethode *k-Means* führte hier zu den schlechtesten Ergebnissen. In keinem Fall wurde eine eindeutige Unterscheidung der Zelllinien erreicht. Die Methoden HCA und *k-Medoids* konnten mit Hilfe einiger Merkmalsextraktionsmethoden eine Klassifikation von 100% erzielen. Die HCA stellte jedoch das stabilere Verfahren dar, da ihre Ergebnisse keiner Standardabweichung unterliegen. Zudem benötigt sie weniger Rechenzeit, da sie keine zufälligen Startparameter besitzt und deshalb nur einmal ausgeführt werden muss.

Eine Analyse mit der LDA ohne vorherige Merkmalsextraktion überschritt die Speicherkapazität des Rechners, wodurch es zum Programmabbruch kam. Für einen nicht reduzierten Datensatz stellt somit der KNN Algorithmus die bessere Variante dar. Beide Verfahren erzielten auch ohne statistische Analyse meistens sehr gute Ergebnisse. Dies führt zu der Schlussfolgerung, dass sie aufgrund der insgesamt geringen Anzahl an Spektren

trotz Aufteilung des Datensatzes in Trainings- und Testdaten „auswendig“ gelernt haben. Für eine sichere Klassifikation ist daher in jedem Fall eine unüberwachte Methode besser geeignet.

**Tabelle 12: Klassifikationsergebnisse der verschiedenen Methoden zur Merkmalsextraktion und Klassifikation**

| Klassifikations-methode | HCA | k-Means | k-Medoids | LDA | KNN |
|---|---|---|---|---|---|
| Ohne Merkmalsextraktion | 0,86 | 0,83 ± 0,06 | 0,91 ± 0,05 | "Out of memory" | 1,00 |
| PCA | 0,86 | 0,80 ± 0,06 | 0,83 ± 0,04 | 0,92 | 1,00 |
| Peak Picking ClinProTools | 0,95 | 0,81 ± 0,05 | 0,95 ± 0,00 | 1,00 | 1,00 |
| Fisher Score Optimized | 1,00 | 0,93 ± 0,06 | 1,00 ± 0,00 | 1,00 | 1,00 |
| Fisher Score | 1,00 | 0,93 ± 0,08 | 1,00 ± 0,00 | 1,00 | 1,00 |
| BLogReg | 1,00 | 0,92 ± 0,08 | 0,99 ± 0,00 | 1,00 | 1,00 |
| Chi Square | 0,42 | 0,42 ± 0,04 | 0,44 ± 0,03 | 0,55 | 0,52 |
| Information Gain | 0,43 | 0,45 ± 0,02 | 0,44 ± 0,02 | 0,56 | 0,34 |
| Kruskal Wallis | 1,00 | 0,93 ± 0,04 | 1,00 ± 0,00 | 1,00 | 1,00 |
| mRMR | 0,95 | 0,90 ± 0,06 | 0,95 ± 0,06 | 1,00 | 1,00 |
| Relief-F | 1,00 | 0,93 ± 0,06 | 1,00 ± 0,00 | 1,00 | 1,00 |
| T-Test | 1,00 | 0,92 ± 0,08 | 1,00 ± 0,00 | 1,00 | 1,00 |

Die besten Klassifikationsergebnisse lieferten die Merkmalsextraktionsmethoden *Fisher Score Optimized*, *Fisher Score*, *Kruskal Wallis* und *Relief-F*. Wie bereits dargestellt, ist davon die Methode *Fisher Score Optimized* in Bezug auf die Rechengeschwindigkeit die stärkste Methode. Folgende Abbildung zeigt noch einmal zusammengefasst die Arbeitsschritte für eine optimale Zelllinien-Klassifikation.


**Abbildung 30: Arbeitsschritte für eine optimale Zelllinien-Klassifikation.** Zunächst werden die Massenspektren auf geeignete Weise vorverarbeitet, so dass die Merkmalsextraktion auf den unreduzierten Daten erfolgen kann. Dies beinhaltet eine Basislinienkorrektur, eine Signalglättung mit dem Savitzky Golay Filter, die Berechnung einer gemeinsamen Massenachse durch eine lineare Interpolation, die Ausrichtung der Spektren mit Hilfe der Korrelationsfunktion, eine Histogramm basierte Rauschunterdrückung und eine Zentrierung der Peaks anhand lokaler Maxima. Für die Merkmalsextraktion wird eine Varianzanalyse mit der *Fisher Score Optimized* Methode durchgeführt. Im Anschluss daran wird die optimale Peak-Anzahl mit der in dieser Arbeit dazu entwickelten Methode bestimmt. Für eine sichere Zelllinien-Klassifikation wird die unüberwachte Methode HCA verwendet. Die Bewertung erfolgt mit dem *Confusion Matrix Maximum*.

Abbildung 31 zeigt das Dendrogamm der Klassifikation der Krebszelllinien einmal ohne Merkmalsextration und einmal mit Datenreduzierung durch die Methode *Fisher Score Optimized*. Ohne Merkmalsextration konnten die Tumorsubtypen GIST T1 und GIST 882 nicht auseinandergehalten werden. Des Weiteren wurde die Brustkrebszelllinie Cal51 in zwei verschiedene Cluster aufgeteilt. Eventuell könnte es sich hierbei um eine Trennung der Zelllinie aufgrund der beiden durchgeführten Experimente handeln. Die Brustkrebs-Zelllinie EFM192A bildet hier ein eigenes Cluster, deutlich getrennt von allen anderen Zelllinien Durch Anwendung der in dieser Arbeit entwickelten Methode können die elf Zelllinien eindeutig getrennt werden. Es konnte sowohl unter den drei Tumortypen Brustkrebs, GIST und Leukämie unterschieden werden, als auch die einzelnen Subtypen den entsprechenden Tumortypen zugeordnet werden. Bei der Klassifikation der GIST-Zelllinien ist sogar die Trennung zwischen der vermeintlich Imatinib-resistenten Zelllinie (GIST 430) und den beiden responsiven Zelllinien (GIST T1 und GIST 882) erhalten geblieben. Die beiden Brustkrebszelllinien Cal51 und MDA-MB-468, welche beide der Tumor-Subklassifizierung „Triple Negativ“ zugeordnet werden können, befinden sich ebenfalls in einem Teil-Cluster.

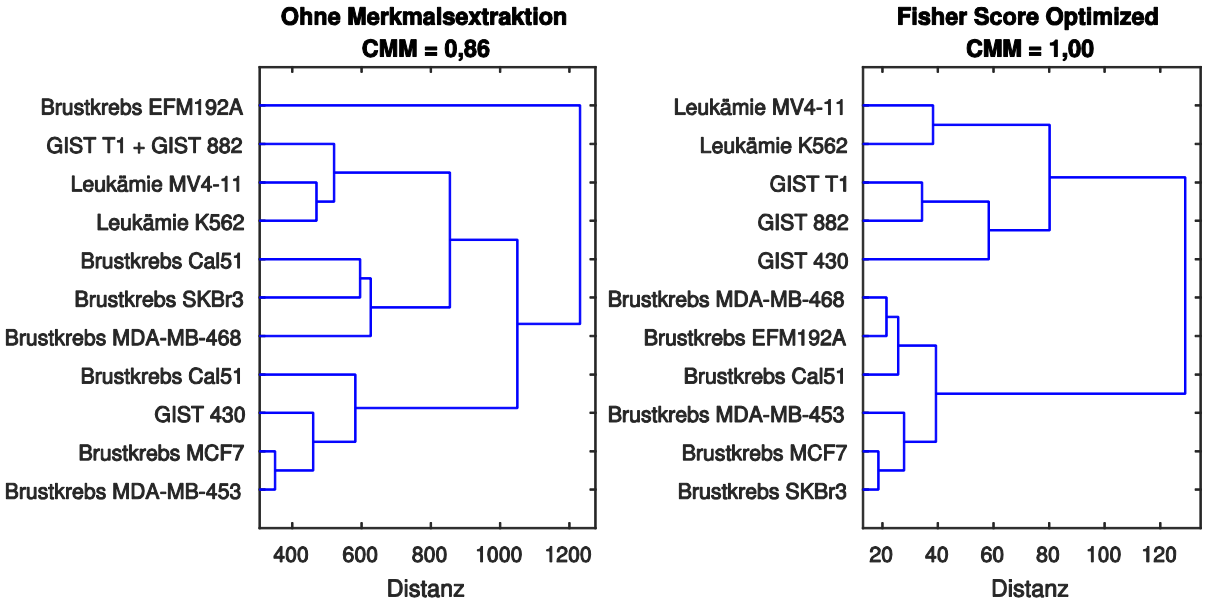


**Abbildung 31: Hierarchische Clusteranalyse der Krebszelllinien ohne und mit Merkmalsextration.** Mit Hilfe des *Fisher Score Optimized* und der in dieser Arbeit entwickelten Methode zur Bestimmung der optimalen Peak-Anzahl zur Klassifikation konnten die drei Tumortypen Brustkrebs, GIST und Leukämie eindeutig getrennt werden. Des Weiteren konnten die Tumorsubtypen eindeutig den entsprechenden Tumortypen zugeordnet werden.

## 4.4 Methode zur automatisierten Bestimmung von Wirkstoffeffekten

Molekular-gerichtete Medikamente besitzen meist eine bessere Wirksamkeit und ersparen dem Patienten unnötige Nebenwirkungen und Zeit, wenn schon gleich zu Beginn der Behandlung eine geeignete Therapie eingesetzt wird. Dabei spielen Biomarker, oder allgemein molekulare Signaturen, eine zentrale Rolle. Prädiktive Biomarker ermöglichen es, das Ansprechen eines individuellen Patienten auf spezielle Wirkstoffe vorherzusagen. Pharmakodynamische Biomarker erlauben es, die Wirksamkeit der molekular-gerichteten Therapeutika anzuzeigen.

Mit Hilfe einer Konzentrations-Wirkungskurve kann die Wirkung eines Medikaments dargestellt werden. Sie hat häufig einen sigmoidalen Verlauf, wobei auf der x-Achse die Wirkstoffkonzentrationen und auf der y-Achse das Messsignal, zum Beispiel die Konzentration eines Biomarkers, aufgetragen wird. In dieser Kurve kann die mittlere, effektive Konzentration abgelesen werden, bei welcher das Medikament die gewünschte Wirkung zeigt. Es existieren einige Softwarepakete, wie zum Beispiel GraphPad Prism, welche eine Darstellung solcher Kurven ermöglichen. Eine Software, welche einen MS-Datensatz automatisch nach Proteinen durchsucht, die eine Wirkung auf ein Medikament zeigen, existiert bisher nicht. Die Intensitäten einzelner m/z-Werte müssen per Hand in die entsprechende Analysesoftware übertragen werden. Des Weiteren gibt es kein Verfahren, welches die Ergebnisse verschiedener Regressionsanalysen miteinander vergleicht.

Folgende Abbildung zeigt beispielhaft einen K562-Zelllinien-Datensatz. Die CML-Zellen wurden vor der MS-Messung mit dem selektiven Wirkstoff Imatinib in verschiedenen Konzentrationen behandelt. Imatinib wird derzeit zur Behandlung der chronisch myeloischen Leukämie eingesetzt, weshalb zu erwarten ist, dass in der Abbildung eine Wirkung zu erkennen ist. Die Spektren werden hier in Form einer *Heatmap* dargestellt, bei welcher die Werte mit Hilfe einer Falschfarbendarstellung repräsentiert werden. So können in einer großen Datenmenge prägnante Werte / Bereiche schnell erfasst werden. Die Zeilen der *Heatmap* stellen hier die einzelnen Massenspektren dar, wohingegen in den Spalten der Intensitätsverlauf eines m/z-Wertes in Abhängigkeit der Wirkstoffkonzentration zu sehen ist. Es sind deutlich einige Bereiche zu erkennen, in welchen die Intensitäten mit zunehmender Wirkstoffkonzentration zu- oder abnehmen. Ein Ziel dieses Dissertationsprojektes war es, diese Bereiche automatisch zu erkennen, zu bewerten und die mittlere, effektive Konzentration (EC50) zu berechnen, bei welcher die Signalintensität für bestimmte Proteine eine Wirkung auf das Medikament zeigt.

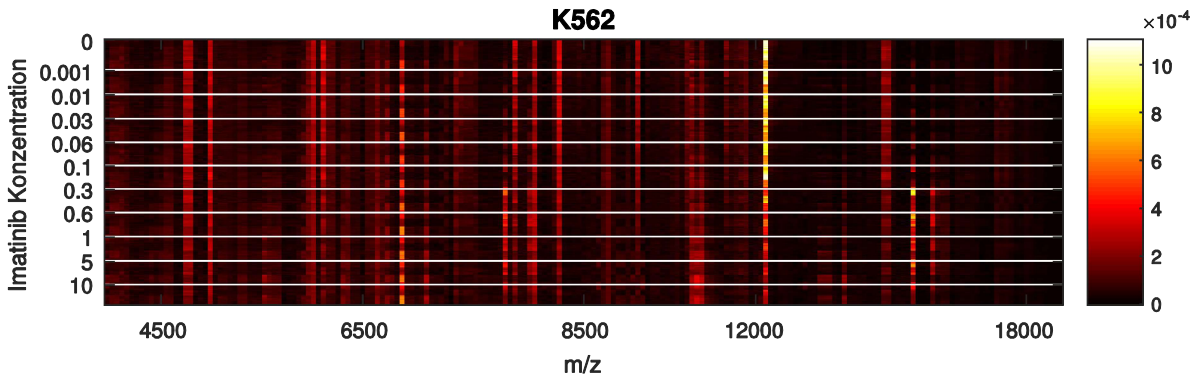


**Abbildung 32: *Heatmap* der wirkstoffbehandelten K562 Zelllinie.** Die Zelllinien K562 wurde mit dem selektiven Wirkstoff Imatinib in verschiedenen Konzentrationen behandelt. Es sind deutlich Bereiche zu erkennen, in welchen der Wirkstoff zu einer Zunahme oder Abnahme der Intensität führt.

Für die Methodenentwicklung wurde wieder das Modell aus massenspektrometrischen Signaturen von Brustkrebs-, GIST- und Leukämie-Zelllinien verwendet, wobei sie mit unterschiedlichen Wirkstoffen in verschiedenen Konzentrationen behandelt wurden. Es wurden pro Zelllinie und Wirkstoff drei Experimente / Datensätze mit jeweils acht technischen Replikaten angefertigt. Die Massenspektren wurden wie zuvor beschrieben vorverarbeitet. Hier wurden sie jedoch einer TIC-Normierung unterzogen, da für die Regression einer Konzentrations-Wirkungskurve die Intensitäten nicht negativ sein sollten (nach einer *Z-Score* Normierung können auch negative Werte auftreten).

### 4.4.1 Methode zur Erstellung von Konzentrations-Wirkungskurven

Abbildung 33 zeigt den typischen Verlauf einer fallenden Konzentrations-Wirkungskurve. Für eine übersichtlichere Darstellung ist es üblich, die Konzentrationen logarithmisch auf die x-Achse aufzutragen. Die Intensitätswerte werden durch die mittlere Intensität einer Konzentration dargestellt. Des Weiteren wurden sie auf die Kontrolle (Intensität der Spektren aus Zellen ohne Wirkstoffbehandlung) normiert. Um sie prozentual im Verhältnis zur Kontrolle angeben zu können, wurden sie zusätzlich mit dem Faktor 100 multipliziert. Zur Berechnung einer Konzentrations-Wirkungskurve werden folgende Parameter benötigt (Wenner et al., 2011): die minimale und maximale Intensität, die Steigung und die mittlere, effektive Konzentration (EC50). Sie kann nach folgender Formel berechnet werden:

$$Y = Bottom + \frac{(Top - Bottom)}{1 + 10^{(LogEC_{50} - X) * HillSlope}}$$

Die minimale und maximale Intensität können direkt aus den Daten abgelesen werden. Die optimale Steigung und der optimale EC50-Wert jedoch können durch ein Regressionsverfahren mit Hilfe des mittleren quadratischen Fehlers bestimmt werden.

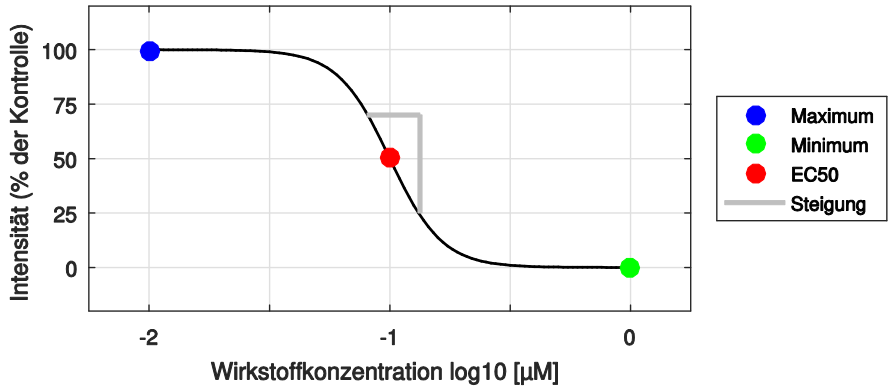


**Abbildung 33: Beispiel Konzentrations-Wirkungskurve.** Die Berechnung einer Konzentrations-Wirkungskurve erfolgt durch den minimalen und maximalen Wert, der Steigung und des EC50-Wertes.

Zur Bestimmung der Steigung und des EC50-Wertes wurde ein Algorithmus entwickelt, welcher die beiden Werte mit Hilfe des mittleren quadratischen Fehlers (engl. *Mean Squared Error*, MSE) optimiert. Dabei wurden die Steigung und der EC50-Wert solange abwechselnd optimiert, bis ein absoluter minimaler MSE erreicht wurde. Bei jeder Optimierung wurden jeweils alle möglichen Werte einmal durchlaufen, während alle anderen Parameter fix geblieben sind (Abbildung 34).

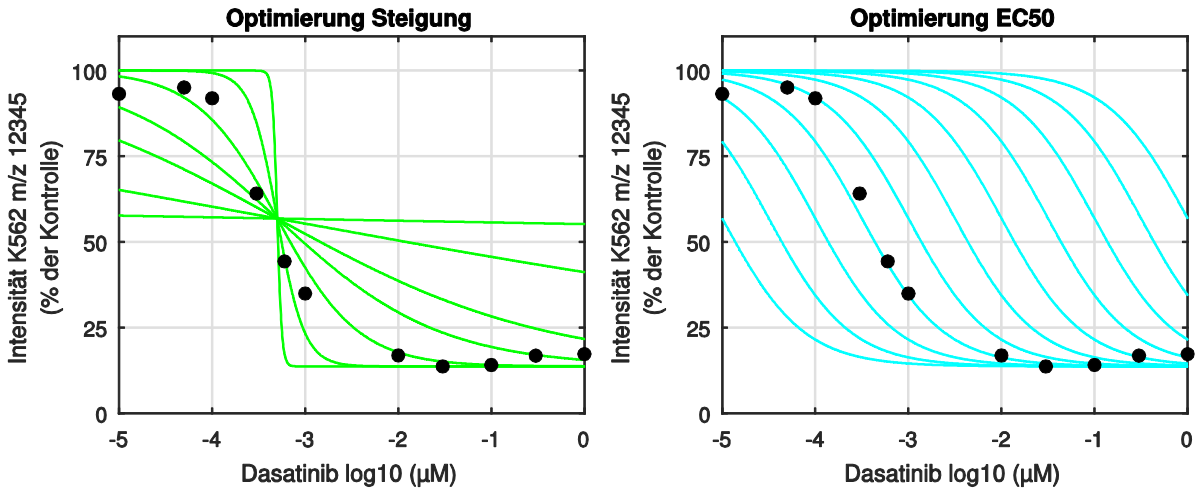


**Abbildung 34: Optimierung der Steigung und des EC50-Wertes einer Konzentrations-Wirkungskurve.** Die Steigung und der EC50-Wert werden mit Hilfe des quadratischen mittleren Fehlers solange optimiert, bis ein absoluter minimaler Fehler erreicht wird.

In manchen Fällen hat das Medikament den Effekt, dass für einige *m/z*-Werte die Intensität mit zunehmender Wirkstoffkonzentration nicht abnimmt, sondern steigt. Da die Zelle aber bei einer gewissen Konzentration funktionell beeinträchtigt ist und schließlich bei noch höheren Konzentrationenden Zelltod erleidet, nimmt die Intensität ab diesem Punkt wieder ab. Um dennoch eine Konzentrations-Wirkungskurve erstellen zu können, wurde der bisher verwendete Algorithmus an diesen Spezialfall angepasst. Wenn die Intensität der Kontrolle kleiner war als die Intensität bei der maximal verwendeten Wirkstoffkonzentration (in diesem Fall handelt es sich um eine steigende Kurve), wurden alle Datenpunkte nach der maximal erreichten Intensität gelöscht (Abbildung 35). Ein weiteres Kriterium zur Erstellung einer steigenden Kurve war, dass zwischen der Kontrolle und der maximalen Intensität mindestens vier weitere Datenpunkte liegen müssen, da zur Berechnung einer sigmoiden Kurve vier Parameter benötigt werden.

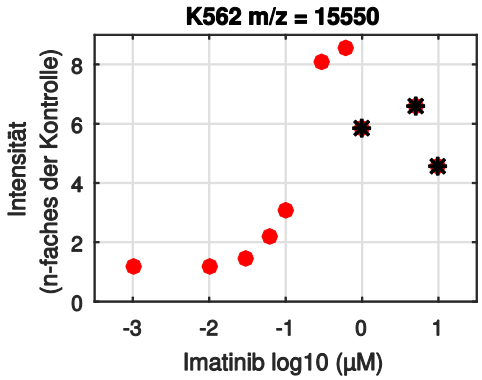


**Abbildung 35: Konzentrations-Wirkungskurve bei steigendem Intensitätsverlauf.** Ab einer bestimmten Wirkstoffkonzentration stirbt die Zelle ab. Daher wurden alle Datenpunkte nach Erreichen der maximalen Intensität gelöscht.

Um zu bewerten, wie gut die realen Daten zum Kurvenverlauf passen, wurde die Korrelation berechnet. Abbildung 36 zeigt beispielhaft den Kurvenverlauf dreier *m/z*-Werte, deren Intensitäten sehr stark, mittelmäßig und schlecht mit der sigmoiden Kurve korrelieren. Bei einer schlechten Korrelation ist davon auszugehen, dass das Medikament keine Wirkung auf das entsprechende Protein ausübt.

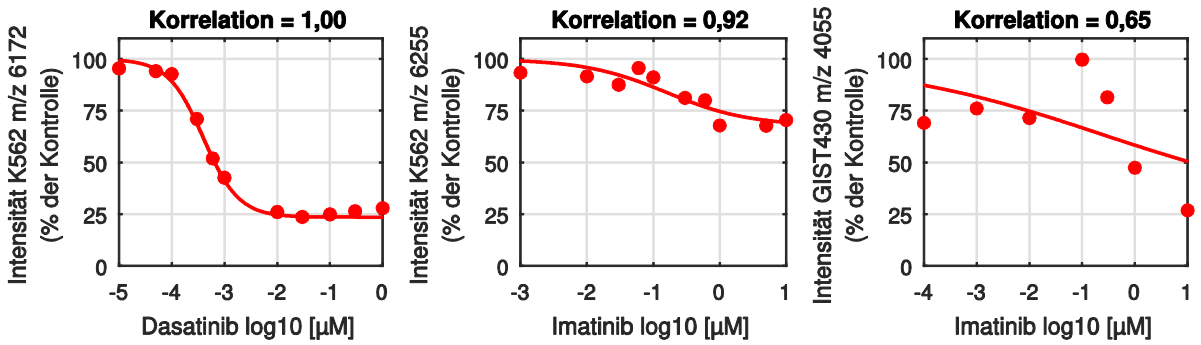


**Abbildung 36: Korrelation zur Bewertung von Konzentrations-Wirkungskurven.** Am Beispiel der Intensitäten von drei m/z-Werten ist hier eine sehr gute (links), mittelmäßige (Mitte) und schlechte (rechts) Korrelation zwischen den realen Daten und der Konzentrations-Wirkungskurve dargestellt.

Eine weitere Kennzahl zur Bewertung des Wirkstoffeffektes ist der sogenannte *Fold Change*. Dieser gibt die Intensitätsänderung in Bezug auf die Kontrolle an. Abbildung 37 zeigt beispielhaft drei weitere Konzentrations-Wirkungskurven mit zwei großen und einem sehr kleinen *Fold Change* (logarithmisch zur Basis 2). Ein sehr kleiner *Fold Change* bedeutet, dass das entsprechende Protein keine oder nur eine sehr kleine Reaktion auf das Medikament zeigt. Da durch die zuvor durchgeführte TIC-Normierung und die Normierung auf die Kontrolle die realen Intensitätsverhältnisse verloren gegangen waren, wurden zur Validierung des *Fold Change* als Bewertungskennzahl die Konzentrations-Wirkungskurven dieser drei Beispiele nochmals für die nicht normierten Intensitäten berechnet. Es hat sich gezeigt, dass auch Signale mit verhältnismäßig niedrigen Intensitäten einen großen *Fold Change* besitzen können. Durch eine Normierung wurde dieser Wert also nicht verfälscht (Abbildung 38).

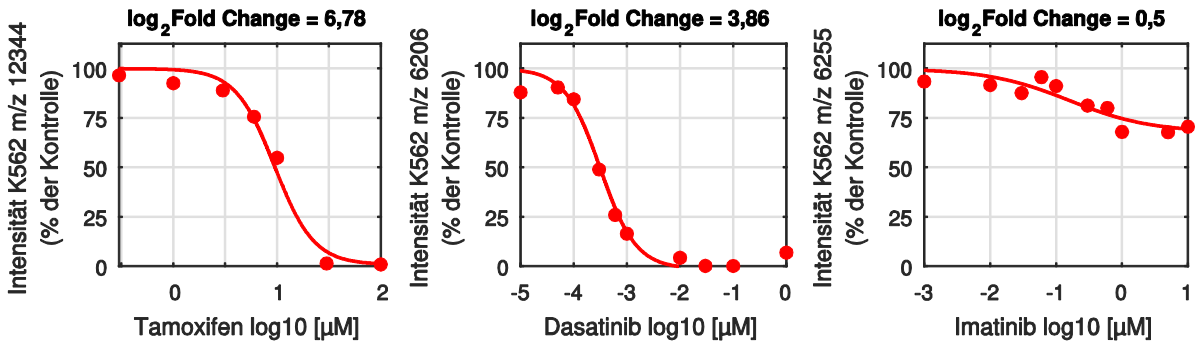


**Abbildung 37: *Fold Change* zur Bewertung der Wirkungsstärke von Medikamenten.** Der *Fold Change* gibt die Intensitätsänderung in Bezug auf die Kontrolle an. Die Abbildung zeigt, dass die Wirkstoffe Tamoxifen und Dasatinib bei den hier dargestellten *m/z*-Werten der K562 Zelllinie eine große Wirkung ausüben. Imatinib hingegen zeigt fast überhaupt keinen Effekt.

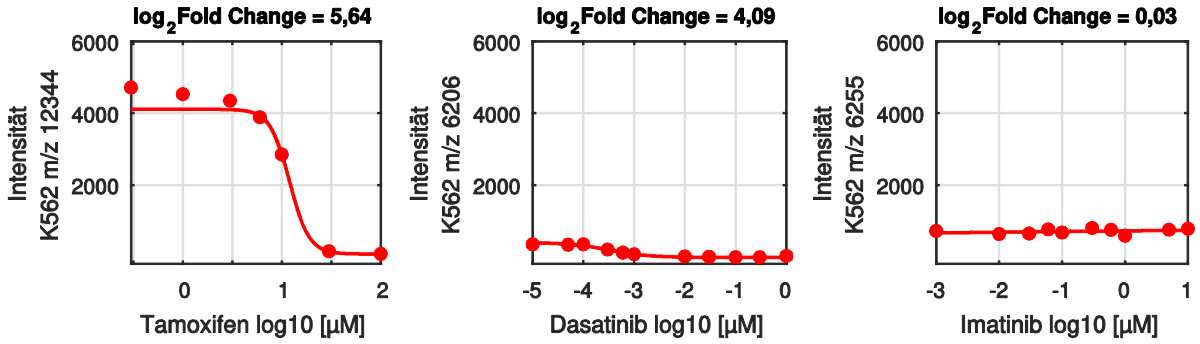


**Abbildung 38: *Fold Change* von Konzentrations-Wirkungskurven bei nicht normierten Daten.** Die Abbildung zeigt ein Signal mit hoher Intensität und großem *Fold Change* (links), ein Signal mit geringerer Intensität und großem *Fold Change* (Mitte) und ein Signal mit hoher Intensität und sehr kleinen *Fold Change*.

Eine Konzentrations-Wirkungs-Kurve besitzt also neben den eigentlichen Parametern zur Berechnung (Minimum, Maximum, Steigung und EC50-Wert) noch viele weitere Eigenschaften. Um auf diese schnell und einfach zugreifen zu können, wurde in MATLAB mit Hilfe objektorientierter Programmierung eine neue Klasse erstellt, welche alle Informationen für weitere Analysen bereitstellt.

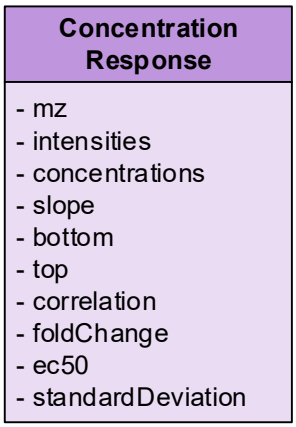


**Abbildung 39: Neue MATLAB Klasse für den einfachen und schnellen Zugriff auf eine Konzentrations-Wirkungskurve.** Neben den Parametern zur Berechnung der sigmoiden Kurve stellt diese Klasse noch weitere Informationen für die Analyse bereit.

### 4.4.2 Methode zur automatischen Auswertung von MS-*Fingerprints*

Im nächsten Schritt wurde diese Methode zur Erstellung von Konzentrations-Wirkungskurven dazu verwendet, in kompletten MS-Datensätzen automatisch nach Massenbereichen zu suchen, welche eine Antwort auf einen Wirkstoff zeigen. Dazu wurde im *Batch*-Verfahren für jeden *m/z*-Wert eines Spektrensatzes eine Konzentrations-Wirkungskurve erstellt und die zugehörigen Parameter und Kenngrößen berechnet. Die erste Kenngröße, die hierbei zur Bewertung des Wirkstoffeffekts eingesetzt wurde, ist die Korrelation. Abbildung 40 zeigt die Korrelationsanalyse eines Datensatzes für Imatinib behandelten K562-Zellen nach einem Such-Durchlauf für fallende Intensitäten bei zunehmender Wirkstoffkonzentration. Es existierten einige *m/z*-Werte, welche eine hohe Korrelation erzielten. Als Schwellwert wurde ein Korrelationskoeffizient von 0,85 festgelegt. Alle *m/z*-Werte, welche darunter lagen, wurden aussortiert, da sie zu sehr vom typischen Verlauf einer Konzentrations-Wirkungskurve abwichen und das Medikament wohl keine große Auswirkung auf sie ausübte. Als zweite Kenngröße zur Bewertung der Wirkstoffreaktion wurde der *Fold Change* der Kurven verwendet, welcher in Abbildung 41 für die verbleibenden m/z-Werte dargestellt ist. Hier wurde der Schwellwert des *Fold Change* auf 1,5 festgesetzt (Abbildung 41).

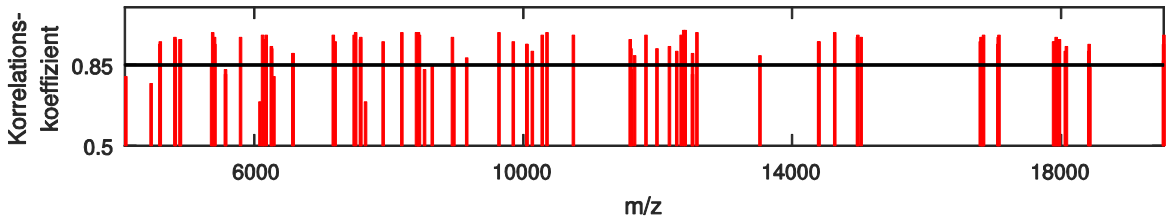


**Abbildung 40: Korrelationsanalyse eines Datensatzes für Imatinib behandelte K562-Zellen nach Durchlauf der IT-Methode zur automatischen Bestimmung von Wirkstoffwirkungen.** Hier wurde nach *m/z*-Werten gesucht, deren Intensitäten mit zunehmender Wirkstoffkonzentration abnehmen. Es wurden einige m/z-Werte gefunden, deren Konzentrations-Wirkungskurven eine hohe Korrelation zu den realen Daten aufweisen. Alle *m/z*-Werte, welche unter dem Schwellwert (Korrelationskoeffizient 0,85) liegen, wurden aussortiert.

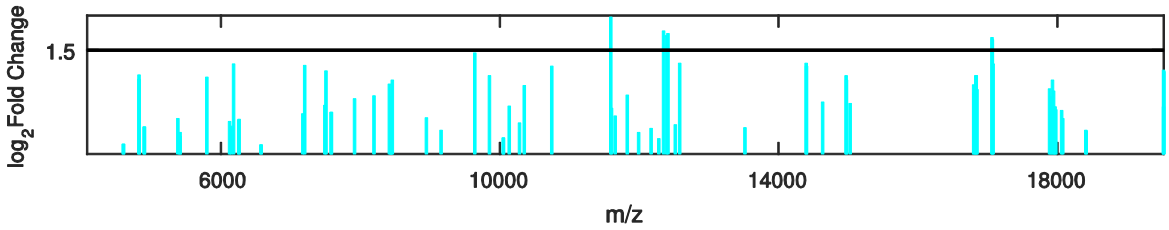


**Abbildung 41: *Fold Change*-Analyse eines Datensatzes für Imatinib behandelte K562-Zellen nach Durchlauf der IT-Methode zur automatischen Bestimmung von Wirkstoffwirkungen.** Hier wurde für jeden *m/z*-Wert, dessen Konzentrations-Wirkungskurve eine Korrelation über 85% erreichte, der entsprechende *Fold Change* berechnet. Alle *m/z*-Werte, welche unter dem Schwellwert (*Fold Change* < 1,5) liegen, wurden aussortiert.

Folgende Tabelle zeigt die *m/z*-Werte, welche nach dem Durchlauf der Methode zur Bestimmung von Wirkstoffwirkungen gefunden wurden. In Abbildung 42 sind die entsprechenden Konzentrations-Wirkungskurven dargestellt.

**Tabelle 13: Ergebnisse der Regressionsanalyse eines Datensatzes für Imatinib behandelte K562-Zellen (fallende Konzentrations-Wirkungskurven).**

| Rangliste | *Fold Change* | Korrelations-koeffizient | *m/z*-Wert |
|---|---|---|---|
| 1 | 1,98 | 0,96 | 11587 |
| 2 | 1,78 | 0,98 | 12347 |
| 3 | 1,74 | 1,00 | 12408 |
| 4 | 1,71 | 1,00 | 12386 |
| 5 | 1,65 | 0,97 | 17056 |

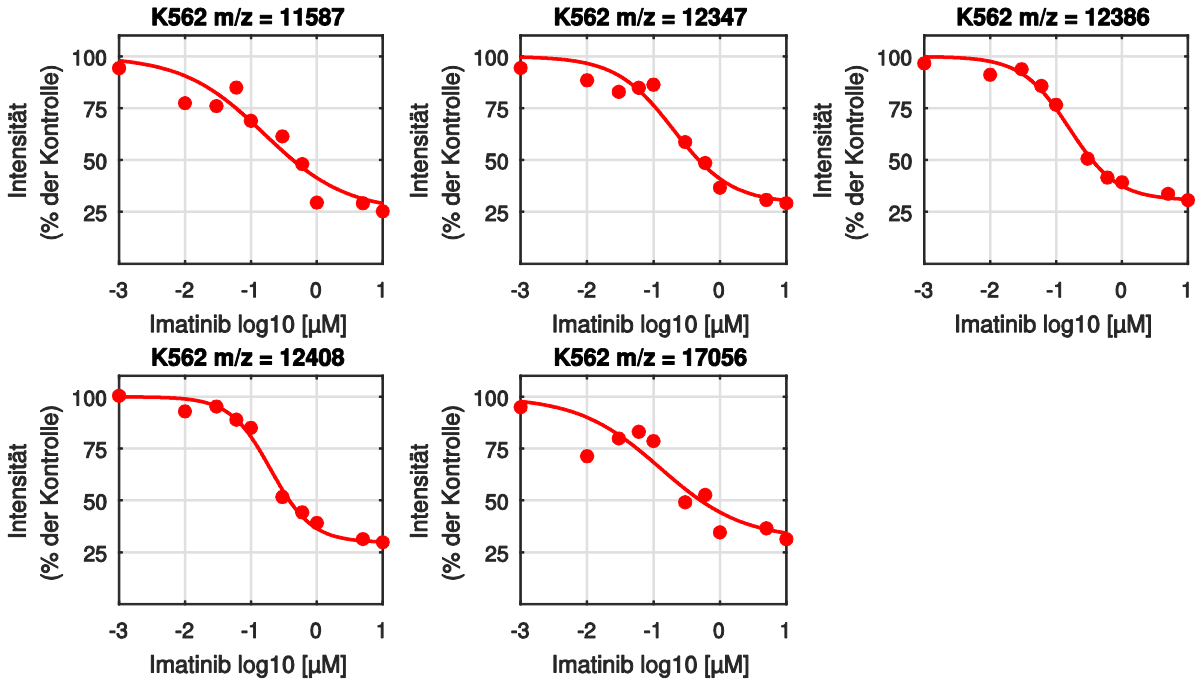


**Abbildung 42: Konzentrationswirkungs-Kurven eines Datensatzes für Imatinib behandelte K562-Zellen nach Durchlauf der IT-Methode zur automatischen Bestimmung von Wirkstoffwirkungen.** Dargestellt sind hier die Konzentrations-Wirkungskurven, welche bei der Berechnung der Kurve einen Korrelationskoeffizienten von mindestens 0,85 erreichten. Des Weiteren besitzen sie einen *Fold Change* größer 1,5.

### 4.4.3 Methode zum Vergleich mehrerer Regressionsanalysen

Nun stellte sich die Frage, ob es sich bei diesen *m/z*-Werten um Proteine handelt, welche ausschließlich eine Reaktion auf den Wirkstoff Imatinib zeigen, oder ob auch ein Effekt bei anderen Wirkstoffen zu erkennen ist. Des Weiteren wurden auch andere Krebs-Zelllinien untersucht, um eventuelle Gemeinsamkeiten festzustellen. Insgesamt wurde zur Bearbeitung dieser Fragestellung eine Sammlung von 13 MS-Datensätzen (bereitgestellt von Dr. Bogdan Munteanu) verwendet, wobei jeder einzelne einmal der automatischen Auswertung unterzogen wurde. Alle *m/z*-Werte, deren Konzentrations-Wirkungskurve eine Korrelation kleiner 85% erreichte, wurden gelöscht. Danach wurde für jeden Datensatz eine Liste aus *m/z*-Werten erstellt, welche prinzipiell einen Wirkstoffeffekt zeigten. Um die Schnittmenge dieser *m/z*-Werte zu bestimmen, musste zunächst für leicht variierende *m/z*-Werte in den verschiedenen Datensätzen (des vermeintlich gleichen Peaks) ein gemeinsamer *m/z*-Wert bestimmt werden. Dazu wurden jeweils *m/z*-Werte aus einem Bereich von fünf Dalton zusammengefasst und ihrem Mittelwert zugeordnet (Tabelle 14).

**Tabelle 14: Auszug aus der Zuordnungstabelle zur Bestimmung der m/z-Schnittmenge nach der Regressionsanalyse**

| m/z-Werte der 13 Datensätze | Gemeinsamer m/z-Wert (gemittelt) | Standard-abweichung |
|---|---|---|
| 12274 | 12276 | |
| 12275 | 12276 | 2,2 |
| 12277 | 12276 | |
| 12279 | 12276 | |
| 12292 | 12296 | |
| 12296 | 12296 | 3,5 |
| 12299 | 12296 | |
| 12343 | 12345 | |
| 12344 | 12345 | |
| 12345 | 12345 | 1,6 |
| 12346 | 12345 | |
| 12347 | 12345 | |

Um zu bestimmen, welcher Schwellwert für den *Fold Change* festgelegt werden soll, wurde ein Histogramm dieser Kennzahl aller verbleibenden *m/z*-Werte erstellt (Abbildung 43). Das Histogramm stellt dar, wie oft ein bestimmter *Fold Change* in allen Datensätzen erreicht wurde. Nun wurde der Schwellwert so gesetzt, dass insgesamt 50% der *m/z*-Werte für die weitere Analyse erhalten bleiben. Die restlichen *m/z*-Werte wurden gelöscht. Dies entspricht her einem *Fold Change* von 1,3.

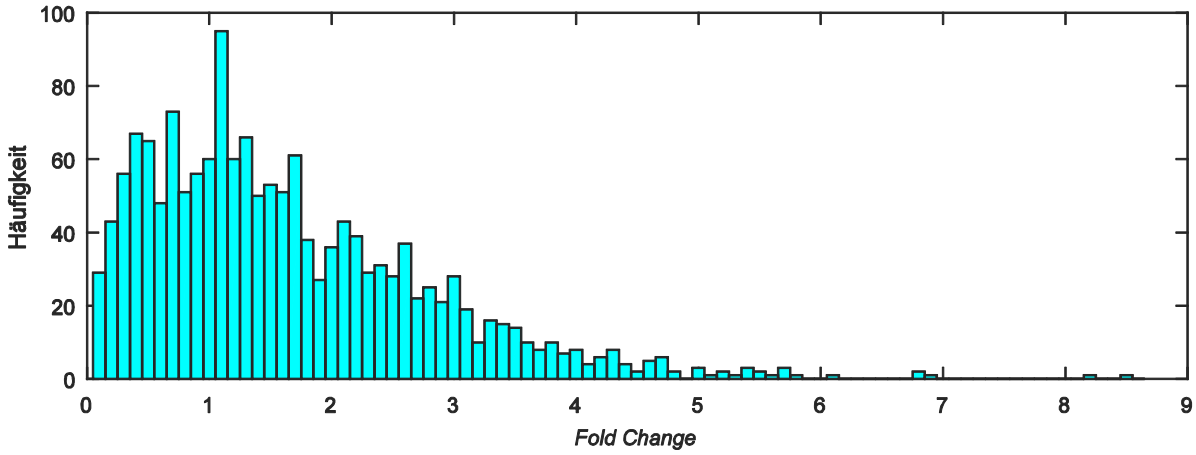


**Abbildung 43: Histogramm der *Fold Change*-Werte für alle Konzentrations-Wirkungskurven nach Durchlauf der IT-Methode zur automatischen Bestimmung von Wirkstoffeffekten.** Damit 50% der m/z-Werte für die weitere Analyse erhalten bleiben, wurde der Schwellwert für den *Fold Change* auf 1,3 gesetzt.

Das Ergebnis der Schnittmengen-Bestimmung für fallende Kurven ist in Abbildung 44 dargestellt. Der *m/z*-Wert 11.584 zeigt in fast jeder Zelllinie eine deutliche Wirkstoffreaktion. Ausnahme sind hier die mit Imatinib behandelten GIST 430 Zellen. Die beiden *m/z*-Werte 6174 und 12.345 zeigen ebenfalls eine Reaktion in vielen Zelllinien auf die Wirkstoffe. Jedoch ist hier neben den mit Imatinib behandelten GIST 430 Zellen auch bei den mit Axitinib behandelten K562-Zellen und den mit Bortezomib behandelten MCF7-Zellen keine Reaktion zu erkennen. Der nicht selektive Wirkstoff Chloroquin zeigt bei keinem *m/z*-Wert eine Reaktion in den K562-Zellen.

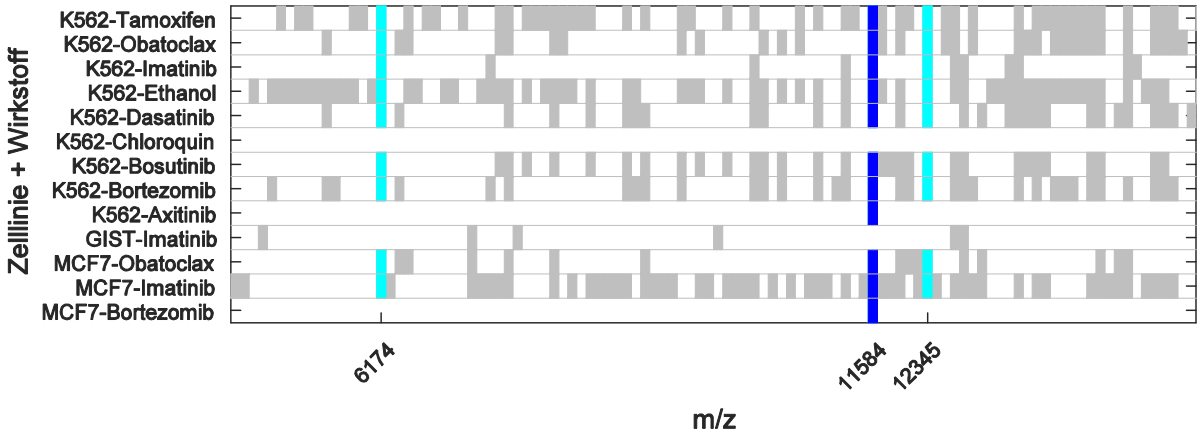


**Abbildung 44: Regressionsanalyse für fallende Kurven**. Der *m/z*-Wert 11.584 zeigt in fast jeder Zelllinie eine deutliche Reaktion auf die Wirkstoffe. Die *m/z*-Werte 6174 und 12.345 zeigen in sehr vielen Zelllinien einen Wirkstoffeffekt. Die K562-Zellen hingegen zeigen überhaupt keine Wirkung auf den nicht selektiven Wirkstoff Chloroquin.

Folgende Abbildung zeigt das Ergebnis der Schnittmengen-Bestimmung für steigende Konzentrations-Wirkungskurven. Allein der *m/z*-Wert 9712 zeigt hier eine deutliche Reaktion in allen Zellenlinien auf alle Wirkstoffe. Die K562-Zellen jedoch zeigen auch hier überhaupt keinen Wirkstoffeffekt.

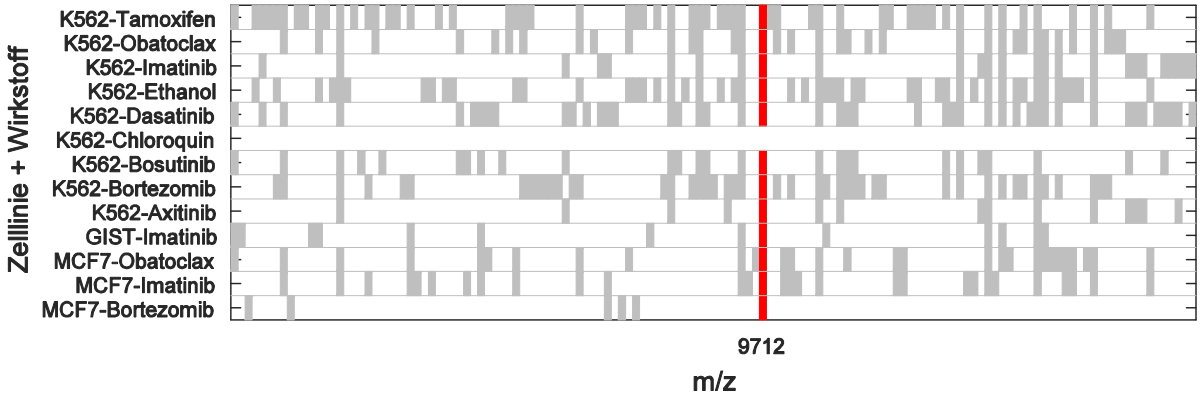


**Abbildung 45: Regressionsanalyse für fallende Kurven.** Es konnte ein *m/z*-Wert (9712) gefunden werden, welcher in allen Zelllinien auf alle Wirkstoffe eine Reaktion zeigt. Die Ausnahme besteht in den K562-Zellen, welche überhaupt keine Wirkung auf den nicht selektiven Wirkstoff Chloroquin zeigen.

### 4.4.4 Entdeckung eines Biomarkers?

Die bei der Regressionsanalyse gefundenen Konzentrations-Wirkungskurven für den *m/z*-Wert 9712 besitzen alle eine Korrelation von mindestens 85% und weisen einen *Fold Change* größer oder gleich 1,3 auf. Dabei ist dieser *m/z*-Wert der einzige, welche in allen drei Krebszellarten für alle Wirkstoffe (bis auf Chloroquin) eine Wirkstoffreaktion zeigt. Daher stellte sich die Frage, ob es sich bei diesem *m/z*-Wert eventuell um einen Biomarker handelt könnte. Im Folgenden sind die Konzentrations-Wirkungskurven der unnormierten Spektren dargestellt.

Abbildung 46 zeigt die Konzentrations-Wirkungskurven der K562-Zelllinie für den *m/z*-Wert 9712. Die y-Achse wurde hier fix auf einen Intensitätsbereich zwischen 0 und 3500 skaliert, um die Kurven besser miteinander vergleichen zu können. Die Konzentrations-Wirkungskurve des Wirkstoffes Chloroquin ähnelt wie erwartet einer geraden Linie, wobei die maximale Intensität ungefähr bei 60 liegt. Die stärkste Reaktion ist bei den Wirkstoffen Dasatinib und Tamoxifen zu erkennen.

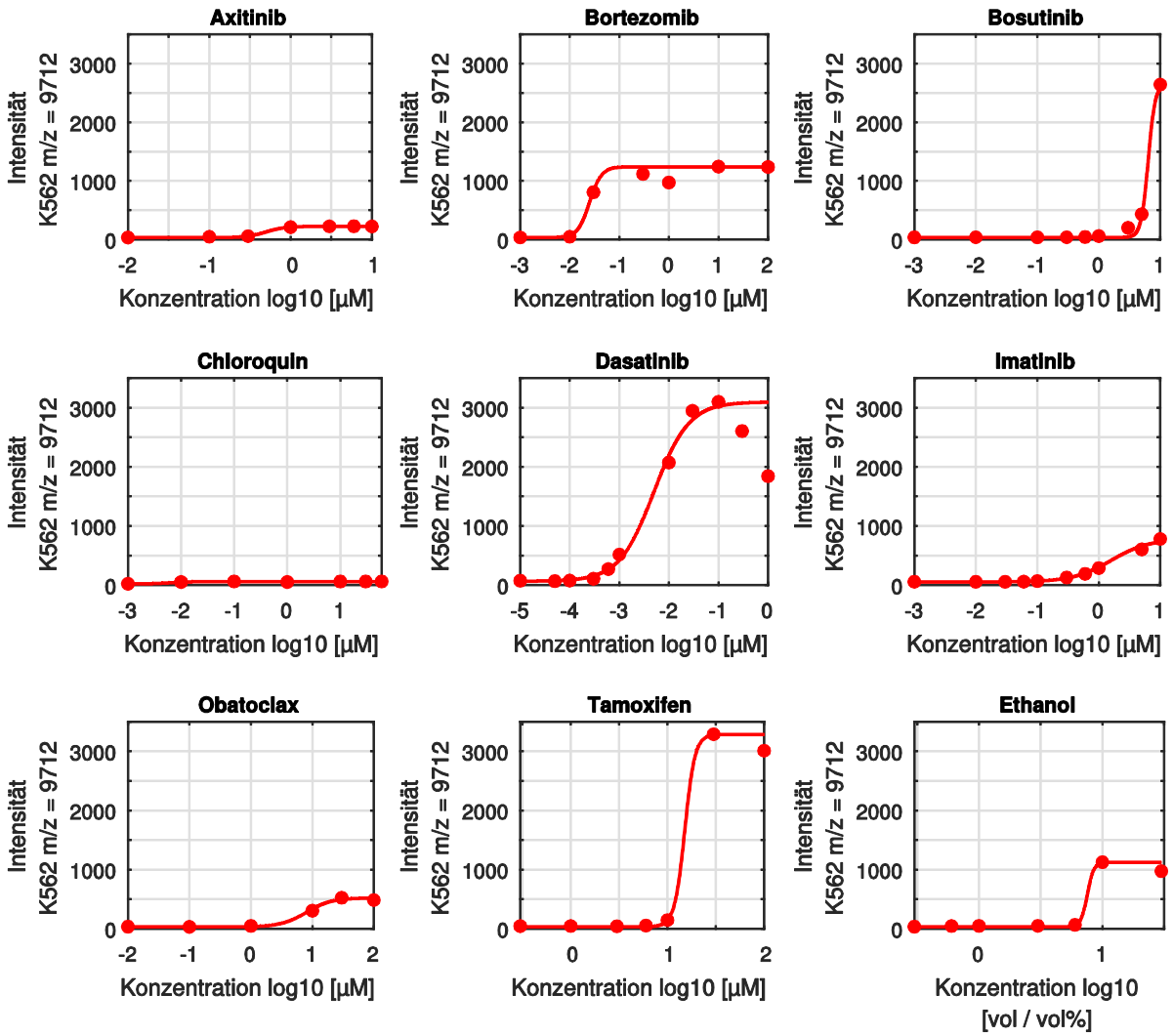


**Abbildung 46: Konzentrations-Wirkungskurven der Zelllinie K562 für den *m/z*-Wert 9712.** Die größte Wirkung zeigen hier die Wirkstoffe Dasatinib und Tamoxifen. Überhaupt keinen Effekt ist beim Wirkstoff Chloroquin zu erkennen.

Abbildung 47 zeigt die Konzentrations-Wirkungskurven der MCF7-Zelllinie für den *m/z*-Wert 9712. Hier wurde die y-Achse fix auf einen Intensitätsbereich zwischen 0 und 20.000 skaliert. Der Wirkstoff Imatinib zeigt einen sehr starken Intensitätsanstieg mit zunehmender Wirkstoffkonzentration. Der Wirkstoff Bortezomib hingegen führt nur zu einem sehr kleinen Intensitätsanstieg, welcher aufgrund der gewählten Skalierung der y-Achse kaum erkennbar ist. Die maximale Intensität liegt hier bei ungefähr 300.

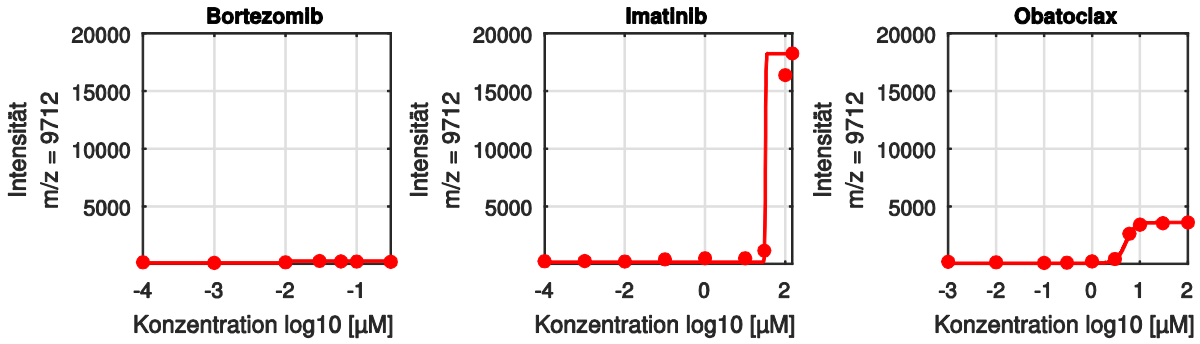


**Abbildung 47: Konzentrations-Wirkungskurven der Zelllinie MCF7 für den *m/z*-Wert 9712.** Eine sehr starke Wirkung wird hier durch den Wirkstoff Imatinib hervorgerufen. Im Gegensatz dazu zeigt der Wirkstoff Bortezomib einen kaum erkennbaren Effekt.

Zuletzt ist hier die Konzentrations-Wirkungskurve der mit Imatinib behandelten Zelllinie GIST 430 für den *m/z*-Wert 9712 dargestellt. Dieser besitzt schon vor der Behandlung der Zellen einen höheren Intensitätswert, welcher aber mit zunehmender Wirkstoffkonzentration noch deutlich weiter ansteigt.

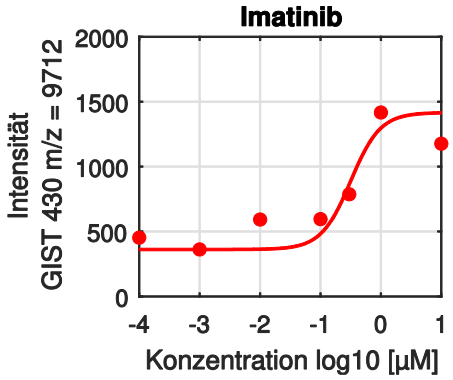


**Abbildung 48: Konzentrations-Wirkungskurve der Zelllinie GIST 430 für den *m/z*-Wert 9712.** Der *m/z*-Wert 9712 besitzt schon vor Behandlung der Zellen einen höheren Intensitätswert. Dieser steigt jedoch mit zunehmender Wirkstoffkonzentration noch weiter an.

Durch die hier entwickelte Methode zur automatischen Bestimmung von Wirkstoffeffekten konnte also ein *m/z*-Wert gefunden werden, welcher in den hier verwendeten Zelllinien für sämtliche Wirkstoffe eine Reaktion zeigt. Um welches Protein es sich hierbei handelt, ist jedoch allein durch informationstechnische Methoden nicht zu bestimmen.

## 4.5 Erstellung einer graphischen Benutzeroberfläche

Um Wissenschaftlern die in diesem Dissertationsprojekt entwickelten Methoden und Algorithmen zugänglich zu machen, wurden alle implementierten Funktionen in einer MATLAB Bibliothek zusammengefasst. Des Weiteren wurde eine graphische Benutzeroberfläche erstellt, welche eine einfache und schnelle Durchführung der Algorithmen ermöglicht. Es können mehrere Datensätze eingelesen und bearbeitet werden. Für die Darstellung der Daten kann zwischen Einzelspektren, Mittelwertspektren und einer *Heatmap* gewählt werden. Die einzelnen Schritte des Vorverarbeitungsprozesses der MS-Spektren können hier individuell zusammengestellt werden. Des Weiteren kann eine Merkmalsextraktion mit Hilfe der Fisher Diskriminante durchgeführt werden. Die beiden Standardmethoden zur Klassifikation, eine hierarchische Clusteranalyse und eine Hauptkomponentenanalyse, können ebenfalls angewendet werden. Zuletzt wird die Möglichkeit einer automatischen Regressionsanalyse zur Bestimmung von Wirkstoffwirkungen angeboten. Dabei können auch die Ergebnisse mehrerer Datensätze miteinander verknüpft werden.

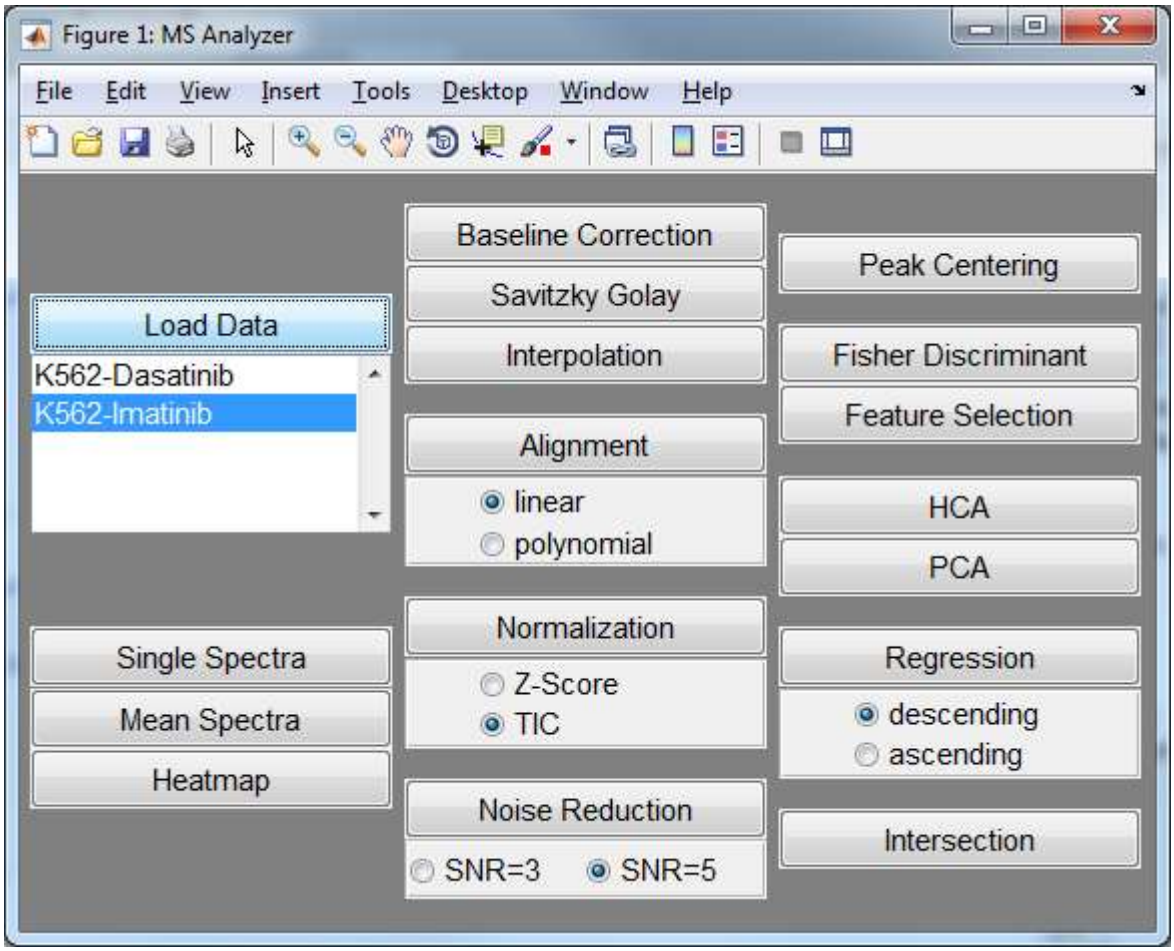


**Abbildung 49: Graphische Benutzeroberfläche für Verarbeitung und Analyse von MS-*Fingerprints*.**

# 5 DISKUSSION

Die Entwicklung molekular-gerichteter Therapieansätze stellt die klinische Forschung vor große Herausforderungen. Das ist Ziel ist es, dem Patient eine unwirksame Behandlung und unnötige Nebenwirkungen zu ersparen. Mit Hilfe von Biomarkern oder allgemein molekularen Signaturen kann die Wirksamkeit molekular-gerichteter Therapeutika angezeigt oder vorhergesagt werden. Die Grundlage dafür besteht in der schnellen, zuverlässigen und kostengünstigen Klassifizierung humaner Tumorproben. Moderne analytische Hochleistungsverfahren, wie zum Beispiel die Massenspektrometrie, generieren jedoch Datensätze im Umfang von mehreren Gigabytes, deren Bearbeitung und Auswertung nur mit Hilfe moderner Signalverarbeitung möglich ist. Die Erforschung und Entwicklung systematischer Arbeitsprozesse zur automatischen Analyse humaner Karzinome wäre hilfreich, um in Zukunft medizinisch relevante Merkmale schneller und zuverlässiger als bisher extrahieren zu können.

## 5.1 Vorverarbeitung von Massenspektren

Nach einer Messung stehen die Massenspektren im firmenspezifischen Datenformat vor und sind nicht unbedingt kompatibel mit dem Datenformat anderer Analyse-Tools. Durch das in diesem Dissertationsprojekt entwickeltem Datenformat, welches einen objektorientierten Ansatz verfolgte, wurde ein einfacher und schneller Zugriff auf die Daten ermöglicht. Dabei wurde der komplette Datensatz auf einmal eingelesen, und die Spektren automatisch den zugehörigen Klassen zugeordnet. Dies stellte eine Zeitersparnis zum Beispiel gegenüber dem Softwarepaket ClinProTools (Bruker Daltronics) dar, bei welchem der Nutzer die Spektren oder Klassen einzeln auswählen muss. Der objektorientierte Ansatz bietet zudem die Möglichkeit der Erweiterung. Auch können Spektren oder Signale anderer Messgeräte eingelesen und verarbeitet werden. Durch die Überführung in diese objektorientierte Datenstruktur können nun alle Methoden zur Vorverarbeitung und zur statistischen Analyse unabhängig des Messverfahrens auf die Daten angewandt werden, weshalb nur eine einzige Implementierung aller weiteren Funktionen notwendig ist. Außerdem ist eine Fusionierung mit anderen Spektraldaten, zum Beispiel aus der Infrarotspektroskopie, möglich. Des Weiteren können Zwischenergebnisse jederzeit problemlos abgespeichert und wieder eingelesen werden.

Martens et al. (Martens et al., 2011) entwickelten ebenfalls eine Datenstruktur für die Verarbeitung von Massenspektren. Es wird *mzML* genannt und soll den Austausch und die Verarbeitung von MS-Daten vereinfachen und eine effiziente Speicherung von großen

Datenmengen ermöglichen. Die Metainformationen der Spektren sind hier jeweils in einer XML-Datei, die spektrale Information in einer binären Datei gespeichert. Um mit diesen Daten in MATLAB arbeiten zu können, wird ein Converter angeboten. Dieses Datenformat bietet jedoch nicht die Vorteile einer objektorientierten Programmierung: Klassen, Vererbung und Polymorphie. Insbesondere ist dieses Datenformat nicht für Daten anderer Messgeräte ausgelegt.

Um sinnvolle statistische Analysen durchführen zu können, müssen die Massenspektren eine gemeinsame Massenachse besitzen. Dies ist jedoch zum Beispiel aufgrund verschiedener Geräteeinstellungen nicht immer der Fall. Im Vergleich zur häufig verwendeten Software *ClinProTools* (Bruker Daltronics), welche im Anschluss an ein *Peak Picking* eine Rekalibration anhand herausragender Peaks durchgeführt, werden die Massenspektren in dieser Arbeit vor der Merkmalsextraktion mit Hilfe einer linearen Interpolation korrigiert. Dies hat den Vorteil, dass keine eventuell wichtige Information verloren gehen und die Auswahl der Peaks anhand statistischer Methoden durchgeführt werden kann. Es wurden zwar schon statistische Methoden zur Merkmalsextraktion zur Klassifikation von Fingerprints angewandt, doch wurde zuvor zusätzlich ein *Peak Picking* durchgeführt (Marvin-Guy et al., 2008; Schwamb et al., 2013).

Die Rekalibration / Ausrichtung der Massenachse wird in dieser Arbeit mit Hilfe der Korrelationsfunktion durchgeführt. Diese Methode kann auch durchgeführt werden, wenn zuvor kein *Peak Picking* stattgefunden hat. Bei der Rekalibration anhand bestimmter Peaks oder Standards müssen deren genauen Massenwerte bekannt sein. Des Weiteren müssen sie auf dem gesamten Massenbereich relativ gleichmäßig verteilt sein (Gobom et al., 2002; Kulkarni et al., 2015). Aufgrund der relativ großen Breite der Peaks und der begrenzten Auflösung im Linearmodus stellt die Ausrichtung der Massenachse mit Hilfe der Korrelationsfunktion eine gute Möglichkeit für die Rekalibration von Protein-*Fingerprints* dar. Zur Rekalibration von Spektren, welche im Reflektormodus, in welchen die Breite der Peaks sehr klein ist, wäre das Verfahren von Gobom et al. die bessere Variante, bei welcher ein Polynom auf mehrere im gesamten Massenbereich verteilte Massen gelegt wird.

Um das chemische und instrumentelle Rauschen in einem Massenspektrum zu reduzieren, wird oft eine Basislinienkorrektur und eine Savitzly Golay Glättung durchgeführt (Munteanu et al., 2012; Ouedraogo et al., 2010; Schwamb et al., 2013). Bereiche, die nur aus Rauschen bestehen, werden dabei jedoch nur unterdrückt und nicht gelöscht. Die Bestimmung des Signal-Rausch-Verhältnisses der Peaks in *Fingerprints* wurde bisher nur in Kombination mit dem *Peak Picking* Verfahrens durchgeführt. Dabei wurde anhand der zugehörigen Datenpunkte eines Peaks das SNR berechnet. Falls dieser unter dem gewünschten Wert lag, wurde dieser Peak aussortiert. Eine Beurteilung dieses Peaks mit Hilfe statistischer

Methoden war dann nicht mehr möglich. Die in dieser Arbeit entwickelte Methode zur Bestimmung des Rauschpegels hat den Vorteil, dass im Vorfeld keine Datenreduktion notwendig ist und nachfolgende statistische Analysen nicht durch Bereiche, welche ausschließlich aus Rauschen bestehen, verfälscht werden können.

Durch die in dieser Arbeit entwickelten Methoden zur Vorverarbeitung konnten die massenspektrometrischen Fingerprints auf geeignete Weise vorverarbeitet werden, so dass sie mit einer gemeinsamen Massenachse, ausgeglichenen Verschiebungen und weitgehend ohne Rauschen in Form einer Matrix für die weiteren Analysen zur Verfügung standen. Die Auswahl der Peak konnte nun mit Hilfe statistischer Methoden statt einem *Peak Picking* getroffen werden.

## 5.2 Kennzahl für die externe Clusteranalyse

Um einen Workflow bewerten zu können, wird ein objektives Kriterium benötigt. Bei der unüberwachten Klassifikation, vor allem wenn die Zahl der Klassen, welche klassifiziert werden sollen, sehr groß ist, ist die Wahl einer Bewertungskennzahl nicht einfach. Wie schon beschrieben, kann die Clusteranalyse mit Hilfe der exakten Klassifikationsrate schnell an die Grenzen der Speicherkapazität und einer akzeptablen Rechenzeit stoßen. Die Alternativen zur exakten Klassifkationsrate, wie zum Beispiel die Reinheit oder die Entropie, können auch bei einer großen Zahl an Klassen schnell berechnet werden, machen jedoch keine genaue Aussage über die Qualität der Klassifikation. Andere Kennzahlen, wie zum Beispiel der Precision / Recall / F-Measure (van Rijsbergen, 1979) oder der Rand Index (Rand, 1971) sind nur auf das Zwei-Klassen-Problem anwendbar. In der Literatur werden oft Klassifikationen durchgeführt, bei welchen kranke von gesunden Proben unterschieden werden sollen. Wu et. al (Wu et al., 2003) zum Beispiel unterscheiden Proben von Patienten mit Eierstockkrebs von gesunden Proben. Sie benutzen die Fehlerrate zur Beurteilung der Klassifikation. Karger et al. (Karger et al., 2010) klassifizierten 66 verschiedene Zelllinien, doch wurde die Beurteilung des Klassifikationsergebnisses nicht anhand einer Bewertungskennzahl, sondern visuell durch Betrachtung des Dendrogramms durchgeführt. Eine systematische Untersuchung von Klassifikationen mit Bewertungskennzahlen, bei welchen sehr viele Zelllinien getrennt werden sollen, wurde bisher noch nicht durchgeführt.

Eine weitere oft verwendete visuelle Darstellung von Klassifikationsergebnissen ist die PCA. Sie veranschaulicht sehr gut die Trennbarkeit der Klassen mit Hilfe von Punktwolken, auch im Fall von mehreren Klassen. Doch auch sie besitzt keine Kennzahl, mit welcher die Ergebnisse von anderen PCA-Analysen verglichen werden können.

Daher stellt die in dieser Arbeit entwickelten Kennzahl, das *Confusion Matrix Maximum (CMM)*, eine sehr gute Möglichkeit zur externen Clusteranalyse dar. Es konnte gezeigt werden, dass sie in Bezug auf die Rechenzeit keinen Engpass darstellt und der exakten Klassifikationsrate sehr nahe kommt. In den meisten Fällen sind sie sogar identisch ist. Sie eignet sich vor allem zur systematischen Entwicklung von Klassifikationsmodellen, da hier zahlreiche Tests durchgeführt werden müssen. Dies wäre aufgrund der hohen Rechenzeit mit der exakten Klassifikationsrate nicht durchführbar. Im Gegensatz zur Reinheit und Entropie liefert sie auch die stabileren Ergebnisse.

## 5.3 Schnelle Klassifikation von Tumorzellen und Tumorzell-Subtypen

Es sollte ein Modell entwickelt werden, welches eine große Zahl an Krebszelllinien optimal klassifizieren kann. Bei den Zelllinien handelte es sich hierbei um Brustkrebs-, Leukämie- und GIST-Zelllinien. Insgesamt bestand der Datensatz aus elf Zelltypen. Es wurde bisher noch nicht untersucht, ob eine Klassifikation von Tumorsubtypen anhand von MS-*Fingerprints* möglich ist, ohne dabei die Zugehörigkeit zum Tumortyp zu verlieren. Zur systematischen Entwicklung eines optimalen Klassifikationsmodells wurden verschiedene Methoden zur Merkmalsextraktion und Klassifikation getestet. Dabei wurden sowohl überwachte und unüberwachte Methoden eingesetzt. Jede Methode zur Merkmalsextraktion wurde einmal mit jeder Klassifikationsmethode kombiniert. Diese wäre aufgrund des hohen Zeitaufwands zur Berechnung der exakten Klassifikationsrate nicht in annehmbarer Zeit möglich gewesen.

Für die Berechnung der Diskriminanten zur Merkmalsextraktion wurden als „Peaks“ die lokalen Maxima der *Skyline* (maximale Intensität eines *m/z*-Wertes) des Datensatzes herangezogen. Die Verwendung der *Skyline* eines Datensatzes statt dem Mittelwertspektrum hatte zur Folge, dass kleine Peaks nicht aufgrund der Mittelung verloren gegangen sind. Des Weiteren konnte durch die Extraktion *aller* lokalen Maxima der *Skyline* ein weiterer Datenverlust verhindert werden. Die Beurteilung der Signifikanz der Peaks konnte nun ausschließlich anhand statistischer Methoden durchgeführt werden.

Die in dieser Arbeit entwickelte Methode zur Bestimmung der Anzahl an Peaks für eine optimale Klassifikation eignete sich sehr gut für die Zellklassifizierung anhand von *Fingerprints*. Durch die Berechnung der Diskriminanten anhand der Tumortypen (Brustkrebs, GIST und Leukämie) und nicht anhand der einzelnen Zelllinien, konnte sichergestellt werden, dass die Trennung der verschiedenen Erkrankungen jederzeit erhalten blieb. Des Weiteren konnte durch Verwendung des CMM die exakte Anzahl an Peaks berechnet werden, mit welcher eine optimale Klassifikation der Zelllinien durchgeführt werden kann. Eine Alternative

zu diesem Verfahren stellen die genetischen Algorithmen dar, bei welchen verschiedene Peak-Kombinationen getestet werden. Aufgrund der zufälligen Startparameter kann es hierbei jedoch passieren, dass zwar ein gutes, aber nicht das optimale Ergebnis gefunden wird. Marvin-Guy et al.(Marvin-Guy et al., 2008) verwendeten nach einer Varianzanalyse ein Konfidenzintervall von 95% zur Auswahl der Peaks. Das 95%-Konfidenzintervall gibt den Bereich an, bei dem mit einer Wahrscheinlichkeit von 95% davon ausgegangen werden kann, dass sich innerhalb dessen der wahre Mittelwert befindet. Hierbei handelte es sich also nicht um einen exakten Wert. Die Anwendung eines genetischen Algorithmus könnte eventuell zu einem exakten Klassifikationsergebnis führen. Schwamb et al. (Schwamb et al., 2013) benutzen ihn, um die beste Peak Kombination zur Identifikation von Signaturen als Indikator für Zellstress und Apoptose anhand von *Fingerprints* zu erhalten. Jedoch wurde hier im Vorfeld ein *Peak Picking* durchgeführt, was eventuell zu einem Informationsverlust geführt haben könnte. Des Weiteren wurden nur drei verschiedene Gruppen untersucht. Für die Peak-Auswahl aus einem sehr großen Merkmalraum und mehreren Gruppen wäre dieser Algorithmus nicht geeignet.

Eine Möglichkeit zur überwachten Klassifikation, welche in dieser Arbeit nicht untersucht wurde, stellen die künstlichen neuronalen Netze (engl. *Artificial Neural Network*, ANN) dar. Künstliche neuronale Netze sind Modelle aus der Informationsverarbeitung (Künstliche Intelligenz), welche die Nervenzellvernetzungen im Gehirn und Rückenmark als biologisches Vorbild betrachten. Sie bestehen aus einer Reihe von Eingangssignalen (Merkmalen) mit zugehörigen Gewichten, eine variablen Menge an Neuronen und den Ausgangssignalen, welche den Klassen des Datensatzes entsprechen. Wie bei den Synapsen der Nervenzellen gibt es auch hier verschiedene Aktivierungsfunktionen, nach deren Regel ein Signal von einem Neuronen an den Ausgang weitergeleitet wird (McCulloch und Pitts, 1943). Valletta et al. (Valletta et al., 2016) benutzen neuronale Netze zur quantitativen Bestimmung von Zelllinien-Kreuzkontaminationen anhand von MS- *Fingerprints*. Diese Klassifikationsmethode wurde in dieser Arbeit nicht in die systematische Untersuchung einbezogen, da insgesamt zu wenige Daten zur Verfügung standen. Üblicherweise werden hier die Daten in Trainings-, Test- und Validierungsdaten aufgeteilt, was bei einem Datensatz bestehend aus zwei Experimenten mit elf Zelllinien zu je acht Spektren zwangläufig zu einem „auswendig lernen“ des neuronalen Netzes geführt hätte.

Die besten Klassifikationsergebnisse wurden unter anderem durch eine Merkmalsextraktion mit Hilfe des *Fisher Score Optimized* erzielt. Diese wurden mit der Funktion aus dem Softwarepaket der Arizona State University verglichen. In beiden Fällen wurden Klassifikationsraten von 100% erreicht, jedoch gab es große Unterschiede in der Rechenzeit. Die enorme Zeitdifferenz der beiden Funktionen war auf die unterschiedliche Art der

Implementierung zurückzuführen. Die Funktion zur Fisher-Diskriminanzanalyse aus dem Softwarepaket der Arizona State University besteht hauptsächlich aus *for*-Schleifen, welche in MATLAB sehr viel Rechenzeit benötigen. Die Implementierung der *Fisher Score Optimized* Variante hingegen stellte die effizientere Umsetzung dar, da sie weniger *for*-Schleifen, dafür aber einige Matrix-orientierte Befehle einsetzt. Diese ermöglichen eine schnelle Berechnung komplexer Probleme.

## 5.4 Automatische Analyse von Konzentrations-Wirkungs-Beziehungen zur Biomarker-Identifikation

In dieser Arbeit wurde ein Regressionsmodell entwickelt, bei welchen die Konzentrations-Wirkungskurve dem typischen Verlauf einer steigenden oder fallenden sigmoiden Kurve folgt. Di Veroli et al. (Di Veroli et al., 2015) entwickelten einen Algorithmus zur automatischen Erstellung von Dosis-Wirkungskurven, welche mehrere Phasen des Steigens und / oder des Fallens beinhalten können. Dieser Algorithmus kann jedoch nur verwendet werden, wenn im Vorfeld bekannt ist, dass der zu untersuchende *m/z*-Wert mehrere solcher Phasen bei einer Wirkstoffbehandlung besitzt. Dies setzt die Kenntnis über das Verhalten bei Wirkstoff-behandlung des entsprechenden Proteins voraus. Die Signale könnten auch rein zufällig, zum Beispiel durch Messschwankungen, mal stärker oder schwächer sein. Ein Algorithmus, welcher mehrere steigende und fallende Phasen erlaubt, wäre bei einer automatischen Analyse von Konzentrations-Wirkungs-Beziehungen praktisch in der Lage, für fast jeden *m/z*-Wert eine Kurve zu erstellen, welche eine sehr hohe Korrelation zu den realen Daten erreicht. Aus diesem Grund eignet sich ein sigmoides Regressionsmodel, wie es in dieser Arbeit verwendet wurde, besser für die Erkennung von Wirkstoffeffekten bei der systematischen Untersuchung des gesamten Massenbereichs. Hierbei soll die Frage beantwortet werden, ob ein Medikament bei einem *m/z*-Wert grundsätzlich zu einer Wirkstoffreaktion führt oder nicht.

Das Softwarepaket *Dr Fit* (Di Veroli et al., 2015), welches für die Erstellung von Konzentrations-Wirkungskurven mit mehreren steigenden und fallenden Phasen entwickelt wurde, und die Statistik-Software *GraphPad Prim* (Motulsky und Christopoulos, 2004), welche die Berechnung einer sigmoiden Konzentrations-Wirkungskurve ermöglicht, gestatten die Analyse einzelner *m/z*-Werte, bieten jedoch nicht die Möglichkeit, einen Datensatz aus massenspektrometrischen *Fingerprints* automatisch nach Proteinen zu durchsuchen, welche eine Wirkung auf das Medikament zeigen. Des Weiteren gibt es kein Verfahren, welches die Ergebnisse verschiedener Regressionsanalysen miteinander vergleicht.

Durch die in dieser Arbeit entwickelten Methode zur automatischen Analyse von Konzentrations-Wirkungs-Beziehungen konnte ein *m/z*-Wert (9712) gefunden werden, welcher in allen hier verwendeten Zelllinien eine deutliche Reaktion auf die Wirkstoffbehandlung zeigt. Die einzige Ausnahme bestand dabei aus dem unselektiven Wirkstoff Chloroquin.

Der *m/z*-Wert 9712 wurde in der Literatur schon des Öfteren dem Apolipoprotein C-III zugeordnet (Bondarenko et al., 1999; Pecks et al., 2014). Als Apolipoproteine bezeichnet man den Proteinanteil der Lipoproteine, welcher die wasserlöslichen Lipide im Blut transportiert. Sie werden hauptsächlich von der Leber und dem Darm produziert. Die Messung der Blutspiegel der verschiedenen Apolipoproteine kann Aufschluss über eine eventuell vorliegende Erkrankung des Fettstoffwechsels geben. Das Apolipoprotein C-III gilt als Hauptregulator des Plasma-Triglyceridspiegels (Ginsberg und Brown, 2011).

Die Berechnung des *m/z*-Wertes für das Apolipoprotein C-III erfolgte jedoch auf der Basis von Blutplasma (Bondarenko et al., 1999), dem nicht-zellulären Anteil des Blutes. Die in dieser Arbeit verwendeten Zellen hingegen stammen aus der Zellkultur. Des Weiteren ist aufgrund der MS-Messung im Linearmodus die Auflösung des Peaks nicht sehr hoch, weshalb nicht mit Sicherheit gesagt werden kann, dass es sich hierbei tatsächlich um das Protein mit dem *m/z*-Wert 9712 handelt. Für eine sichere Identifikation dieses Peaks müssten weitere biotechnologische Analysen durchgeführt werden.

## 5.5 Ausblick: Übertragung auf MS-Bildgebung und Infrarot-Spektroskopie-Fingerprinting

Mit Hilfe der Massenspektrometrie können anhand von *Fingerprints* erfolgreich Zellklassifikationen und Wirkstoffbestimmungen durchgeführt werden. Durch die bildgebende Massenspektrometrie kann diese Anwendung auf die molekulare Analyse von Geweben erweitert werden (Norris und Caprioli, 2013). Ähnlich eines Rasters werden auf einem Gewebeschnitt eine Vielzahl an Spektren generiert und mit den entsprechenden Ortskoordinaten abgespeichert. Trotz der hohen Dynamik der Intensitätswerte können sämtliche m/z-Werte mithilfe hochkomplexer Signal- und Bildverarbeitung in Falschfarben dargestellt werden. Diese Methode gibt somit einen wichtigen Aufschluss über die Ortsauflösung wichtiger Biomoleküle oder anderer Substanzen im Gewebe. Sie ermöglicht die Suche nach neuen krankheitsspezifischen Biomarkern (Balluff et al., 2012). Des Weiteren kann sie Aufschluss über metabolomische Bestandteile des Stoffwechsels geben (Miura et al., 2012). Eine weitere Anwendung ist die Detektion von Pharmaka und deren Metaboliten in Geweben (Schwamborn, 2012). Eine Übersicht verschiedener Methoden zur

Verarbeitung und Analyse von MS-Bilddaten wurde von Alexandrov (Alexandrov, 2012) in einem *Review* zusammengestellt. Im Vergleich zu den MS-Fingerprints entstehen hierbei jedoch Datensätze in viel größeren Volumen, was die Speicherung und Verarbeitung erheblich erschwert. Die Übertragung der 1D-Analysemethoden auf die Bilddaten führt in manchen Fällen zu nicht mehr annehmbaren Rechenzeiten. Interessant wäre auch eine Untersuchung, ob m/z-Werte, welche zur Klassifikation von *Fingerprints* auch zur Klassifikation der entsprechenden MS-Bilddaten geeignet sind.

Eine weitere Methode, welche ebenfalls erfolgreich in der Mikrobiologie (Lasch und Naumann, 2006) und zur Untersuchung von Zellen und erkranktem Gewebe (Diem et al., 2008) eingesetzt wurde, ist die Infrarotspektroskopie (IR). Sie ist im Gegensatz zur Massenspektrometrie ein zerstörungsfreies Messverfahren, mit welchem ohne Probenvorbereitung gearbeitet werden kann. Sie basiert auf der Wechselwirkung elektromagnetischer Strahlung aus dem infraroten Spektralbereich und der zu untersuchenden Probe. Insbesondere funktionelle Gruppen von Molekülen, welche eine polare Bindung aufweisen, können durch IR-Spektroskopie nachgewiesen und somit biomolekulare Strukturen aufgeklärt werden. Speziell der sogenannte Fingerprint-Bereich im mittleren IR (MIR) in welchem viele Banden zu beobachten sind, ermöglicht oft eine eindeutige Zuordnung der zu analysierenden Probe. Im Vergleich zur Massenspektrometrie besteht das spektrale Fenster nicht nur aus Lipiden und Proteinen, sondern es enthält auch Fettsäuren und Polysaccharide. Die Infrarotspektroskopie eignet sich hervorragend zur Untersuchung von Bakterienzellen (Davis und Mauer, 2010). Interessant wäre eine Untersuchung, ob durch die Fusionierung der Massenspektrometrie und der Infrarotspektroskopie die Klassifikationsergebnisse in der Zellanalytik von Säugerzellen noch weiter verbessert werden könnten. Die in dieser Arbeit entwickelten Methoden für die MS-*Fingerprint* Verarbeitung und Analyse konnten teilweise auch auf die Infrarotspektroskopie übertragen werden (Lux et al., 2013).

# 6 ZUSAMMENFASSUNG

Personalisierte Medizin steht im besonderen Fokus der modernen Krebsforschung. Anhand spezifischer Signaturen soll dem Patient einer molekular-gerichteten und persönlich zugeschnittenen Therapie zugeführt werden, welche ihm eine unwirksame Behandlung und unnötige Nebenwirkungen ersparen soll. Dies kann nur durch moderne analytische Hochleistungs-Verfahren wie zum Beispiel die Massenspektrometrie realisiert werden. Die Klassifikation humaner Tumorsubtypen soll schneller, möglichst kostengünstig und trotzdem zuverlässig durchführbar sein. Mit Hilfe von massenspektrometrischen *Fingerprints* konnten bisher erfolgreich Zellklassifikationen von Säugerzellen durchgeführt werden. Eine systematische Untersuchung verschiedener Merkmalsextraktions- und Klassifikations-methoden zur optimierten Klassifikation humaner Tumorsubtypen wurde jedoch bisher noch nicht durchgeführt. Des Weiteren existiert noch kein automatisiertes Verfahren, welches die Wirkung von selektiven und molekular-gerichteten Medikamenten untersucht. Oftmals führt erst ein zeitintensives Austesten verschiedener Einflussgrößen oder gar ein mühsames Durchsuchen des zu analysierenden Datensatzes „per Hand“ zu den erwünschten Ergebnissen.

Dieses Dissertationsprojekt verfolgte daher folgende Ziele:

1. Überführung der MS-Daten in ein MATLAB-kompatibles Datenformat sowie die Optimierung der Vorverarbeitung von massenspektrometrischen *Fingerprints*
2. Entwicklung einer geeigneten Kennzahl zur Bewertung der systematischen Untersuchung verschiedener Merkmalsextraktions- und Klassifikationsmethoden
3. Entwicklung und Optimierung einer Methodik für die Merkmalsextraktion zur Identifizierung geeigneter Biomarker-Kandidaten
4. Entwicklung eines auf spektralen, molekularen *Fingerprints* basierenden Verfahrens zur schnellen Klassifikation von Tumorsubtypen anhand von der in 2 entwickelten Kennzahl
5. Entwicklung einer Methode zur automatisierten Erstellung von Konzentrations-Wirkungskurven zur Bewertung der zellulären Wirkung von zielgerichteten Medikamenten

Es wurde ein objektorientiertes Datenformat entwickelt, welches einen schnellen und einfachen Umgang mit den *Fingerprints* ermöglicht. Die Vorverarbeitung der *Fingerprints* konnte optimiert werden, sodass die Bewertung der Signifikanz einzelner Peaks ausschließlich mit Hilfe statistischer Methoden durchgeführt werden konnte. Zur

systematischen Untersuchung verschiedener Kombinationen einiger Merkmalsextraktions- und Klassifikationsmethoden wurde eine Kennzahl entwickelt werden, welche genaue Ergebnisse liefert und in einer annehmbarer Zeit berechnet werden kann. Anhand von Brustkrebs-, Leukämie- und GIST-Zelllinien wurde ein Verfahren zur schnellen Klassifikation von Tumorsubtypen entwickelt werden, wobei die Zugehörigkeit zum Tumortyp erhalten bleiben konnte. Des Weiteren wurde eine automatisierte Methode zur Bestimmung von Wirkstoffeffekten entwickelt, welche in verschiedenen MS-Datensätzen nach Bereichen mit einer Wirkstoffreaktion und nach Korrelationen zwischen den verschiedenen Datensätzen sucht. Die Algorithmen wurden in MATLAB implementiert werden.

# 7 LITERATURVERZEICHNIS

Alexandrov, T. (2012). MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. BMC bioinformatics *13*, S11.

Allen, D.M. (1971). Mean square error of prediction as a criterion for selecting variables. Technometrics *13*, 469-475.

Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician *46*, 175-185.

Anhalt, J.P., und Fenselau, C. (1975). Identification of bacteria using mass spectrometry. Analytical Chemistry *47*, 219-225.

Avila, C., Almeida, F., und Palmisano, G. (2016). Direct identification of trypanosomatids by matrix-assisted laser desorption ionization–time of flight mass spectrometry (DIT MALDI-TOF MS). Journal of Mass Spectrometry *51*, 549-557.

Balluff, B., Rauser, S., Ebert, M.P., Siveke, J.T., Höfler, H., und Walch, A. (2012). Direct molecular tissue analysis by MALDI imaging mass spectrometry in the field of gastrointestinal disease. Gastroenterology *143*, 544-549. e542.

Bantscheff, M., Hopf, C., Savitski, M.M., Dittmann, A., Grandi, P., Michon, A.-M., Schlegl, J., Abraham, Y., Becher, I., und Bergamini, G. (2011). Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. Nature biotechnology *29*, 255-265.

Barreiro, J.R., Braga, P.A.C., Ferreira, C.R., Kostrzewa, M., Maier, T., Wegemann, B., Boeettcher, V., Eberlin, M.N., und dos Santos, M.V. (2012). Nonculture-based identification of bacteria in milk by protein fingerprinting. Proteomics *12*, 2739-2745.

Bondarenko, P.V., Cockrill, S.L., Watkins, L.K., Cruzado, I.D., und Macfarlane, R.D. (1999). Mass spectral study of polymorphism of the apolipoproteins of very low density lipoprotein. Journal of lipid research *40*, 543-555.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilita (Libreria internazionale Seeber).

Buchanan, C.M., Malik, A.S., und Cooper, G.J. (2007). Direct visualisation of peptide hormones in cultured pancreatic islet alpha-and beta-cells by intact-cell mass spectrometry. Rapid Communications in Mass Spectrometry *21*, 3452-3458.

Cawley, G.C., und Talbot, N.L.C. (2006). Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics *22*, 2348-2355.

Chiu, N.H., Jia, Z., Diaz, R., und Wright, P. (2015). Rapid differentiation of in vitro cellular responses to toxic chemicals by using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Environmental Toxicology and Chemistry *34*, 161-166.

Consolino, L., Longo, D.L., Sciortino, M., Dastrù, W., Cabodi, S., Giovenzana, G.B., und Aime, S. (2016). Assessing tumor vascularization as a potential biomarker of imatinib resistance in gastrointestinal stromal tumors by dynamic contrast-enhanced magnetic resonance imaging. Gastric Cancer, 1-11.

Cornish, T.J., und Cotter, R.J. (1997). High-order kinetic energy focusing in an end cap reflectron time-of-flight mass spectrometer. Analytical chemistry *69*, 4615-4618.

Cover, T.M., und Thomas, J.A. (2006). Elements of Information Theory (New Jersey: John Wiley & Sons).

Croxatto, A., Prod'hom, G., und Greub, G. (2012). Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS microbiology reviews *36*, 380-407.

Davis, R., und Mauer, L. (2010). Fourier transform infrared (FT-IR) spectroscopy: a rapid tool for detection and analysis of foodborne pathogenic bacteria. Current research, technology and education topics in applied microbiology and microbial biotechnology *2*, 1582-1594.

Dawson, M.A., Prinjha, R.K., Dittmann, A., Giotopoulos, G., Bantscheff, M., Chan, W.-I., Robson, S.C., Chung, C.-w., Hopf, C., und Savitski, M.M. (2011). Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. Nature *478*, 529-533.

Deepa, M., Revathy, P., und Student, P. (2012). Validation of Document Clustering based on Purity and Entropy measures. International Journal of Advanced Research in Computer and Communication Engineering *1*, 147-152.

Di Veroli, G.Y., Fornari, C., Goldlust, I., Mills, G., Koh, S.B., Bramhall, J.L., Richards, F.M., und Jodrell, D.I. (2015). An automated fitting procedure and software for dose-response curves with multiphasic features. Scientific reports *5*, 14701.

Diem, M., Griffiths, P.R., und Chalmers, J.M. (2008). Vibrational spectroscopy for medical diagnosis, Vol 40 (Wiley Chichester).

Dong, H., Shen, W., Cheung, M.T.W., Liang, Y., Cheung, H.Y., Allmaier, G., Au, O.K.-C., und Lam, Y.W. (2011). Rapid detection of apoptosis in mammalian cells by using intact cell MALDI mass spectrometry. Analyst *136*, 5181-5189.

Druker, B.J., Guilhot, F., O'brien, S.G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M.W., Silver, R.T., Goldman, J.M., und Stone, R.M. (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. New England Journal of Medicine *355*, 2408-2417.

Du, P., Kibbe, W.A., und Lin, S.M. (2006). Improved peak detection in mass spectrum by incorporating continous wavlet transform-based pattern matching. Bioinformatics *22*, 2059-2065.

Dubreuil, P., Letard, S., Ciufolini, M., Gros, L., Humbert, M., Castéran, N., Borge, L., Hajem, B., Lermet, A., und Sippl, W. (2009). Masitinib (AB1010), a potent and selective tyrosine kinase inhibitor targeting KIT. PloS one *4*, e7258.

Duda, R.O., Hart, P.E., und Stork, D.G. (2001). Pattern Classification Second Edition John Wiley & Sons. New York *58*.

Ebert, M.P., Tänzer, M., Balluff, B., Burgermeister, E., Kretzschmar, A.K., Hughes, D.J., Tetzner, R., Lofton-Day, C., Rosenberg, R., und Reinacher-Schick, A.C. (2012). TFAP2E–DKK4 and chemoresistance in colorectal cancer. New England Journal of Medicine *366*, 44-53.

Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., und Goldstein, T.C. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353-360.

Feng, H.t., Wong, N.S., Sim, L.C., Wati, L., Ho, Y., und Lee, M.M. (2010). Rapid characterization of high/low producer CHO cells using matrix-assisted laser desorption/ionization time-of-flight. Rapid Communications in Mass Spectrometry *24*, 1226-1230.

Finley, R.S. (2003). Overview of targeted therapies for cancer. American journal of health-system pharmacy *60*.

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics *7*, 179-188.

Gámez-Pozo, A., Sánchez-Navarro, I., Nistal, M., Calvo, E., Madero, R., Díaz, E., Camafeita, E., de Castro, J., López, J.A., und González-Barón, M. (2009). MALDI profiling of human lung cancer subtypes. PLoS One *4*, e7731.

Gerber, D.E. (2008). Targeted therapies: a new generation of cancer treatments. Am Fam Physician *77*, 311-319.

Gilbert, E.G., Johnson, D.W., und Keerthi, S.S. (1988). A fast procedure for computing the distance between complex objects in three-dimensional space. Robotics and Automation, IEEE Journal of *4*, 193-203.

Ginsberg, H.N., und Brown, W.V. (2011). Apolipoprotein Ciii (Am Heart Assoc).

Gobom, J., Mueller, M., Egelhofer, V., Theiss, D., Lehrach, H., und Nordhoff, E. (2002). A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS. Analytical chemistry *74*, 3915-3923.

Goldberg, D.E., und Holland, J.H. (1988). Genetic algorithms and machine learning. Machine learning *3*, 95-99.

Goldhirsch, A., Wood, W., Coates, A., Gelber, R., Thürlimann, B., und Senn, H.-J. (2011). Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Annals of oncology, mdr304.

Gross, J.H. (2012). Massenspektrometrie: Ein Lehrbuch (Springer-Verlag).

Guilhaus, M., Mlynski, V., und Selby, D. (1997). Perfect timing: time-of-flight mass spectrometry. Rapid communications in mass spectrometry *11*, 951-962.

Hanrieder, J., Wicher, G., Bergquist, J., Andersson, M., und Fex-Svenningsen, Å. (2011). MALDI mass spectrometry based molecular phenotyping of CNS glial cells for prediction in mammalian brain tissue. Analytical and bioanalytical chemistry *401*, 135-147.

Hartung, J., und Elpelt, B. (2007). Multivariate Statistik: Lehr-und Handbuch der angewandten Statistik (Oldenbourg Verlag).

Henson, R., und Cetto, L. (2005). The MATLAB bioinformatics toolbox. Encyclopedia of genetics, genomics, proteomics and bioinformatics (Natick, MA, USA: The MathWorks, Inc).

Herrero, M., Simó, C., García-Cañas, V., Ibáñez, E., und Cifuentes, A. (2012). Foodomics: MS-based strategies in modern food science and nutrition. Mass spectrometry reviews *31*, 49-69.

Hilario, M., Kalousis, A., Pellegrini, C., und Mueller, M. (2006). Processing and classification of protein mass spectra. Mass spectrometry reviews *25*, 409-449.

Hochberg, Y., und Tamhane, A.C. (2009). Multiple comparison procedures.

Holland, J.H. (1975). Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. Ann Arbor, MI: University of Michigan Press.

Hollinger, M.A. (2007). Introduction to pharmacology (CRC Press).

Jain, A.K., und Dubes, R.C. (1988). Algorithms for clustering data (Prentice-Hall, Inc.).

Kannan, L., Rath, N.C., Liyanage, R., und Lay, J.O. (2007). Identification and Characterization of Thymosin β-4 in Chicken Macrophages Using Whole Cell MALDI-TOF. Annals of the New York Academy of Sciences *1112*, 425-434.

Karas, M., Bachmann, D., und Hillenkamp, F. (1987). Matrix-Assisted Ultraviolet-Lase Desorption of Nonvolatile Compounds. International Journal of Mass Spectrometry and Ion Process *78*, 53-68.

Karger, A., Bettin, B., Lenk, M., und Mettenleiter, T.C. (2010). Rapid characterisation of cell cultures by matrix-assisted laser desorption/ionisation mass spectrometric typing. Journal of virological methods *164*, 116-121.

Kaufman, B., Mackey, J.R., Clemens, M.R., Bapsy, P.P., Vaid, A., Wardley, A., Tjulandin, S., Jahn, M., Lehle, M., und Feyereislova, A. (2009). Trastuzumab plus anastrozole versus anastrozole alone for the treatment of postmenopausal women with human epidermal growth factor receptor 2–positive, hormone receptor–positive metastatic breast cancer: Results from the randomized phase III TAnDEM study. Journal of Clinical Oncology *27*, 5529-5537.

Kempka, M., Sjödahl, J., Björk, A., und Roeraade, J. (2004). Improved method for peak picking in matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry *18*, 1208-1212.

Ketterlinus, R., Hsieh, S.-Y., Teng, S.-H., Lee, H., und Pusch, W. (2005). Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProToolsTM software. Biotechniques *6*, 37.
Kira, K., und Rendell, L.A. (1992). A practical approach to feature selection. Paper presented at: Proceedings of the ninth international workshop on Machine learning.

Kober, S.L., Meyer-Alert, H., Grienitz, D., Hollert, H., und Frohme, M. (2015). Intact cell mass spectrometry as a rapid and specific tool for the differentiation of toxic effects in cell-based ecotoxicological test systems. Analytical and bioanalytical chemistry *407*, 7721-7731.

Koubek, J., Uhlik, O., Jecna, K., Junkova, P., Vrkoslavova, J., Lipov, J., Kurzawova, V., Macek, T., und Mackova, M. (2012). Whole-cell MALDI-TOF: rapid screening method in environmental microbiology. International Biodeterioration & Biodegradation *69*, 82-86.

Kremer, H., Kranen, P., Jansen, T., Seidl, T., Bifet, A., Holmes, G., und Pfahringer, B. (2011). An effective evaluation measure for clustering on evolving data streams. Paper

presented at: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM).

Kruse, U., Bantscheff, M., Drewes, G., und Hopf, C. (2008). Chemical and Pathway Proteomics Powerful Tools for Oncology Drug Discovery and Personalized Health Care. Molecular & Cellular Proteomics *7*, 1887-1901.

Kruskal, W.H., und Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American statistical Association *47*, 583-621.

Kulkarni, P., Kaftan, F., Kynast, P., Svatoš, A., und Böcker, S. (2015). Correcting mass shifts: A lock mass-free recalibration procedure for mass spectrometry imaging data. Analytical and bioanalytical chemistry *407*, 7603-7613.

Lancet, J., Ravandi, F., Ricklis, R., Cripe, L., Kantarjian, H., Giles, F., List, A., Chen, T., Allen, R., und Fox, J. (2011). A phase Ib study of vosaroxin, an anticancer quinolone derivative, in patients with relapsed or refractory acute leukemia. Leukemia *25*, 1808-1814.

Lasch, P., und Naumann, D. (2006). Infrared Spectroscopy in Microbiology. In Encyclopedia of Analytical Chemistry (John Wiley & Sons, Ltd).

Liao, C.C., Ward, N., Marsh, S., Arulampalam, T., und Norton, J.D. (2010). Mass spectrometry protein expression profiles in colorectal cancer tissue associated with clinico-pathological features of disease. BMC cancer *10*, 410.

Liu, H., und Motoda, H. (2008). Computational Methods of Feature Selection (Chapman & Hall / CRC Press).

Liu, H., und Setiono, R. (1995). Chi2: Feature Selection and Discretization of Numeric Attributes. In Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence (Herndon, Virgina), pp. 388-391.

Lloyd, S.P. (1982). Least squares quantization in PCM. Information Theory, IEEE Transactions on *28*, 129-137.

Love, M.I., Huber, W., und Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology *15*, 550.

Lux, A., Müller, R., Tulk, M., Olivieri, C., Zarrabeita, R., Salonikios, T., und Wirnitzer, B. (2013). HHT diagnosis by Mid-infrared spectroscopy and artificial neural network analysis. Orphanet journal of rare diseases *8*, 94.

Mamyrin, B. (1994). Laser assisted reflectron time-of-flight mass spectrometry. International Journal of Mass Spectrometry and Ion Processes *131*, 1-19.

Mann, H.B., und Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, 50-60.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpp, A., Neumann, S., und Pizarro, A.D. (2011). mzML—a community standard for mass spectrometry data. Molecular & Cellular Proteomics *10*, R110. 000133.

Marty, M., Cognetti, F., Maraninchi, D., Snyder, R., Mauriac, L., Tubiana-Hulin, M., Chan, S., Grimes, D., Antón, A., und Lluch, A. (2005). Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2–positive metastatic breast cancer administered as first-line treatment: the M77001 study group. Journal of Clinical Oncology *23*, 4265-4274.

Marvin-Guy, L.F., Duncan, P., Wagniere, S., Antille, N., Porta, N., Affolter, M., und Kussmann, M. (2008). Rapid identification of differentiation markers from whole epithelial cells by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry and statistical analysis. Rapid Communications in Mass Spectrometry *22*, 1099-1108.

McCulloch, W.S., und Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics *5*, 115-133.

McLachlan, G. (2004). Discriminant analysis and statistical pattern recognition, Vol 544 (John Wiley & Sons).

Miura, D., Fujimura, Y., und Wariishi, H. (2012). In situ metabolomic mass spectrometry imaging: recent advances and difficulties. Journal of proteomics *75*, 5052-5060.

Moja, L., Tagliabue, L., Balduzzi, S., Parmelli, E., Pistotti, V., Guarneri, V., und D'Amico, R. (2012). Trastuzumab containing regimens for early breast cancer. The Cochrane Library.

Motulsky, H., und Christopoulos, A. (2004). Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting (Oxford University Press).

Munteanu, B., Meyer, B.r., von Reitzenstein, C., Burgermeister, E., Bog, S., Pahl, A., Ebert, M.P., und Hopf, C. (2014). Label-free in situ monitoring of histone deacetylase drug target engagement by matrix-assisted laser desorption ionization-mass spectrometry biotyping and imaging. Analytical chemistry *86*, 4642-4647.

Munteanu, B., Reitzenstein, C.v., Hänsch, G.M., Meyer, B., und Hopf, C. (2012). Sensitive, robust and automated protein analysis of cell differentiation and of primary human blodd cells by intact cell MALDI mass spectrometry biotyping. Analytical and Bioanalytical Chemistry *404*, 2277-2286.

Nannini, M., Biasco, G., Astolfi, A., und Pantaleo, M.A. (2013). An overview on molecular biology of KIT/PDGFRA wild type (WT) gastrointestinal stromal tumours (GIST). Journal of medical genetics *50*, 653-661.

Norris, J.L., und Caprioli, R.M. (2013). Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. Chemical reviews *113*, 2309.

Ouedraogo, R., Daumas, A., Ghigo, E., Capo, C., Mege, J.-L., und Textoris, J. (2012). Whole-cell MALDI-TOF MS: A new tool to assess the multifaceted activation of macrophages. Journal of proteomics *75*, 5523-5532.

Ouedraogo, R., Flaudrops, C., Amara, A.B., Capo, C., Raoult, D., und Mege, J.-L. (2010). Global analysis of circulating immune cells by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. PLoS One *5*, e13691.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., und Park, T. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine *351*, 2817-2826.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 240-242.

Pearson, K. (1901). Principal components analysis. The London, Edinburgh and Dublin Philosophical Magazine and Journal *6*, 566.

Pearson, K. (1992). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. In Breakthroughs in Statistics (Springer), pp. 11-28.

Pecks, U., Kirschner, I., Wölter, M., Schlembach, D., Koy, C., Rath, W., und Glocker, M.O. (2014). Mass spectrometric profiling of cord blood serum proteomes to distinguish infants with intrauterine growth restriction from those who are small for gestational age and from control individuals. Translational Research *164*, 57-69.

Peng, H., Long, F., und Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on *27*, 1226-1238.

Portevin, D., Pflüger, V., Otieno, P., Brunisholz, R., Vogel, G., und Daubenberger, C. (2015). Quantitative whole-cell MALDI-TOF MS fingerprints distinguishes human monocyte sub-populations activated by distinct microbial ligands. BMC biotechnology *15*, 1.

Povey, J.F., O'Malley, C.J., Root, T., Martin, E.B., Montague, G.A., Feary, M., Trim, C., Lang, D.A., Alldread, R., und Racher, A.J. (2014). Rapid high-throughput characterisation, classification and selection of recombinant mammalian cell line phenotypes using intact cell MALDI-ToF mass spectrometry fingerprinting and PLS-DA modelling. Journal of biotechnology *184*, 84-93.

Rammohan, A., Sathyanesan, J., Rajendran, K., Pitchaimuthu, A., Perumal, S.-K., Srinivasan, U., Ramasamy, R., Palaniappan, R., und Govindan, M. (2013). A gist of gastrointestinal stromal tumors: A review. World journal of gastrointestinal oncology *5*, 102.

Ramsden, N., Perrin, J., Ren, Z., Lee, B.D., Zinn, N., Dawson, V.L., Tam, D., Bova, M., Lang, M., und Drewes, G. (2011). Chemoproteomics-based design of potent LRRK2-selective lead compounds that attenuate Parkinson's disease-related toxicity in human neurons. ACS chemical biology *6*, 1021-1028.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association *66*, 846-850.

Rendón, E., Abundez, I., Arizmendi, A., und Quiroz, E. (2011). Internal versus external cluster validation indexes. International Journal of computers and communications *5*, 27-34.
Rokach, L., und Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook (Springer), pp. 321-352.

Ross, J.S., Schenkein, D.P., Pietrusko, R., Rolfe, M., Linette, G.P., Stec, J., Stagliano, N.E., Ginsburg, G.S., Symmans, W.F., und Pusztai, L. (2004). Targeted therapies for cancer 2004. American journal of clinical pathology *122*, 598-609.

Ruh, H., Salonikios, T., Fuchser, J., Schwartz, M., Sticht, C., Hochheim, C., Wirnitzer, B., Gretz, N., und Hopf, C. (2013). MALDI imaging MS reveals candidate lipid markers of polycystic kidney disease. Journal of Lipid Research *54*, 2785-2794.

Sakai, M., Martinez-Arguelles, D.B., Patterson, N.H., Chaurand, P., und Papadopoulos, V. (2015). In search of the molecular mechanisms mediating the inhibitory effect of the GnRH antagonist degarelix on human prostate cell growth. PloS one *10*, e0120670.

Salonikios, T. (2012). Ortsabhängige Verarbeitung und Analyse von zweidimensionalen Massenspektrometerdaten. (Institut für Digitale Signalverarbeitung, Hochschule Mannheim).

Savitzky, A., und Golay, M.J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry *36*, 1627-1639.

Schwamb, S., Munteanu, B., Meyer, B., Hopf, C., Hafner, M., und Wiedemann, P. (2013). Monitoring CHO cell cultures: cell stress and early apoptosis assessment by mass spectrometry. Journal of biotechnology *168*, 452-461.

Schwamborn, K. (2012). Imaging mass spectrometry in biomarker discovery and validation. Journal of proteomics *75*, 4990-4998.

Sripada, S.C., und Rao, D.M.S. (2011). Comparison of purity and entropy of k-means clustering and fuzzy c means clustering. Indian journal of computer science and engineering *2*, 343-346.

Student (1908). The probable error of a mean. Biometrika, 1-25.

Theodoridis, S., Pikrakis, A., Koutroumbas, K., und Cavouras, D. (2010). Introduction to Pattern Recognition: A MATLAB Approach (London, UK: Academic Press).

Torrisi, R., Bagnardi, V., Pruneri, G., Ghisini, R., Bottiglieri, L., Magni, E., Veronesi, P., D'Alessandro, C., Luini, A., und Dellapasqua, S. (2007). Antitumour and biological effects of letrozole and GnRH analogue as primary therapy in premenopausal women with ER and PgR positive locally advanced operable breast cancer. British journal of cancer *97*, 802-808.

Valletta, E., Kučera, L., Prokeš, L., Amato, F., Pivetta, T., Hampl, A., Havel, J., und Vaňhara, P. (2016). Multivariate Calibration Approach for Quantitative Determination of Cell-Line Cross Contamination by Intact Cell Mass Spectrometry and Artificial Neural Networks. PloS one *11*, e0147414.

van Rijsbergen, C. (1979). Information retrieval (2nd edit.) butterworths. London, UK.

van Veen, S.Q., Claas, E., und Kuijper, E.J. (2010). High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization-time of flight mass spectrometry in conventional medical microbiology laboratories. Journal of clinical microbiology *48*, 900-907.

Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association *58*, 236-244.

Wei, L.J. (1981). Asymptotic Conservativeness and Efficiency of Kruskal-Wallis Test for K Dependent Samples. Journal of the American Statistical Association *76*, 1006-1009.

Weickhardt, C., Moritz, F., und Grotemeyer, J. (1996). Time-of-flight mass spectrometry: State-of the-art in chemical analysis and molecular science. Mass spectrometry reviews *15*, 139-162.

Wenner, M.M., Wilson, T.E., Davis, S.L., und Stachenfeld, N.S. (2011). Pharmacological curve fitting to analyze cutaneous adrenergic responses. Journal of Applied Physiology *111*, 1703-1709.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics bulletin, 80-83.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., und Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics *19*, 1636-1643.

Wu, H.-C., Chang, D.-K., und Huang, C.-T. (2006). Targeted therapy for cancer. J Cancer Mol *2*, 57-66.

Yang, C., He, Z., und Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC bioinformatics *10*, 4.

Zhang, X., Scalf, M., Berggren, T.W., Westphall, M.S., und Smith, L.M. (2006). Identification of mammalian cell lines using MALDI-TOF and LC-ESI-MS/MS mass spectrometry. Journal of the American Society for Mass Spectrometry *17*, 490-499.

## Publikationen

Ruh, H.; Salonikios, T.; Fuchser, J.; Schwartz, M.; Sticht, C.; Hochheim, C.; Wirnitzer, B.; Gretz, N., und Hopf, C. (2013). MALDI imaging MS reveals candidate lipid markers of polycystic kodney disease. J Journal of Lipid Research *54*, 2785-2794

Lux, A., Müller, R., Tulk, M., Olivieri, C., Zarrabeita, R., Salonikios, T., und Wirnitzer, B. (2013). HHT diagnosis by Mid-infrared spectroscopy and artificial neural network analysis. Orphanet journal of rare diseases *8*, 1-15.

*In Vorbereitung:*

Munteanu, B.; Salonikios, T.; Zolg, D.; Weigt, d.; Noll, E.; von Reitzenstein, C.; Paulitschke, V.; Trumpp, A.; Levesque, M.; Sprick, M.; Küster, B. und Hopf. C. (2017). Automated, Label-Free Drug-Response Profiling by Whole-Cell-MALDI-MS Biotyping Reveals [XXXXX] as a Mass Spectrometry [Imaging] Accessible Cell Viability Marker in Cancer Cells.

# 8 LEBENSLAUF

PERSONALIEN

| | |
|---|---|
| Name und Vorname: | Salonikios, Theresia |
| Geburtsdatum: | 09.09.1982 |
| Geburtsort: | Freiburg im Breisgau |
| Familienstand: | ledig |

UNIVERSITÄRER WERDEGANG

| | |
|---|---|
| seit 08/2012 | **Institut für Medizintechnologie der Universität Heidelberg und der Hochschule Mannheim**<br>Dissertation<br>Automatische Datenanalyse von massenspektrometrischen Signaturen zur Klassifikation von Krebszellen und Bestimmungen von Wirkstoffwirkungen |
| 03/2010 – 08/2012 | **Hochschule Mannheim**<br>Studiengang Informationstechnik |
| 1. August 2012 | Abschluss **Master of Science**, Note: 1,2 |
| 09/2006 – 02/2010 | **Hochschule Mannheim**<br>Studiengang Technische Informatik |
| 22. März 2010 | Abschluss **Bachelor of Science**, Note: 2,5 |
| 09/2002 – 08/2006 | **Universität Mannheim**<br>Fachrichtung Technische Informatik |

SCHULISCHER WERDEGANG

| | |
|---|---|
| 08/1993 – 07/2002 | **Carl-Benz-Gymnasium Ladenburg** |
| 5. Juni 2002 | Abitur, Note: 3,1 |

BERUFLICHER WERDEGANG

| | |
|---|---|
| seit 05/2017 | **Universitätsklinikum Heidelberg**<br>*Radiologische Klinik, Nuklearmedizin*<br>Wissenschaftliche Mitarbeiterin |
| 06/2016 – 12/2016 | **Hochschule Mannheim**<br>*Zentrum für Angewandte Forschung - „Applied Biomedical Mass Spectrometry“ (ZAFH-ABIMAS)*<br>Wissenschaftliche Mitarbeiterin |
| 08/2015 – 05/2016 | **Hochschule Mannnheim**<br>*Projekt: MITIGATE (Closed-loop Molecular Enviroment for Minimally Invasive Treatment of Patients with Metastatic Gastrointestinal Stromal Tumours)*<br>Wissenschaftliche Mitarbeiterin |

# 9 DANKSAGUNG

An dieser Stelle möchte ich mich bei allen bedanken, die zum Erfolg dieses interdisziplinären Dissertationsprojekts beigetragen haben.

Herrn Prof. Dr. Carsten Hopf danke ich für die Überlassung des Themas dieser Arbeit und die Bereitstellung des Arbeitsplatzes. Diese Doktorarbeit wurde erst durch seine Unterstützung als Doktorvater möglich.

Zudem möchte ich mich bei Dr. Carolina von Retzenstein, Dr. Bogdan Munteanu und Jan-Hinrich Rabe für die Hilfe in der Zellkultur und die Bereitstellung der massenspektrometrischen Datensätze bedanken.

Besonders danke ich auch Matthias Schwartz für seine Diskussionsbereitschaft, die vielen praktischen Tipps und Anregungen für den IT-Teil dieser Arbeit. Ebenso danke ich Frau Dr. Hermelindis Ruh, die mir zum Verständnis zahlreicher biomedizinischer Fragestellungen verhalf und mir viele sachliche Hinweise gab.

Für die finanzielle Unterstützung und das dadurch ausgedrückte Vertrauen möchte ich gerne der Albert-und-Anneliese-Konanz-Stiftung und Prof. Dr. Carsten Hopf danken. Ebenso bedanke ich mich für die Förderung durch das EU-FP7-Projekt „MITIGATE“.

Diese Arbeit wurde durch das Ministerium für Wissenschaft und Kultur, Baden-Württemberg (INST 874/2-1 LAGG.) unterstützt, und im Rahmen von “ZAFH ABIMAS”, gemeinschaftlich durch das ZO IV der Landesstiftung Baden-Württemberg und den Europäischen Fonds für regionale Entwicklung (EFRE) getragen.