In silico Fragmentierung für die computergestützte Auswertung von Tandem-Massenspektrometrie Daten

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlichen Fakultät III
(Institut für Informatik)
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Sebastian Wolf
geb. am 24. Februar 1984 in Karl-Marx-Stadt (Chemnitz)

Halle (Saale), 29. Februar 2012

Gutachter:
1. Prof. Dr. Matthias Müller-Hannemann
2. Prof. Dr. Oliver Kohlbacher

Vorgelegt am: 29. Februar 2012
Datum der Verteidigung: 1. Juni 2012

Danksagung

Besonders bedanken möchte ich mich bei Herrn Prof. Dr. Matthias Müller-Hannemann und Dr. Steffen Neumann, die mit viel Engagement und guten Ideen meine Dissertation betreut haben. Weiter danke ich meinen Kollegen Franziska Taruttis, Michael Gerlich, Carsten Kuhl und Christian Hildebrandt für die vielen hilfreichen Diskussionen. Auch möchte ich Frau Dr. Nadine Strehmel, Herrn Dr. Stephan Schmidt und PD Dr. Wolfgang Brandt für die zahlreichen Gespräche und Anregungen danken.

Ein besonderer Dank gilt meiner Frau Franziska Wolf, die mich in der gesamten Zeit großartig unterstützt hat.

Inhaltsverzeichnis

1 Einführung

1.1 Metabolomik

Die *Metabolomik* befasst sich mit der Erforschung der am Stoffwechsel (*Metabolismus*) beteiligten Substanzen (*Metabolite*) und wurde in Analogie zu den Begriffen Genomik und Proteomik geprägt. Die Gesamtheit aller *Metabolite* wird als *Metabolom* bezeichnet. Deren Identifizierung und Quantifizierung ist somit das Anwendungsgebiet der *Metabolomik*.

Beispiele für die genutzten analytischen Methoden zur Separierung sind die Gaschromatographie (GC) und Flüssigkeitschromatographie (HPLC und UPLC). Die am weitesten verbreiteten Methoden zur Detektion von Metaboliten sind die Kernspinresonanzspektroskopie (NMR-Spektroskopie) und die Massenspektrometrie (MS).

1.2 Strukturaufklärung mittels LC/MS

Ein typisches Vorgehen zur Identifizierung einer unbekannten Verbindung ist in Abbildung 1.1 dargestellt. Damit nicht alle Verbindungen zur gleichen Zeit am Massenspektrometer detektiert werden, und dadurch Koelutionen und somit Ungenauigkeiten in der Summenformelbestimmung entstehen, wird die Probe durch ein spezielles Einlasssystem nach ihrer Polarität aufgetrennt (zum Beispiel LC/MS). Die Probe wird danach im Massenspektrometer gemessen. Das Gerät nimmt viele hundert bis tausende Spektren zu verschiedenen Zeitpunkten (Retentionszeiten) auf. Diese Spektren enthalten Signale (Ionen), welche einen Rückschluss auf die Masse der gemessenen Verbindungen ermöglichen. Strukturhinweise von interessanten Peaks, zum Beispiel aus verschiedenen Experimenten hoch- oder herunterregulierte,

bekommt man durch die Tandem-Massenspektrometrie (MS/MS) Messung. Hierfür wird das entsprechende Vorläuferion (Peak) ausgewählt (Abbildung 2.2) und durch Kollision mit einem Stoßgas (CID - „collision induced dissociation“) unter Verwendung verschiedener Stoßenergien fragmentiert. Das resultierende MS/MS Spektrum enthält Fragmentionen (Peaks), die Substrukturen des Analyten sind.

Durch die hohe Massenauflösung neuer Instrumente kann außerdem die Summenformel von Verbindungen bestimmt oder eingegrenzt werden. Mit diesen MS/MS Spektrum wird im Anschluss eine Spektrendatenbank (siehe Kapitel 2.2.2) durchsucht und womögliche Treffer durch Messung einer Referenzsubstanz bestätigt. In den meisten Fällen reicht eine Spektrensuche (siehe Kapitel 2.2.2) nicht aus, da diese Datenbanken noch nicht genügend Verbindungen (ca. 15 000) enthalten (Abschnitt 2.2.2). Es wird vermutet, dass es alleine in Pflanzen über 200 000 *Metabolite* vorhanden sind [PG00]. Daher werden auch Strukturdatenbanken (siehe Abschnitt 2.2.1), die keine Spektreninformation sondern unter anderem Summenformel und Struktur enthalten, abgefragt. Eine solche Suche liefert keine bis viele tausend Strukturen, was das Identifizieren von Analyten äußerst schwierig und sehr zeitaufwendig macht. Außerdem gibt es noch viele unbekannte Verbindungen, die noch nicht in den Datenbanken enthalten sind. In Frage kommende Kandidaten können gekauft oder synthetisiert werden und durch eine erneute MS/MS Messung, können die Retentionszeit und das Spektrum mit dem Analyten abgeglichen und verifiziert werden.

1.3 Ziele und Aufbau der Arbeit

Ziel dieser Arbeit ist es, die Strukturaufklärung durch MS/MS mit Hilfe von Verbindungen aus verschiedensten Quellen zu erleichtern und zu beschleunigen. Dafür werden neben Moleküldatenbanken auch Strukturgeneratoren genutzt, um geeignete Kandidatenmoleküle zu finden. MetFrag [WSMHN10], die im Rahmen dieser Arbeit entwickelte Software, führt eine Vorverarbeitung und Fragmentierung der Kandidaten durch und ordnet den generierten Fragmenten die gemessenen Peaks zu. Das Ergebnis ist eine nach Score sortierte Liste von Kandidaten. MetFrag erzielt mit hochauflösenden MS/MS Daten bessere Ergebnisse als MassFrontier 4 und

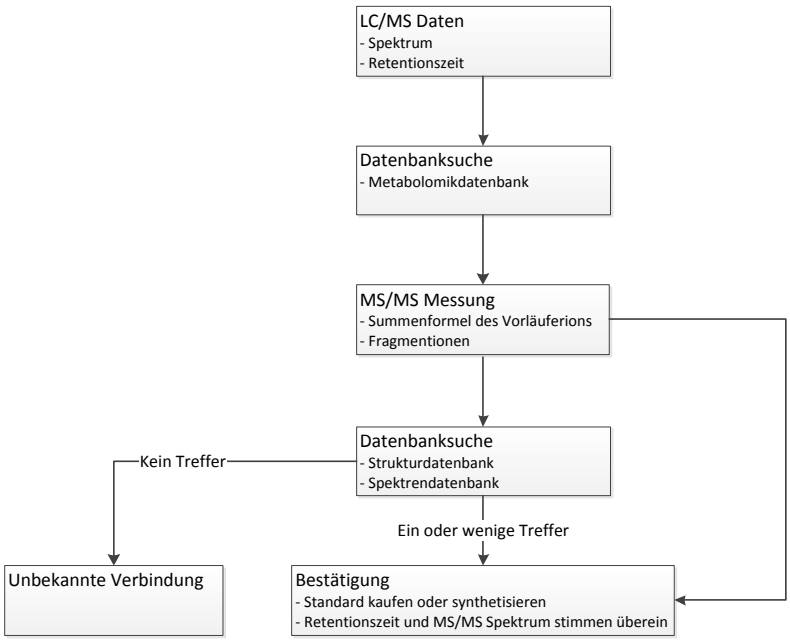


Abbildung 1.1. Strategie zur Identifizierung von Verbindungen nach [WHD+08]. Eine erste LC/MS Messung sucht nach interessanten Peaks, auf die man in Metabolomikdatenbanken Hinweise finden kann. Die anschließende Identifizierung des Analyten kann mit einer MS/MS Messung durchgeführt werden. Mit dem resultierenden Spektrums können Kandidaten aus Spektren- und Strukturdatenbanken gesucht werden. Zur Verifizierung des Kandidaten wird eine Messung mit einem Standard durchgeführt.

kann darüber hinaus mit GC/MS Daten verwendet werden. Außerdem ist im Rahmen dieser Arbeit MassStruct [HWN11] entwickelt worden, was es ermöglicht durch MetFrag annotierte Peak → Fragment Assoziationen zu lernen und dadurch eine Reihenfolge der Kandidaten zu bestimmen. Da zuerst besonders relevante Verbindungen zurückgeliefert werden, kann die Laufzeit von MetFrag verringert werden. In [SGK+] wurde neben dem MetFrag Score auch andere (experimentellen) Daten verwendet, um potentielle Kandidaten auszuschließen bzw. in die engere Auswahl zu nehmen. Es wurde von [SGK+] gezeigt, wie ein solches Vorgehen zu einer erfolgreichen Identifizierung von Unbekannten führen kann. Durch die im Rahmen dieser Dissertation entwickelte Software kann der zeitaufwendige Schritt der Identifizierung des gemessenen Analyten erheblich beschleunigt und erleichtert werden.

Diese Arbeit beschäftigt sich mit den Grundlagen (Kapitel 2) der Massenspektrometrie von niedermolekularen Verbindungen, beschreibt deren Identifizierung, gibt eine Übersicht über die wichtigsten Molekülrepräsentationen und der verwendeten Algorithmen. Außerdem wird die Geometrieoptimierung von Verbindungen genauer betrachtet. Weiterhin werden bereits verfügbare Algorithmen zur Fragmentvorhersage vorgestellt. Kapitel 3 beschreibt die Architektur und Implementation von MetFrag sowie die entwickelte Weboberfläche. MassStruct, ein Verfahren zur intelligenten Kandidatensuche wird in Abschnitt 3.3 vorgestellt. Im darauffolgenden 4. Kapitel wird beschrieben, wie die Scoring Funktion von MetFrag aufgebaut, optimiert und evaluiert wurde. Maße zur Evaluierung von Software zur Fragmentvorhersage und Vergleiche von MetFrag mit kommerzieller Software auf GC/MS und MS/MS Daten sind auch Bestandteil dieses Kapitels. Außerdem werden die Ergebnisse von MassStruct vorgestellt. Schließlich werden Anwendungsgebiete von MetFrag in der Massenspektrometrie Community beschrieben.

2 Grundlagen

Das folgende Kapitel gibt einen Überblick über die Massenspektrometrie. Dabei wird vor allem auf MS/MS eingegangen, da dies das hauptsächliche Anwendungsgebiet, der im Rahmen der Arbeit entwickelten Software, darstellt. Außerdem wird das „Chemistry Development Kit“ [SHK+03], eine in Java geschriebene Cheminformatik Bibliothek (siehe Kapitel 2.4), näher vorgestellt und verschiedene Molekülrepräsentationen beschrieben. Im letzten Teil dieses Kapitels werden Methoden zur Geometrieoptimierung von Verbindungen eingeführt, die später in der Vorverarbeitung von Molekülen eine Rolle spielen. In der vorliegenden Arbeit werden für die Moleküle eindeutige PubChem CIDs verwendet, da diese frei verfügbar sind und sich daraus auch die CAS („Chemical Abstracts Service“) Nummer ableiten lässt. Dies ist zum Beispiel mit dem „Chemical Identifier Resolver“[1] einem Service zum Übersetzen von chemischen IDs möglich.

2.1 Massenspektrometrie (MS)

Massenspektrometer sind Instrumente, um kleinste Konzentrationen von chemischen Verbindungen in einer Probe festzustellen und zu analysieren. Weiterhin sind es wichtige Werkzeuge zur Strukturaufklärung. Die heutigen Anwendungsbereiche umfassen vor allem biochemische Fragestellungen (z.B. Proteom-, Metabolom- und Pharmaforschung), kriminaltechnische Untersuchungen, sowie Lebensmittel- und Dopingkontrollen. Diese Arbeit befasst sich mit der Identifizierung von niedermolekularen Molekülen, bei der die Massenspektrometrie die Methode der Wahl ist [Dun08].

[1] `http://cactus.nci.nih.gov/chemical/structure` - Abgerufen im November 2011

Die wesentlichen Elemente eines Massenspektrometers nach [BS05] sind das Einlasssystem für die Probe, die Ionenquelle, der Analysator und der Detektor.

Über das Einlasssystem des Massenspektrometers wird die Probe in das Hochvakuum des Instrumentes eingebracht. In der Ionenquelle findet der Prozess der Ionisierung statt, welcher positiv bzw. negativ geladene Teilchen, die Ionen, erzeugt, da nur diese im Massenspektrometer detektiert werden können. Der Analysator dient zur Trennung der Ionen nach ihrem Masse-zu-Ladung Verhältnis (m/z). Das Spektrum einer gemessenen Probe wird durch einen Detektor aufgenommen und erlaubt Rückschlüsse auf dessen Masse.

In der Massenspektrometrie sind unterschiedliche Massenbegriffe gebräuchlich. Die *Nominalmasse* entspricht der Summe der auf ganze Zahlen gerundeten Masse der Elemente eines Moleküls. Die *exakte Masse* addiert (mit Nachkommastellen) die Massen individueller Isotope einer Verbindung. Im Gegensatz dazu nimmt die *monoisotopische Masse* für die Berechnung das am häufigsten auftretende Isotop der Elemente einer Verbindung. Desweiteren wird bei der *durchschnittlichen Masse* die relative Auftrittswahrscheinlichkeit der Isotope mit einbezogen. Dabei gilt folgende Definition:

$$1\ \mathrm{u} = 1\ \mathrm{Da} = 1,660540 \cdot 10^{-27}\mathrm{kg} = \frac{m(^{12}C)}{12}\ .$$

Tabelle 2.1 zeigt die unterschiedlichen Massen am Beispiel von CH_3Cl . Chlor besteht zu 24,23% aus ^{37}Cl mit einer exakten Masse von 36,966 Da und zu 75,77% aus ^{35}Cl mit 34,969 Da (durchschnittliche Masse: $34,969 \cdot 0,7577 + 36,966 \cdot 0,2423 \approx 35,453$ Da). Die durchschnittliche Masse von Kohlenstoff beträgt 12,011 Da (^{12}C = 12,0 Da mit 98,93% und ^{13}C = 13,003 Da mit 1,07% $\rightarrow 12 \cdot 0,9893 + 13,003 \cdot 0,0107 \approx 12,011$ Da). Wasserstoff besitzt eine durchschnittliche Masse von 1,008 Da (^{1}H = 1,008 Da mit 99,989% und ^{2}H = 2,014 Da mit 0,0115% $\rightarrow 1,008 \cdot 0,99989 + 2,014 \cdot 0,000115 \approx 1,008$ Da).

Alle Massen (gerundet) und Isotopenverhältnisse stammen aus den Veröffentlichungen [BW11, AWT03, WB09].

durchschnittliche Masse	monoisotopische Masse	Nominalmasse
$12,011 + (3 \cdot 1,008) + 35,453$ $\approx 50,488$ Da	$12,0 + (3 \cdot 1,008) + 34,969$ $\approx 49,993$ Da	$12 + (3 \cdot 1) + 35$ $= 50$ Da

Tabelle 2.1. Berechnung der durchschnittlichen, monoisotopischen und Nominalmasse von CH_3Cl

2.1.1 Probentrennungs- und Ionisierungsverfahren

LC/MS Systeme benutzen zur Auftrennung komplexer Stoffgemische einen Flüssigchromatographen (LC - „liquid chromatography“), der mit dem Massenspektrometer gekoppelt ist. In diesem Verfahren wird die gelöste Probe (auch Gemische) nach ihrer Polarität aufgetrennt und dann durch ein Ionisierungsverfahren, zum Beispiel Elektrospray Ionisierung (ESI), fein zerstäubt.

Als Beispiele für einen Flüssigchromatographen seien hier die „high performance liquid chromatography“ (HPLC) und neuere „ultra performance liquid chromatography“ (UPLC) genannt. Letztere kann unter anderem einen höheren Druck in der Säule aufbauen und besitzt mit einer Partikelgröße von $1,7\mu$m wesentlich feineres chromatographisches Material im Vergleich zur HPLC. Letztendlich ermöglicht dies eine bessere und stabilere Trennung der Verbindungen, wie in [CTMD05] gezeigt wurde.

Im Gegensatz zum Flüssigchromatographen kann die Probe auch mit einem Gaschromatographen verdampft werden. Die Trennung erfolgt hier nach dem Siedepunkt. Sie wird durch ein Trägergas (meist Helium) und ein Trägermaterial in der Säule hervorgerufen. Ein solches GC/MS wurde unter anderen von [FKD+00] verwendet. Typischerweise wird die Probe durch Elektronenstoßionisation (EI) ionisiert, wodurch im Gegensatz zur sanften ESI, der Analyt stark fragmentiert wird. Weitere Ionisierungsmethoden und Details zur ESI werden im folgenden Kapitel näher betrachtet.

Ionisierungstechniken

Massenspektrometer können nur Ionen, d.h. elektrisch geladene Teilchen, messen. Daher muss die zu messende Probe ionisiert werden, bevor sie im Gerät analy-

siert werden kann. Es gibt viele unterschiedliche Ionisierungsverfahren, da nicht jede Technik alle Klassen von Verbindungen, zum Beispiel polare oder unpolare, ionisieren kann. Elektronenstoßionisation (EI), chemische Ionisation (CI), MALDI (Matrix-Assisted Laser Desorption Ionisation) oder auch Atmosphärendruck Ionisation (API) sind Beispiele verschiedener Ionisierungsmethoden. Letztere beinhaltet beispielsweise „atmospheric pressure chemical ionization“ (APCI) und die Elektrospray Ionisierung (ESI), die im folgenden genauer betrachtet wird.

Für die Entdeckung dieser weichen Ionisierungsmethode, die im Gegensatz zur (harten) Elektronenstoßionisation (EI) Moleküle mit nur geringer Energie anregt, wurde 2002 der Nobelpreis für Chemie an John B. Fenn [FMM+89], als einer von drei Preisträgern, vergeben. Diese Methode transferiert unter Atmosphärendruck Ionen aus einer Lösung in die Gasphase (Desolvatisierung). Dabei finden folgende Prozesse statt (nach [LEL06]):

1. Bildung kleiner, geladener Tröpfchen aus Elektrolyten

2. kontinuierlicher Lösungsmittelverlust durch Verdampfung

3. wiederholter, spontaner Zerfall der Tröpfchen in Mikrotröpfchen

4. Desolvatisierung von Molekülen beim Transfer in das Massenspektrometer

Abbildung 2.1 zeigt beispielhaft diesen Prozess für den positiven Modus, d.h. es werden nur positiv geladene Ionen im Massenspektrometer beschleunigt. Die gelöste Probe passiert die Zerstäuberkapillare, wobei das elektrische Feld zwischen Kapillarspitze und Massenspektrometer ein Flüssigkeitskonus bildet. Daraus entstehen mikrometergroße Tröpfchen, die durch zunehmende Verdampfung des Lösungsmittels schrumpfen. Durch Abnahme des Tröpfchenradius wird die Oberfläche mit positiven Ladungen angereichert. Anschließend gelangen die Ionen in das Vakuum des Massenspektrometers. Der entsprechende negative Modus kann durch Umpolung des elektrischen Feldes erreicht werden.

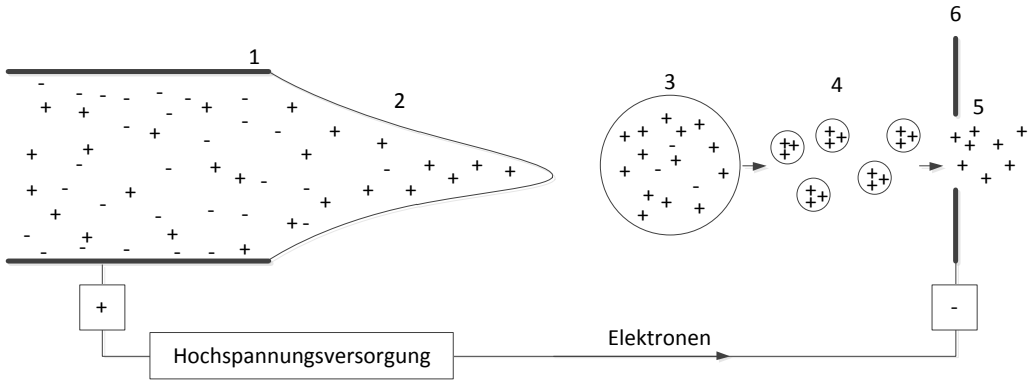


Abbildung 2.1. Bildung positiv geladener Ionen mit einem Elektrospray nach [BS05]: Das elektrisches Feld zwischen Kapillarspitze (1) und Gegenelektrode (6) erzeugt den sogenannten Taylor-Konus (2). Das daraus entstehende Tröpfchen (3) wird durch Verdampfung des Lösungsmittels kleiner (4). Anschließende Weiterleitung der Ionen (5) an das Massenspektrometer.

2.1.2 Tandem-Massenspektrometrie (MS/MS)

Spezielle Massenspektrometer können neben der Masse des Molekülions (MS1 - ohne bzw. nur geringe Fragmentierung) auch Fragmentionen von Vorläuferionen erzeugen. Dadurch kann auf die Struktur der gemessenen Verbindung geschlossen werden. Das Prinzip einer solchen Tandem-Massenspektrometrie (MS/MS) Messung ist in Abbildung 2.2 veranschaulicht. Das obere Spektrum zeigt einen MS1 Peak, von dem zusätzlich zur Masse noch die Struktur von Interesse ist. Daher selektiert das MS/MS Gerät in einem ersten Schritt die Masse dieses Peaks (Vorläuferion - „precursor ion“), das durch eine Fragmentierung (CID) in kleinere Ionen aufgespalten wird, die im MS/MS Spektrum als Peaks erkennbar sind.

Der Aufbau eines solchen MS/MS Gerätes ist am Beispiel eines QqTOF Massenspektrometers in Abbildung 2.3 dargestellt. Ein solches Instrument besteht aus drei Quadrupolen (Q0, Q1, q2) gefolgt von einem Flugzeitanalysator (TOF). Die Ionen werden durch eine ESI-Schnittstelle in das Vakuum des Massenspektrometers geleitet und im ersten Quadrupol (Q0) fokussiert. Die Selektion des Vorläuferions findet schließlich im Q1 statt. In der nachfolgenden Kollisionszelle (q2) werden die selektierten Ionen mit Hilfe eines Stoßgases (zum Beispiel Argon oder Stickstoff)

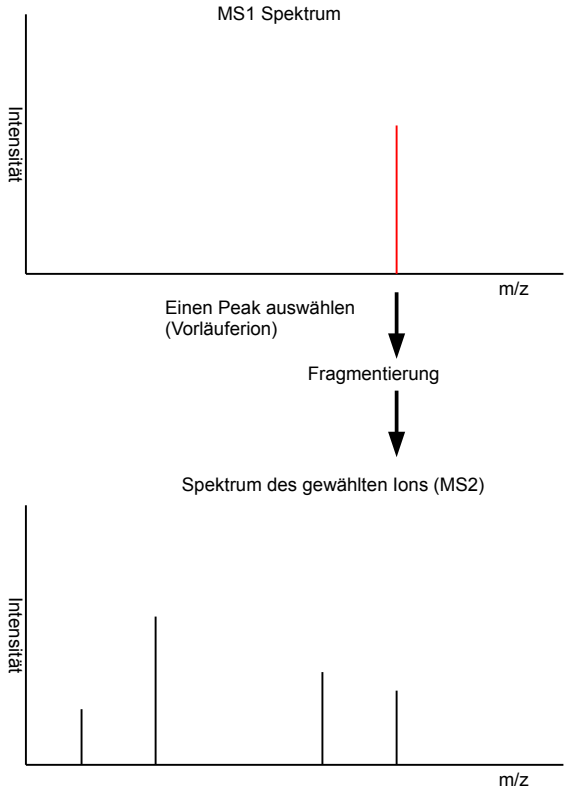


Abbildung 2.2. Prinzip der MS/MS nach [HS07]: Ausgangspunkt ist ein MS1 Spektrum mit einem zu analysierenden Peak (rot). Um die Struktur des Ions aufklären zu können wird eine MS/MS Messung durchgeführt, bei welcher die Struktur der Verbindung fragmentiert wird. Das resultierende Spektrum enthält Peaks (Fragmentionen), die Hinweise auf die gemessene Struktur geben.

fragmentiert. Diese kollisions-induzierte Fragmentierung (CID) kann in ihrer Stärke durch Regelung der Kollisionsenergie variiert werden. Je höher diese gewählt ist und je nachdem wie stabil das Molekül ist, desto stärker fragmentiert es. Im Flugzeitanalysator (TOF) werden die Ionen anschließend nach ihrem Masse-zu-Ladung Verhältnis, durch die unterschiedliche Flugzeit, aufgetrennt und schließlich im Detektor gemessen. Wenn statt eines TOF als Analysator ein weiterer Quadrupol eingesetzt wird, dann spricht man von einem sogenannten Triple-Quadrupol Gerät, wobei dieses Instrument eine geringere Massengenauigkeit hat (siehe Kapitel 2.1.3).

Ein Beispiel für ein derartiges MS/MS Spektrum ist in Abbildung 2.4 dargestellt. Dieses wurde im positiven Modus mit 10 eV Kollisionsenergie auf einem Bruker

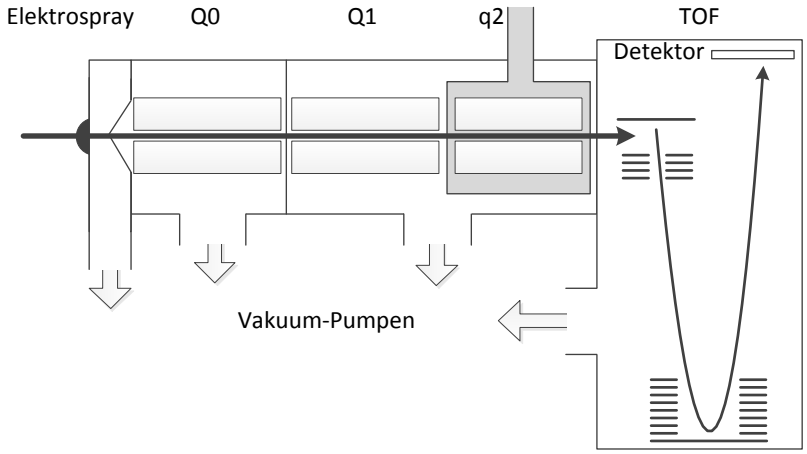


Abbildung 2.3. Vereinfachtes Schema eines QqTOF Massenspektrometers (nach [CLT01]): Das Elektrospray-Interface dient zur Probenionisierung und im Q0 werden die Ionen fokussiert. Im Quadrupol Q1 findet die Massenselektion statt. Die Kollisionszelle q2 fragmentiert die Ionen mit Hilfe eines Stoßgases (CID). Der Flugzeitmassenanalysator (TOF) dient zur Trennung der Ionen nach ihrem Masse-zu-Ladung Verhältnis (m/z) und das resultierende Spektrum wird am Detektor aufgenommen.

micrOTOF II aufgenommen. Das Spektrum von Epicatechin (CID: 72276) zeigt neben den annotierten Peaks auch das Vorläuferion mit 291,0758 m/z. Im Gegensatz dazu ist im PubChem Eintrag dazu eine monoisotopische Masse von 290,079038 Da angegeben, welche der Masse des neutralen Moleküls entspricht. Diese Massendifferenz ist durch die zusätzliche Ladung des Moleküls zu erklären.

2.1.3 Massengenauigkeit

Die Massengenauigkeit eines Massenspektrometers spielt eine entscheidende Rolle bei der späteren Analyse der Spektren. Diese gibt die Abweichung der berechneten exakten Masse zur gemessenen Masse an und wird üblicherweise in ppm („parts per million“) angegeben. Das Vorläuferion in Abbildung 2.4 weist eine gemessene Masse von 291,0758 Da auf. Die berechnete exakte Masse von $C_{15}H_{14}O_6$ beträgt 290,079 Da (geladen: 290,079 - 5,486E-4 + 1,008 = 291,0865 Da). Dies entspricht einer Abweichung von -0,011 Da ($\frac{291{,}0758\ \text{Da} - 291{,}086\ \text{Da}}{291{,}086\ \text{Da}} \cdot 1000000 \approx -35$ ppm). GC/MS Geräte messen oft nur Nominalmassen, was die spätere Identifizierung schwerer macht, weil

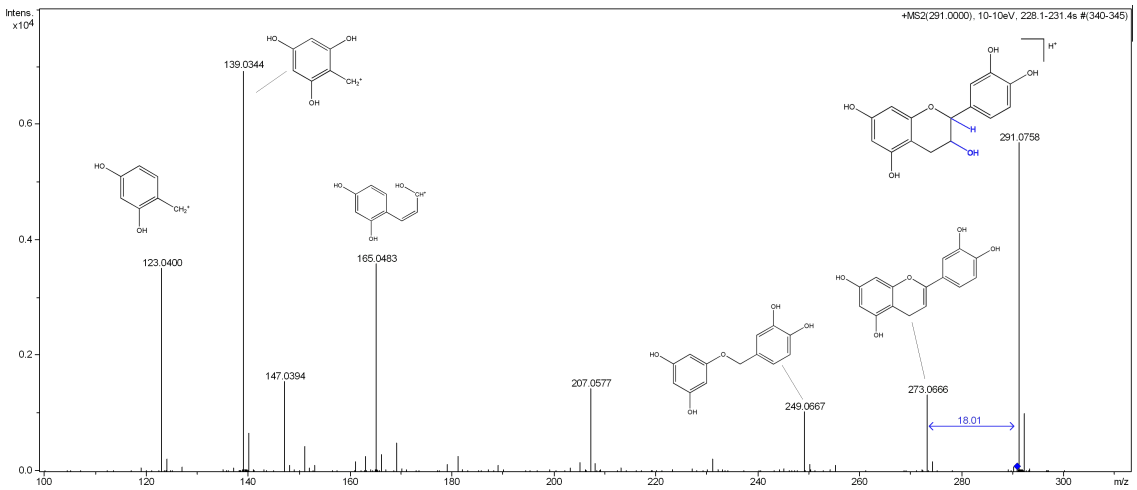


Abbildung 2.4. Handannotiertes MS/MS Spektrum von Epicatechin. Das Vorläuferion mit 291,0758 m/z und dessen Neutralverlust von Wasser (Massendifferenz von 18,01 Da), sowie weitere Fragmentionen sind dargestellt. (nach [WSMHN10])

ein Peak mehrere Erklärungen haben kann. Beispielsweise hat C_6H_6 eine exakte Masse von 78,047 Da und wäre dadurch nicht unterscheidbar von C_3H_7FO mit einer exakten Masse von 78,048 Da.

2.2 Identifizierung von Metaboliten

Aus den aufgenommenen LC/MS Spektren kann die Retentionszeit, die Masse und im besten Fall die Summenformel einer Verbindung bestimmt (Abbildung 1.1) werden. Mit diesen Daten kann in Metabolomikdatenbanken, die auf bestimmte biologische Kontexte zugeschnitten sind, gesucht werden. HMDB [WTK+07] und Metlin [SMW+05] sind Beispiele solcher Datenbanken und enthalten vor allem Spektren von humanen Metaboliten. Die Anzahl der Einträge ist allerdings sehr gering (siehe Kapitel 2.2.2). Daher müssen weitere Schritte durchgeführt werden, um eine Verbindung eindeutig zu bestimmen. Durch eine MS/MS Messung werden Strukturhinweise gewonnen, die zur Identifizierung des Analyten herangezogen werden können.

Aus einem MS/MS Spektrum kann man auf die Struktur eines Moleküls schließen, wenn man eine Idee hat, um welche Substanzklasse es sich handelt bzw. bereits ein Referenzspektrum gemessen wurde. Um mögliche Kandidaten zu finden, kann

auf verschiedene Datenbanken zurückgegriffen werden. Zum einen gibt es Spektrendatenbanken, die durch einen Spektrenvergleich mögliche Kandidaten finden können. In Strukturdatenbanken kann man eine Verbindung unter anderem nach der Summenformel oder der exakten Masse heraussuchen. Außerdem gibt es noch die Möglichkeit der Strukturgenerierung, die völlig auf Datenbanken verzichtet und alle möglichen Strukturen zu einer Summenformel erzeugen kann. Im folgenden Abschnitt werden drei Möglichkeiten zur Kandidatensuche näher betrachtet.

2.2.1 Strukturdatenbanken

Strukturdatenbanken beinhalten Strukturformeln, sowie verschiedene chemische Eigenschaften von bekannten Molekülen. Drei bekannte Vertreter sind ChemSpider [PW10], PubChem [WXS+09, BWT+08] und KEGG („Kyoto Encyclopedia of Genes and Genomes“) [OGS+99, KG00, KGKN02]. Tabelle 2.2 zeigt die Anzahl der Einträge, die von 16 262 (KEGG) bis 28,3 Millionen (PubChem) Strukturen reicht, und wie auf die entsprechenden Datenbanken zugegriffen werden kann. KEGG ist eine auf Reaktionsnetzwerke spezialisierte Datenbank, die Genominformationen („KEGG GENES“) mit chemischen Strukturen und Reaktionen („KEGG LIGAND“) verknüpfen und schließlich in Stoffwechselnetzwerken („KEGG PATHWAY“) darstellen kann. KEGG COMPOUND ist ein auf chemische Strukturen spezialisierter Teil von KEGG LIGAND und beinhaltet nur stoffwechselrelevante Verbindungen, deren biologische Funktion geklärt ist.

Die Strukturdatenbank PubChem besteht aus drei Teilen: „Substance“, „Compound“ und „BioAssay“ und beinhaltet viele Millionen Strukturen. „Substance“ enthält Strukturen von externen Quellen, die einer SID zugeordnet werden. „Compound“ ordnet diesen SIDs eindeutigen CIDs zu, um Redundanzen aus verschiedenen Quellen zu vermeiden. Auf diese Weise sind viele Datenbanken, zum Beispiel KEGG, in PubChem eingefügt worden. Des Weiteren enthält die „BioAssay“ Datenbank Informationen über die Bioaktivität von Verbindungen.

Genau wie PubChem ist auch ChemSpider eine Datenbank für chemische Verbindungen, welche zu 400 verschiedenen Datenquellen verlinkt [PW10] sind. Dabei wird

vor allem auf die Mithilfe der Nutzer zur Verbesserung und Ergänzung der Daten gesetzt.

Datenbank	Verbindungen	Zugriff	Lizenz
KEGG	16 262[2]	Download, Webservice	Freie akademische Lizenz
PubChem	28 389 738[3]	Download, Webservice	Freie Lizenz
ChemSpider	≈ 26 000 000[4]	Weboberfläche, Webservice	Open Source Lizenz

Tabelle 2.2. Aktuelle Statistiken zu den Strukturdatenbanken auf dem Stand von Oktober 2011.

Als Beispiel für einen Eintrag in einer Strukturdatenbank zeigt Abbildung 2.5 den PubChem Eintrag von Naringenin (CID: 932) mit Summenformel und einer Auswahl an chemischen Eigenschaften.

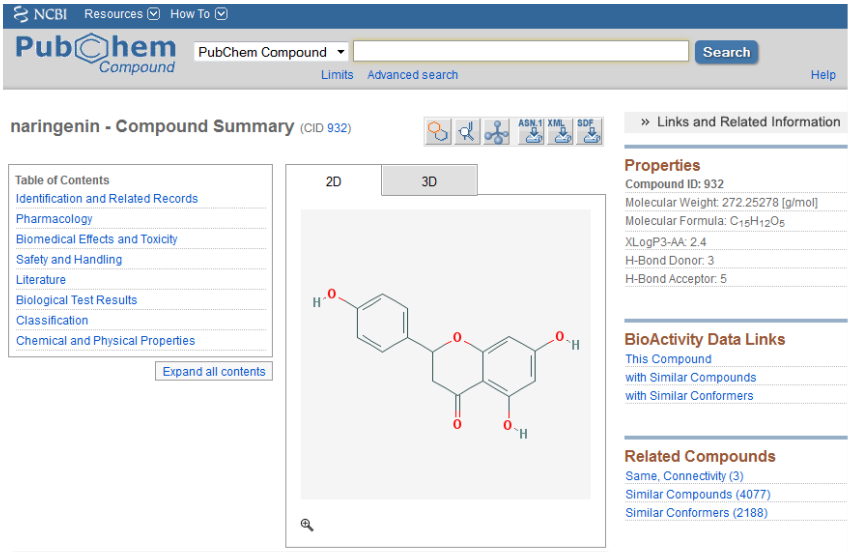


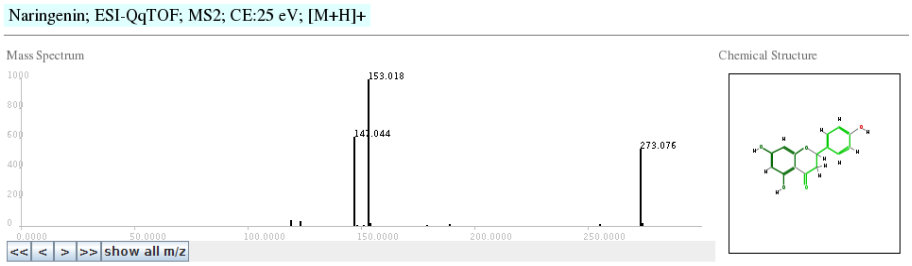
Abbildung 2.5. PubChem Eintrag von Naringenin mit CID 932. Auf dem Screenshot ist die Molekülstruktur und eine Auswahl an chemischen Eigenschaften dargestellt.

Auf alle genannten Datenbanken kann per Webservice zugegriffen werden, um entsprechende Strukturen zu suchen. Doch nur KEGG (kostenlos für akademische Benutzer) und PubChem erlauben die kompletten Daten herunterzuladen und somit einen lokalen Spiegel anzulegen.

[2] `http://www.ncbi.nlm.nih.gov/sites/entrez?term=all%5Bfilt%5D&cmd=search&db=pccompound` - Abgerufen im Oktober 2011
[3] `http://www.genome.jp/dbget-bin/www_bfind?compound` - Abgerufen im Oktober 2011
[4] `http://www.chemspider.com/About.aspx` - Abgerufen im Oktober 2011

2.2.2 Spektrendatenbanken

Spektrendatenbanken enthalten neben der chemischen Struktur unter anderem auch Daten über das verwendete Massenspektrometer, die Kollisionsenergie sowie die eigentlichen Spektren. Daher kann man Messungen einer unbekannten Verbindung mit den bereits aufgenommenen Spektren vergleichen, um dadurch Rückschlüsse auf dessen Struktur zu ziehen. Abbildung 2.6 zeigt als Beispiel ein Spektrum von Naringenin (CID: 932) aus der MassBank [HAK+10] mit der ID PB000123[5].




```
ACCESSION: PB000123
RECORD_TITLE: Naringenin; ESI-QqTOF; MS2; CE:25 eV; [M+H]+
DATE: 2011.12.13
AUTHORS: Boettcher C, Institute of Plant Biochemistry, Halle, Germany
LICENSE: CC BY-SA
COPYRIGHT: Copyright(C) 2008 Institute of Plant Biochemistry, Halle, Germany
COMMENT: IPB_RECORD: 83
COMMENT: CONFIDENCE: confident structure

CH$NAME: Naringenin
CH$NAME: 5,7-dihydroxy-2-(4-hydroxyphenyl)chroman-4-one
CH$COMPOUND_CLASS: Natural Product; Flavanone
CH$FORMULA: C15H12O5
CH$EXACT_MASS: 272.06847
CH$SMILES: C1C(OC2=CC(=CC(=C2C1=O)O)O)C3=CC=C(C=C3)O
CH$IUPAC: InChI=1S/C15H12O5/c16-9-3-1-8(2-4-9)13-7-12(19)15-11(18)5-10(17)6-14(15)20-13/h1-6,13,16-18H,7H2
CH$LINK: PUBCHEM CID:932
CH$LINK: INCHIKEY FTVWIRXFELQLPI-UHFFFAOYSA-N
CH$LINK: KEGG C00509

AC$INSTRUMENT: API QSTAR Pulsar i
AC$INSTRUMENT_TYPE: ESI-QqTOF
AC$ANALYTICAL_CONDITION: MS_TYPE MS2
AC$ANALYTICAL_CONDITION: IONIZATION ESI
AC$ANALYTICAL_CONDITION: MODE POSITIVE
AC$ANALYTICAL_CONDITION: PRECURSOR_TYPE [M+H]+
AC$ANALYTICAL_CONDITION: PRECURSOR_SELECTION Q
AC$ANALYTICAL_CONDITION: COLLISION_ENERGY 25 eV

PK$NUM_PEAK: 12
PK$PEAK: m/z int. rel.int.
  119.051 467.616 45
  123.044 370.662 36
  147.044 6078.145 606
  ...
  273.076 5286.093 527
  274.081 246.689 23
//
```


Abbildung 2.6. MassBank Eintrag mit der ID PB100123 von Naringenin. Unter anderen ist das verwendete Massenspektrometer, messungsrelevante Einstellungen sowie die Peakliste abgebildet.

EI-MS hat den Vorteil, vergleichbare Spektren aufzunehmen, da hier meist eine Kollisionsenergie von 70 eV eingesetzt wird. Daher ist es möglich eine Spektrendaten-

[5] `http://www.massbank.jp/jsp/FwdRecord.jsp?id=PB000123` - Abgerufen im November 2011

bank, zum Beispiel NIST '11[6] mit 243 893 EI Spektren, aufzubauen. Durch Vergleich von einem gemessenen Spektrum mit einem bereits in einer Datenbank vorhandenen kann einen starken Hinweis auf die gemessene Verbindung liefern. ESI-MS Spektren sind in der Regel hochaufgelöst und mit unterschiedlichen Kollisionsenergien aufgenommen. Unterschiedliche Instrumente und Kollisionsenergien erzeugen allerdings verschiedene Spektren. Abbildung 2.7 zeigt einen Vergleich von einem API QSTAR Pulsar i (ESI-QqTOF, grün, PB000123, 25 eV) und einem Q-Tof Premier von Waters (ESI-QTOF, rot, PR040043, 30 eV). Die beiden MS/MS Spektren von Naringenin unterscheiden sich sowohl in der Intensität als auch in der Anzahl der Peaks.

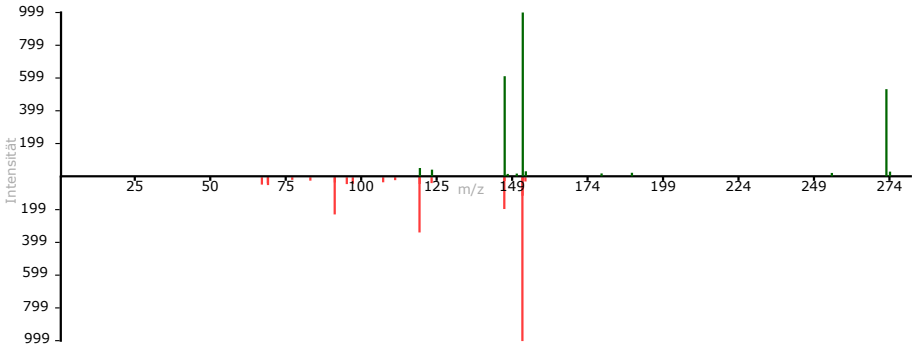


Abbildung 2.7. Vergleich zweier MS/MS Spektren von Naringenin mit den MassBank IDs PB000123 (grün) und PR040043 (rot). Ersteres wurde mit einer Kollisionsenergie von 25 eV auf einem API QSTAR Pulsar i (ESI-QqTOF) am IPB in Halle gemessen. Das rote MS/MS Spektrum wurde vom RIKEN Plant Science Center auf einem UPLC Q-Tof Premier von Waters (ESI-QTOF) mit einer Kollisionsenergie von 30 eV gemessen.

Dies macht den Aufbau einer umfassenden Spektrendatenbank schwierig. Tabelle 2.3 gibt einen Überblick über derzeitige (ESI) Spektrendatenbanken und zeigt die Anzahl der gemessenen Verbindungen, wobei pro Verbindung mehrere Messungen möglich sind. NIST '11 besitzt eine kommerzielle Lizenz, wobei METLIN, HMDB und MassBank auch kostenlos durchsucht werden können.

Insgesamt umfassen diese Datenbanken 14 273 verschiedene Verbindungen, wobei nicht ausgeschlossen werden kann, dass Redundanzen vorhanden sind. Selbst im besten Fall decken diese Datenbanken nur einen Bruchteil von Verbindungen ab, die in einer Strukturdatenbank vorhanden sind.

[6] `http://www.sisweb.com/software/ms/nist.htm` - Abgerufen im Oktober 2011

Datenbank	Verbindungen	Spektren	Genauigkeit	Kommentar
NIST '11	8 505	95 409	Exakt/Nominal	Kommerzielle Lizenz
METLIN	5 327	29 500	Exakt	Weboberfläche
HMDB	921	2 565	Nominal	Weboberfläche
MassBank	2 189	9 218	Exakt/Nominal	Weboberfläche, Webservice

Tabelle 2.3. Übersicht über die Spektrendatenbanken mit der Anzahl der enthaltenen ESI-MS/MS Spektren und Verbindungen [NB10].

2.2.3 Strukturgenerierung

Strukturgeneratoren können datenbankunabhängig Strukturen generieren und dadurch bisher unbekannte Verbindungen gefunden werden. MOLGEN 4.0 [KLGM98] ist ein Strukturgenerator, der ohne Redundanzen alle Strukturen einer bestimmten Summenformel aufzählen kann. Ohne Einschränkungen zu der gemachten Formel kann die Zahl der generierten Strukturen sehr groß werden. Zum Beispiel generiert MOLGEN 4.0 zu der Summenformel $C_6H_8O_6$ 2 558 517 Isomere [KLGM98]. Eine Reduzierung dieser kann durch das Erzwingen chemischer Merkmale in den generierten Strukturen erreicht werden.

2.3 Systeme und Algorithmen zur Spektreninterpretation

Das folgende Kapitel beschäftigt sich mit den bereits verfügbaren Methoden zur Fragmentvorhersage von EI-MS und MS/MS Spektren. Kombinatorische Algorithmen haben in der Vergangenheit zu viel Rechenzeit benötigt. Deshalb ist vor allem Software entwickelt worden, dass sich auf Expertenwissen stützt. Aus den Veröffentlichungen sind Regeln zur Fragmentierung erstellt worden, die in einem Regelwerk abgespeichert sind. Im folgenden Abschnitt wird Software zur regelbasierten Fragmentvorhersage vorgestellt.

2.3.1 Regelbasierte Fragmentvorhersage

Regelbasierte Algorithmen versuchen Verbindungen so zu fragmentieren, wie es in einem Regelwerk abgespeichert wurde. Außerdem sind unterschiedliche Regelmengen für bestimmte Ionisierungstechniken, zum Beispiel ESI oder EI, vorhanden. Das hat den Vorteil bereits bekannte Bindungsbrüche zuverlässig vorhersagen zu können, liefert aber bei unbekannten Molekülen oder neuen Ionisierungstechniken höchstwahrscheinlich unbefriedigende Ergebnisse.

MassFrontier[7], ACD/MS Fragmenter[8] und MOLGEN-MS [KLMV01] sind Programme, die nach diesem Prinzip arbeiten. Ersteres hat eine aus der Literatur gesammelte Regelmenge gespeichert und kann außerdem um eine durch den Benutzer angelegte Regeldatenbank erweitert werden. Dies verlängert die Laufzeit so sehr, dass nur wenige Kandidaten prozessiert werden können [SMB09, HKF$^+$08]. Abbildung 2.8 zeigt ein Beispiel einer solchen Regel. Das Experiment von [HKF$^+$08] hat gezeigt, dass MassFrontier (getestete Version 4.0) in der Lage ist für im positiven Modus gemessene MS/MS Spektren gute Ergebnisse zu erreichen. Die Autoren haben 102 Verbindungen auf einem Micromass Q-TOF II (MS/MS) mit jeweils fünf verschiedenen Kollisionsenergien (positiver Modus) gemessen und die Kandidaten sind nach der Anzahl der erklärten Peaks geordnet worden. Abschnitt 4.6.2 zeigt diese Ergebnisse und vergleicht diese mit der im Rahmen dieser Arbeit entwickelten Software. Auf der anderen Seite hat [HRM$^+$08] gezeigt, dass MassFrontier 5.0 keine Peaks von MS/MS Daten, die im negativen Modus gemessen worden sind, annotieren konnte. Die Ursache könnte an dem Fehlen von Regeln zur Fragmentvorhersage im negativen Modus begründet sein.

ACD/MS Fragmenter ist ein Teil des ACD/MS Managers, der ähnlich MassFrontier Regeln zur Fragmentierung aus der Literatur enthält. Diese werden genutzt, um Fragmente von Kandidatenstrukturen vorherzusagen. Außerdem erlauben es beide Programme, die Anzahl der Fragmentierungschritte festzulegen, was einen großen Einfluss auf das Ergebnis und die Laufzeit hat [SMB09]. Die genauen Details der Algorithmen, sowie der Quellcode der kommerziellen Programme MassFrontier und ACD/MS Fragmenter sind nicht verfügbar.

[7] `http://www.thermoscientific.com/massfrontier/` - Abgerufen im November 2011, aktuelle MassFrontier Version: 7.0

[8] `http://acdlabs.com/products/adh/ms/ms_frag/` - Abgerufen im November 2011

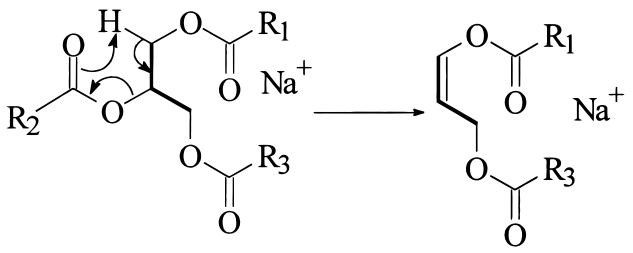


Abbildung 2.8. Beispiel einer Fragmentierungsregel aus MassFrontier. Abspaltung von zwei Acylgruppen. (nach [CGP98])

MOLGEN-MS [KLMV01] hat sich auf die Interpretation von EI-MS Spektra (Nominalmassen) spezialisiert. Das besondere ist, dass die Kandidatenstrukturen zu einer bestimmten Summenformel unabhängig von Molekül datenbanken mit MOLGEN generiert werden können. Anschließend werden diese fragmentiert und eine Rangordnung berechnet. Auch dieses Programm wird im Abschnitt 4.7 mit MetFrag verglichen.

Ein Vergleich zwischen MassFrontier, ACD/MS Fragmenter und MOLGEN-MS wurde von [SMB09] mit EI-MS (Nominalmassen) Spektren angefertigt. Dabei wurden die Kandidaten mit MOLGEN 3.5 generiert, mit allen drei Programmen prozessiert und schließlich mit den experimentellen Spektren verglichen. Zur Auswertung ist die „relative ranking position“ RRP (siehe Kapitel 4.2) berechnet worden. Es konnte gezeigt werden, dass MOLGEN-MS den besten RRP hat, wenn man maximal 500 Kandidaten betrachtet. MassFrontier mit drei Fragmentierungsschritten ist ähnlich gut und sogar besser als MOLGEN-MS je mehr Kandidaten prozessiert werden (bis 10000 Strukturen). Auffällig ist, dass ACD/MS Fragmenter stets den größten RRP liefert und damit am schlechtesten abschneidet. Somit ist diese Software für EI-MS nur bedingt geeignet, da auch die Laufzeit bei wenigen Kandidaten sehr lang ist [SMB09].

2.3.2 Kombinatorische Fragmentvorhersage

Kombinatorische Algorithmen brauchen viel Rechenleistung und wurden erst mit schneller werdenden Rechnern durchführbar, da sie jede mögliche Substruktur eines

Kandidaten aufzählen. Die Anzahl an resultierenden Fragmenten wächst sehr stark, wenn keine geeigneten Einschränkungen vorgenommen werden. Außerdem muss die Isomorphie von bereits erstellten Substrukturen beachtet werden, da sonst viele Fragmente doppelt generiert werden.

In [Swe03] wurde festgestellt, dass sich viele Peaks durch einfache Fragmentierung des Molekülgraphen erklären lassen. Die Autoren haben Annahmen zur Vereinfachung des Problems gemacht: Neutralverluste und Umlagerungen, die innerhalb der gemessenen Verbindungen auftreten können, werden nicht beachtet. Außerdem ist angenommen worden, dass die einfachste Lösung einen Peak mit einem Fragment zu erklären die Beste sei. Alle Vereinfachungen treten jedoch mehr oder weniger häufig in realen Experimenten auf, vor allem bei ESI-MS Messungen werden Neutralverluste sehr häufig beobachtet (siehe Abbidung 2.4).

EPIC [HM05] ist eine nicht öffentlich verfügbare Software, die mit Hilfe eines Bindungsbrechungsalgorithmus versucht Peaks zu annotieren. Es ist für eine Kandidatenstruktur jedes mögliche Fragment generiert und mit den Peaks aus dem gemessenen Spektrum verglichen worden. Außerdem können Fragmente zu Peaks zugeordnet werden, die sich nur durch das Gewicht von wenigen Wasserstoffen unterscheiden. Wenn ein Peak von mehreren Fragmentstrukturen erklärt wird, dann ist eine Scoring Funktion verwendet worden: Diese bestraft Wasserstoffadditionen (oder Subtraktionen) und schwer zu brechende Bindungen. Zur Auswertung ist die Anzahl der mit Strukturen annotierten Peaks von zwei gemessenen Verbindungen verwendet worden. EPIC hat mehr Peaks erklärt als MassFrontier 4 und ACD MS Manager Version 8.13, wobei keine weitere Evaluierung durchgeführt worden ist.

„Fragment Identificator“ (FiD) von [HRM$^+$06, HRM$^+$08] ist ein Prgramm zur Identifizierung von Fragmentionen aus MS/MS Spektren. Die Software versucht Fragmente der Kandidatenstruktur den gemessenen Peaks zuzuordnen. [HRM$^+$06] verwendet Bindungsdissoziationsenergien (BDE) als Kantengewichte im Molekülgraphen, um einfach bzw. schwer brechende Bindungen zu annotieren. FiD kann alle Fragmente direkt aus dem Ausgangsmolekül erzeugen (Einzelschrittverfahren) oder bereits generierte Fragmente weiter fragmentieren (Mehrschrittverfahren). Es sind durch einen Experten annotierte MS/MS Spektren von 27 Verbindungen untersucht worden. FiD ist in der Lage 90% der Fragmente im Einzelschrittverfahren richtig zuzuordnen. Das Mehrschrittverfahren liefert schlechtere Ergebnisse. Es ist auch festgestellt worden,

CDK
Java Bibliothek
Lesen und schreiben vieler Datenformate für Molekülstrukturen
Größter gemeinsamer Subgraph (MCS)
Integration in Statistiksoftware R
OpenBabel
C++ Bibliothek, die in den meisten Linux Distributionen enthalten ist
Spezialisiert auf Umwandlung zwischen Molekül datenformaten
2D und 3D Struktur Layout
Strukturoptimierung mit Kraftfeldern
Open Babel Version 2.3 kann 110 Formate lesen und schreiben.

Tabelle 2.4. Ausgewählte Funktionen vom Chemistry Development Kit (CDK) und OpenBabel. Ersteres ist eine Java Bibliothek, die viele Funktionen im Umgang mit Molekülen bereitstellt. OpenBabel ist spezialisiert auf das Umwandeln in verschiedene Molekül datenformate.

dass die Laufzeit extrem ansteigt je größer die Kandidatenstruktur ist. Deshalb ist eine Obergrenze von 50 Atomen (ohne Wasserstoffe) festgelegt worden, die maximal mit FiD berechenbar ist. Eine zweite Untersuchung [HRM+08] hat gezeigt, dass FiD mehr Peaks als MassFrontier 5.0 erklären kann, wobei mit sinkender Genauigkeit des Massenspektrometers beide Programme vergleichbare Ergebnisse erreichen.

2.4 Cheminformatik Software

Um die Programmierung von MetFrag zu beschleunigen, wurde bei der Entwicklung von MetFrag auf bereits verfügbare Software Bibliotheken gesetzt, die verschiedene chemische Datenformate, zum Beipiel MDL Moldateien, SMILES oder SD Dateien, lesen, schreiben und visualieren können. Da nur quelloffene Software für MetFrag verwendet werden soll, kommen keine kommerziellen Pakete wie OEChem[9] oder Daylight[10] in Frage. Tabelle 2.4 gibt einen Überblick über die aktiv weiterentwickelten Cheminformatik Open Source Pakete „OpenBabel“[11] [OBJ+11] und das „Chemistry Development Kit“[12] [SHK+03, SHK+06].

[9] `http://www.eyesopen.com/oechem-tk/`
[10] `http://www.daylight.com/`
[11] `http://openbabel.org/`
[12] `http://cdk.sf.net/`

Im folgenden Abschnitt werden einige Funktionen des CDK, das als Cheminformatik Bibliothek für MetFrag verwendet wurde, näher betrachtet.

2.4.1 Graphentheorie in der Cheminformatik

Chemische Moleküle werden üblicherweise als molekulare Graphen gespeichert. Ein Molekül ist definiert als ungerichteter, verbundener, gewichteter und beschrifteter Graph $G = (V, E, t_V, t_E, w_V, w_E)$ [HRM+06]. Die Atome werden durch die Knotenmenge V („vertices“) beschrieben und die Kantenmenge E („edges“) entspricht den Bindungen eines Moleküls. Die Funktion $t_V : V \to A$ ordnet jedem Atom den Typ (Kohlenstoff, Wasserstoff, Stickstoff,...) zu und $t_E : E \to B$ bestimmt den Bindungstyp (Einfachbindung, aromatische Bindung,...). Knoten besitzen Atomgewichte, die mit der Funktion $w_V : V \to \mathbb{R}_+$ zugeordnet werden und die Bindungsgewichte, zum Angeben der Stärke einer Bindung, werden durch die Funktion $w_E : E \to \mathbb{R}$ angegeben. Im folgenden Abschnitt wird anhand einer CML Datei gezeigt, wie die Informationen abgespeichert werden können.

CML

CML („chemical markup language“) bietet die Möglichkeit Moleküle computerlesbar abzuspeichern. Dieses Format stützt sich auf XML („extensible markup language“) und bringt alle Vorteile (Lesbarkeit, Erweiterbarkeit, verfügbare Software) und Nachteile (Dateigröße) von XML mit sich [MRR99]. Es wurden verschiedene Erweiterungen auf CML aufbauend von [MRR01, GMRRW01, MRR03, MRRWW04, HMRR06, KHL+07] publiziert, was zeigt wie flexibel und erweiterbar dieses Format ist. In Abbildung 2.9 ist der Inhalt der CML Datei von Ethanol (CID: 702) dargestellt. Der Abschnitt `atomArray` enthält neben den Atomsymbolen auch Atomkoordinaten und um welches Isotop es sich handelt. Die Bindungen werden innerhalb des Tags `bondArray` gespeichert und können mit der Bindungsordnung (`bondOrder`) (siehe Kapitel 3.1.2), ein heuristisches Maß zur Bestimmung der Bindungsstärke, annotiert sein.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<molecule id="m1" title="702" xmlns="http://www.xml-cml.org/schema">
  <atomArray>
    <atom id="0" elementType="O" x2="2.5369" y2="-0.25" formalCharge="0" isotopeNumber="16"/>
    <atom id="1" elementType="C" x2="3.403" y2="0.25" formalCharge="0" isotopeNumber="12"/>
    <atom id="2" elementType="C" x2="4.269" y2="-0.25" formalCharge="0" isotopeNumber="12"/>
    <atom id="3" elementType="H" x2="3.8015" y2="0.7249" formalCharge="0" isotopeNumber="1"/>
    <atom id="4" elementType="H" x2="3.0044" y2="0.7249" formalCharge="0" isotopeNumber="1"/>
    <atom id="5" elementType="H" x2="3.959" y2="-0.7869" formalCharge="0" isotopeNumber="1"/>
    <atom id="6" elementType="H" x2="4.8059" y2="-0.56" formalCharge="0" isotopeNumber="1"/>
    <atom id="7" elementType="H" x2="4.579" y2="0.2869" formalCharge="0" isotopeNumber="1"/>
    <atom id="8" elementType="H" x2="2.0" y2="0.06" formalCharge="0" isotopeNumber="1"/>
    </atom>
  </atomArray>
  <bondArray>
    <bond id="0" atomRefs2="0 1" order="S">
      <scalar title ="bondOrder" dataType="xsd:string">0.593136</scalar>
    </bond>
    <bond id="1" atomRefs2="0 8" order="S"></bond>
    <bond id="2" atomRefs2="1 2" order="S">
      <scalar title ="bondOrder" dataType="xsd:string">1.029438</scalar>
    </bond>
    <bond id="3" atomRefs2="1 3" order="S"></bond>
    <bond id="4" atomRefs2="1 4" order="S"></bond>
    <bond id="5" atomRefs2="2 5" order="S"></bond>
    <bond id="6" atomRefs2="2 6" order="S"></bond>
    <bond id="7" atomRefs2="2 7" order="S"></bond>
  </bondArray>
</molecule>
```

Abbildung 2.9. CML Datei von Ethanol mit der Bindungsordnung annotierten Bindungen.

SD Datei

Eine weitere sehr weit verbreitete Art Moleküle auszutauschen sind SD Dateien, die standardmäßig von sehr vielen Programmen unterstützt werden. Diese bestehen aus einer oder mehreren Moldateien, die Strukturinformationen enthalten, sowie dazugehöriger Eigenschaften, welche pro Verbindung angegeben werden können. Die Spezifikation[13] wurde ursprünglich 1992 von [DNH+92] entwickelt. Ein Beispiel einer SD Datei von Ethanol ist in Beispielcode A.1 dargestellt.

[13] `http://download.accelrys.com/freeware/ctfile-formats/ctfile-formats.zip` von 2010 - Abgerufen im Oktober 2011

Ringsuche in Molekülgraphen und Aromatendetektion

Das Auffinden von Ringen spielt eine große Rolle in der Cheminformatik, da es dadurch ermöglicht wird, Strukturen zu klassifizieren, zu benennen, graphisch darzustellen und Aromaten zu detektieren. Das CDK [SHK+03] verwendet den Algorithmus von [HJK96] zur schnellen Suche von Ringen in Molekülgraphen. Dieser Algorithmus basiert auf der Kontraktion des Pfadgraphen, der am Anfang das Abbild des Molekülgraphen (Abbildung 2.10) darstellt.

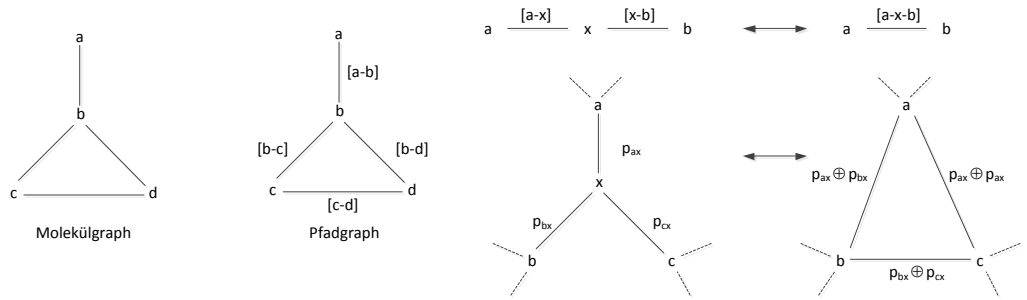


Abbildung 2.10. Zusammenführen von Kanten und deren Beschriftung nach dem Algorithmus von [HJK96]. Links ist der Molekülgraph dargestellt, aus dem der Pfadgraph abgeleitet werden kann. Die Zusammenfassung der Kanten erfolgt nach dem rechten Schema. Falls ein Weg $a - x - b$ im Molekülgraph existiert, dann kann im Pfadgraph diese Kante durch eine mit der Beschriftung [a-x-b] repräsentiert werden ($\oplus$ ist der Operator für die Konkatenation zweier Zeichenketten). Der Knoten x und dessen Kanten können demnach aus dem Pfadgraph ohne Verlust von Informationen entfernt werden, da eine neue Kante zwischen $a - b$ hinzugefügt wurde. (Abbildung nach [HJK96])

Durch die Reduktion werden entsprechende Knoten des Pfadgraphen entfernt und in der Bezeichnung der neuen Kante gespeichert (Abbildung 2.10). Pfade, d.h. alle Knoten v_i auf dem Weg W sind unterschiedlich ($\forall\, i, j \in \{1, \ldots, n\}$ gilt $v_i \neq v_j$ falls $i \neq j$), werden aus dem Pfadgraphen entfernt, damit nur Kreise übrig bleiben. Das Label eines Zyklus entspricht dem gefundenen Kreis. Der Algorithmus terminiert sobald der Pfadgraph nicht weiter reduziert werden kann. Abbildung 2.11 zeigt dieses Vorgehen für ein einfaches Beispiel.

Durch die Ringdetektion kann im Algorithmus von MetFrag unterschieden werden, ob eine linearen oder zyklische Bindung vorliegt. Dies spielt bei der Fragmentierung von Molekülen eine Rolle (siehe Abschnitt 3.1.3).

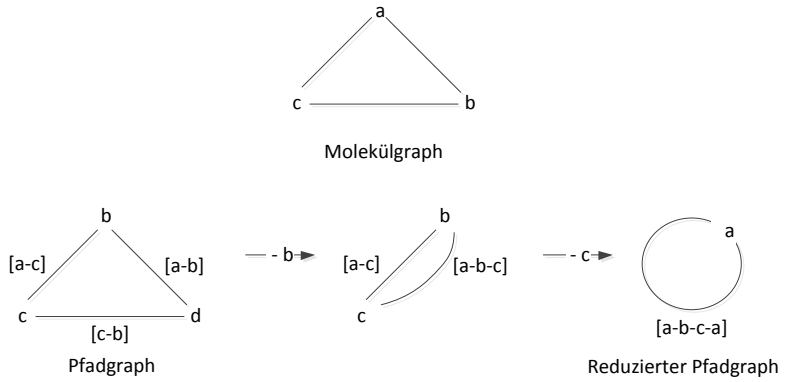


Abbildung 2.11. Beispiel zur Ringsuche in einem Beispielgraphen mit dem Algorithmus von [HJK96]. Nach der Reduzierung des Pfadgraphen entspricht das Label [a-b-c-a] des Zyklus dem gefundenen Kreis (Abbildung nach [HJK96]).

Aromatendetektion

Durch die Ringdetektion aus dem vorherigen Abschnitt, kann nun die Annotation von Aromaten durchgeführt werden. Beginnend mit dem größten Ring werden die Elektronen der alternerienden Doppel- und Dreifachbindungen und die freien Elektronenpaare von Heteroatomen gezählt. Danach wird überprüft, ob der Ring entsprechend der Hückel-Regel $4n+2$ π-Elektronen besitzt. Benzol ist demnach aromatisch, da es 6 ($n = 1$) π-Elektronen besitzt. Entsprechend dieser Regel werden alle Bindungen des Ringes als aromatisch markiert [SHK+03].

2.4.2 Molekülrepräsentation

Serialisierung von Molekülgraphen mit SMILES™ und InChI™

Die SMILES™ [Wei88] („Simplified Molecular Input Line Entry Specification“) Notation bietet eine einfache Möglichkeit, um Moleküle in einer Zeichenkette zu re-

präsentieren. Dabei wird die Konnektivität des Molekülgraphen beschrieben, wobei die Wasserstoffatome nicht explizit mit angegeben werden müssen. Atome werden durch ihre Symbole aus dem Periodensystem der Elemente angegeben. Nicht organische Substanzen werden durch eckige Klammern „[Au]“ beschrieben, wobei auch die verbundenen Wasserstoffe durch diese Notation angegeben werden können. Einfach- („-“), zweifach- („=“), dreifach- („#“) und aromatische- („:“) Bindungen werden durch die jeweiligen Symbole angegeben. Einfachbindungen werden implizit angenommen, wenn kein explizites Bindungssymbol angegeben wird, wobei aromatische Bindungen üblicherweise durch Groß- und Kleinschreibung der Atomsymbole angezeigt werden. Verzweigungen im Molekülgraph werden durch Klammern „()“ gekennzeichnet, wie in Abbildung 2.12 am Beispiel von Mesityloxid dargestellt ist. Ringe werden durch das Weglassen einer Bindung und der Nummerierung des ringöffnenden bzw. ringschließenden Atomes notiert. Abbildung 2.12 zeigt Cyclohexan, bei dem das Start- und Endatom mit einer 1 gekennzeichnet ist. Der resultierende SMILES™ dieser Verbindung ist `C1CCCCC1` nach [Wei88].

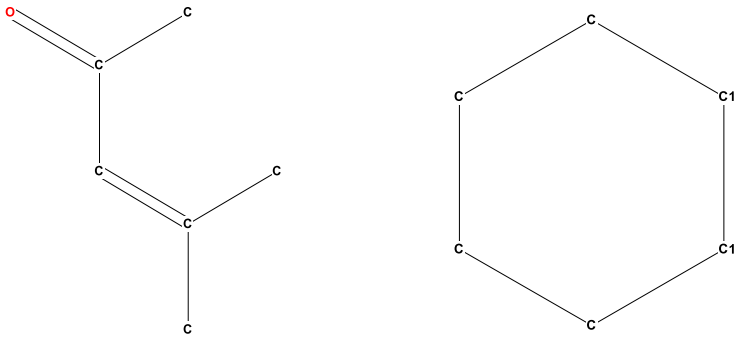
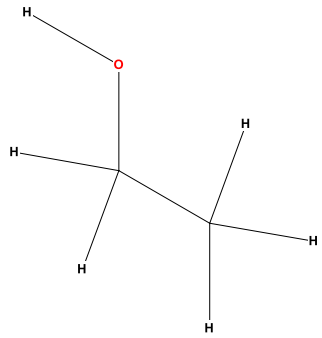


Abbildung 2.12. Der SMILES™ von Mesityloxid (CID: 885, links) ist `CC(=CC(=O)C)C`. Cyclohexan (CID: 8078, rechts) besitzt folgenden SMILES™: `C1CCCCC1`

Da ein Molekül viele unterschiedliche SMILES™ haben kann, wurden von [WWW89] die eindeutigen („canonicalized“) SMILES™ entwickelt, bei denen jedes Atom kanonisch geordnet und bezeichnet wird. Daher gibt es für eine Verbindung nur einen eindeutigen SMILES™, was in Abbildung 2.13 dargestellt ist.

[14] `http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=702` - Abgerufen im Oktober 2011



SMILES mit explizitem Wasserstoff:	[H]OC([H])([H])C([H])([H])[H]
Kanonischer SMILES:	CCO
InChI:	InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3
InChIKey:	LFQSCWFLJHTTHZ-UHFFFAOYSA-N

Abbildung 2.13. Ethanol (PubChem ID 702[14] dargestellt als Strukturformel mit expliziten Wasserstoffen, sowie der dazugehörige (kanonische) SMILES, InChI™ und InChIKey.

Ein InChI™ [SHT03] ist ein eindeutiger Identifizierungsstring der IUPAC[15] für chemische Verbindungen. Dieser wurde eingeführt, um Molekülinformationen zu speichern und diese im Internet und in Datenbanken leicht zugänglich zu machen. Der Quellcode zum Generieren eines InChI™ wurde unter einer Open Source Lizenz (LGPL) veröffentlicht. Ein InChI™ (siehe Beispielcode 2.1) besteht aus verschiedenen Info Blöcken, die mit einem „/“ getrennt sind. Jeder abgetrennte Teil steht für eine bestimmte Klasse von strukturellen Informationen über das Molekül, z.B. Summenformel, Konnektivität, Ladung und Stereochemie. Abbildung 2.13 zeigt den Standard (`S` in der Versionsschicht) InChI™ von Ethanol.

$$InChI = \underbrace{1S}_{\text{Version}} / \underbrace{C2H6O}_{\text{Summenformel}} / \underbrace{c1-2-3}_{\text{Konnektivität}} / \underbrace{h3H,2H2,1H3}_{\text{Verbundene H}} \tag{2.1}$$

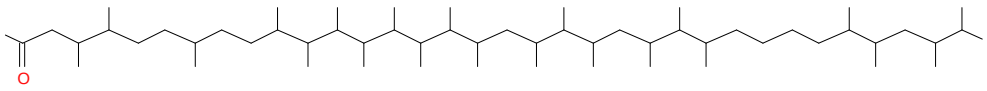
Standard InChI werden mit festgelegten Optionen, zum Beispiel für Stereochemie, generiert, damit diese untereinander vergleichbar sind.

Der InChIKey ist ein 25 Zeichen langer Hash (Beispielcode 2.2) eines InChI™, der vor allem zum Suchen im Internet genutzt wird.

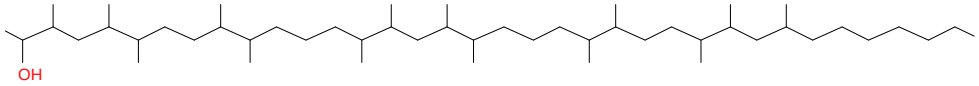
[15] International Union of Pure and Applied Chemistry - Institution für einheitliche Standards und Empfehlungen für verschiedene Bereiche der Chemie

$$\underbrace{LFQSCWFLJHTTHZ}_{\text{14-stellige Konnektivität}}-\underbrace{UHFFFAOYS}_{\text{Eigenschaften}}\ \underbrace{A}_{\text{InChI Version}}\ -\ \underbrace{N}_{\text{Prüfsumme}} \tag{2.2}$$

Ein InChIKey ist eindeutig für eine Verbindung, aber Kollisionen können mit geringer Wahrscheinlichkeit auftreten. Abbildung 2.14 zeigt eine Kollision zweier InChIKeys von Molekülen mit unterschiedlichen InChI™, aber gleichen InChIKey[16].



InChI: InChI=1S/C57H114O/c1-36(2)42(8)32-43(9)38(4)26-24-25-27-40(6)50(16)45(11)33-46(12)51(17)47(13)34-48(14)53(19)55(21)57(23)56(22)54(20)52(18)41(7)31-29-37(3)28-30-39(5)44(10)35-49(15)58/h36-48,50-57H,24-35H2,1-23H3
InChIKey: OCPAUTFLLNMYSX-UHFFFAOYSA-N



InChI: InChI=1S/C50H102O/c1-17-18-19-20-21-24-36(2)33-45(11)43(9)31-29-41(7)37(3)25-22-27-39(5)46(12)34-47(13)40(6)28-23-26-38(4)42(8)30-32-44(10)48(14)35-49(15)50(16)51/h36-51H,17-35H2,1-16H3
InChIKey: OCPAUTFLLNMYSX-UHFFFAOYSA-N

Abbildung 2.14. Gleicher InChIKey trotz unterschiedlicher InChI™ (siehe [16]).

Substruktursuche mit SMARTS

Eine weitere wichtige Funktion von Cheminformatik Software ist die Suche von Substrukturen, die durch SMARTS („SMiles ARbitrary Target Specification“ [JWD]) ermöglicht wird. SMARTS sind eine Erweiterung von SMILES™, wobei fast jeder SMILES™ auch ein gültiger SMARTS ist.

SMARTS ähneln regulären Ausdrücken zur Suche von Mustern in einer Zeichenkette. Statt einer Zeichenfolge wird ein Molekül durchsucht. Beispielsweise sucht folgender SMARTS `[c,n;H1]` [JWD] nach einem aromatischen Kohlenstoff `c` oder einem aromatischen Stickstoff `n` mit genau einem verbundenen Wasserstoff `H1`. SMARTS

[16]Kollision zweier InChIKeys von Antony John Williams: `http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry/` - Abgerufen im November 2011

können aus Symbolen zum Beschreiben von Atomen (z.B. C - aliphatischer Kohlenstoff, c - aromatischer Kohlenstoff, [#6] - beliebiges Kohlenstoffatom, [R] - beliebiges Atom in einem Ring) und Bindungen (z.B. [#6]-[#6] - Einfachbindung zwischen zwei Kohlenstoffatomen, [#6]~[#6] - beliebige Bindung zwischen zwei Kohlenstoffatomen) bestehen. Außerdem ist es möglich, logische Operatoren zu verwenden: `[!c;R]` sucht nach einem nicht-aromatischen Kohlenstoff in einem Ring. Rekursive SMARTS können benutzt werden, um bestimmte Atomeigenschaften zu beschreiben und zusammenzufassen: SMARTS `$([OH1][#6])` beschreibt (siehe Abbildung 3.7) einen Hydroxylrest `[OH1]`, der mit einem Kohlenstoffatom `[#6]` verbunden ist. Durch rekursive SMARTS `$()` ist es möglich, ein spezielles Atom mit einer bestimmten „Eigenschaft“ zu finden, ohne dies mit in das Ergebnis aufzunehmen. Umfangreichere Informationen, weitere Atom- und Bindungssymbole, sowie weiterführende SMARTS Beispiele sind unter [JWD] zu finden. MetFrag verwendet SMARTS Regeln zum Auffinden von Substrukturen häufiger Neutralverluste (siehe Tabelle 3.2).

2.4.3 Fingerprints und Strukturähnlichkeit

Fingerprints stellen eine abstrakte Repräsentation von speziellen Molekülmerkmalen dar [JWD]. Diese werden verwendet, um möglichst schnell ähnliche Verbindungen zu finden, oder dienen als Vorfilter für eine zeitaufwendige Isomorphieüberprüfung. Es existieren unterschiedliche Algorithmen zur Fingerprintgenerierung und diese liefern unterschiedlich gute Ergebnisse[17]. Die Funktionsweise eines Fingerprinter wird beispielhaft an dem in PubChem[18] verwendeten beschrieben: Grundlage ist ein Bitvektor der Länge 881. Jede Position ist binär kodiert und beschreibt, ob ein bestimmtes Merkmal vorhanden ist (1) oder nicht (0). Wenn beispielsweise das Molekül ≥ 16 H-Atome enthält, dann wird der Bit an Position 2 auf 1 gesetzt. Insgesamt werden 881 verschiedene Merkmale, die beispielsweise die Art und das Vorhandensein von Ringen oder die Nachbarschaft von Atomen beschreiben, analysiert. Um diese Merkmale ausfindig zu machen, werden unter anderem SMARTS (siehe Abschnitt 2.4.2) ver-

[17] `http://rguha.wordpress.com/2008/10/11/do-the-cdk-fingerprints-work/` - Abgerufen im November 2011

[18] `ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt` - Version 1.3, Abgerufen im Januar 2012

wendet. Die resultierenden Bitvektoren können genutzt werden, um eine chemische Ähnlichkeit zwischen Molekülen zu bestimmen.

Das Review von [WBD98] gibt einen Überblick über die gebräuchlichen Ähnlichkeitssuchen in chemischen Datenbanken. Um die chemische Ähnlichkeit zu bestimmen werden Fingerprints (Bitvektor bestimmter Länge) von beiden Molekülen berechnet. Diese können genutzt werden, um mit einem geeigneten Distanzmaß verglichen zu werden. [JWD] hat festgestellt, das in diesem Fall der Tanimoto Koeffizient (Gleichung 2.3) die besten Ergebnisse liefert. Die folgenden Gleichungen beschreiben die Berechnung des Tanimoto Koeffizienten aus den Fingerprints von Molekül A und B.

$$
\begin{aligned}
X_A &= (x_{1A}, x_{2A}, \ldots, x_{jA}, \ldots, x_{nA}) \text{ Vektor des Moleküls A} \\
X_B &= (x_{1B}, x_{2B}, \ldots, x_{jB}, \ldots, x_{nB}) \text{ Vektor des Moleküls B} \\
a &= \sum_{j=1}^{j=n} x_{jA} \text{ Anzahl der Bits in A, die auf 1 gesetzt sind} \\
b &= \sum_{j=1}^{j=n} x_{jB} \text{ Anzahl der Bits in B, die auf 1 gesetzt sind} \\
c &= \sum_{j=1}^{j=n} x_{jA} x_{jB} \text{ Anzahl der Bits, die auf 1 gesetzt sind in A und B} \\
S_{A,B} &= \frac{c}{a+b-c} \text{ Tanimoto Ähnlichkeit}
\end{aligned}
\tag{2.3}
$$

Der Tanimoto Koeffizient für dichotome Werte liefert eine Ähnlichkeit zwischen 0 und 1, wobei 0 sehr unähnliche und 1 identische Moleküle beschreibt. Abbildung 2.15 zeigt als Beispiel zwei Bitvektoren von den Molekülen A und B, die aus Darstellungsgründen 18 lang sind. Mit 1 sind erfüllte Strukturmerkmale notiert und mit 0 nicht vorhandene. Die Anzahl der gemeinsamen Merkmale beträgt $c = 8$. Daraus lässt mit Hilfe des Tanimoto Koeffizienten die chemische Ähnlichkeit $S_{A,B}$ bestimmen, die von MetFrag benutzt wird, um gleichartigee Strukturen zusammenzufassen (siehe Abschnitt 3.1.7)

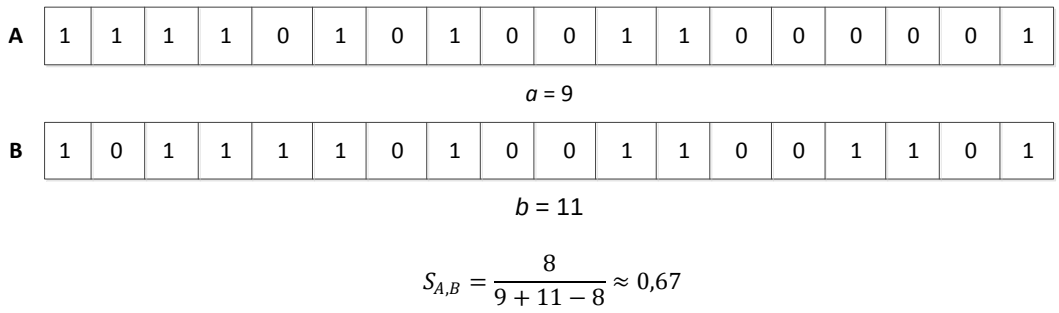


Abbildung 2.15. Beispiel zur Berechnung der Tanimoto Distanz von zwei Bitvektoren der Länge 18 nach [LG03].

2.5 Energieoptimierung von Molekülen

Ein weiteres Gebiet der Cheminformatik bzw. des Molecular Modelling beschäftigt sich mit der Strukturoptimierung von Molekülen. Durch diese ist es möglich, aus dem Molekülgraph die dreidimensionale Struktur eines Moleküls zu bestimmen, die von MetFrag in der Vorverarbeitung (siehe Abschnitt 3.1.2) verwendet wird.

2.5.1 Empirische Methode

Um eine erste Näherung der 3D Struktur eines Moleküls zu erreichen, werden häufig Kraftfelder eingesetzt. Dabei handelt es sich um eine Methode der Molekülmechanik, die im Gegensatz zur Quantenmechanik (Kapitel 2.5.2) Elektronen und Nuklei von Atomen nicht mit in der Berechnung berücksichtigt. Es wird angenommen, dass die Atome eines Moleküles untereinander durch harmonische Kräfte interagieren. Das Hookesche Gesetz (elastische Verformung einer Feder) bildet dabei die Grundlage zur Berechnung. Vereinfacht kann man sich Atome als Gummibälle unterschiedlicher Größe, die durch Federn unterschiedlicher Länge (Bindungen) verbunden sind, vorstellen. Zur Berechnung der Geometrie wird die Gesamtenergie (E_{tot}) des Moleküles minimiert [HSRF08]:

$$E_{tot} = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{elec} + \dots$$

Dabei werden die einzelnen Energieterme E_{str} (Bindungslänge), E_{bend} (Bindungswinkel), E_{tors} (Torsionswinkel), E_{vdw} (Van-der-Waals-Wechselwirkungen) und E_{elec} (elektrostatische Wechselwirkungen) addiert. Unterschiedliche Kraftfelder nutzen andere Energieterme und Berechnungen dieser für die Ermittlung der Gesamtenergie. Die Kraftfelder beinhalten empirisch bestimmte Idealwerte für die Parameter und jede Abweichung von diesen erhöht die Gesamtenergie. Zur Minimierung der Energiefunktion kann ein Gradientenabstieg verwendet werden. Dabei wird die Gesamtenergie des initialen Moleküles berechnet und bei der Bewegung eines Atomes in verschiedene Richtungen wird diese weiter beobachtet. Dieser Prozess wird für alle Atome wiederholt, bis die Abbruchbedingung erfüllt ist und damit ein lokales Minimum erreicht wurde. Eine nachfolgende weitere Optimierung ist unerlässlich. Als Beispiel für verschiedene Kraftfelder sei MMFF94 [Hal96], Ghemical [HP01] und UFF [RCC+92] genannt, die alle in OpenBabel [OBJ+11] enthalten sind.

2.5.2 Ab-initio und semi-empirische Methoden

Im Gegensatz zur Molekülmechanik verwenden ab-initio Methoden keine empirischen Parameter und werden vor allem in Bereichen ohne experimentelle Daten eingesetzt. Als Beispiel einer solchen quantenchemischen Methode ist die Dichtefunktionaltheorie (DFT), die eine große Laufzeit hat, aber sehr genaue Ergebnisse liefert. Zwischen den ab-initio Berechnungen und der Molekülmechanik gibt es riesige Unterschiede in Genauigkeit und Geschwindigkeit. Semi-empirische Methoden versuchen das Beste aus den beiden Welten, d.h. Schnelligkeit und Präzision, zu vereinen. Ähnlich den quantenmechanischen Berechnungen, aber auf empirische Werte für rechenintensive Aufgaben zurückgreifend, eignen sie sich auch für größere Moleküle. Des Weiteren werden nur Valenzelektronen (Außenelektronen) bei der Berechnung beachtet, das in einer weiteren Beschleunigung resultiert. MOPAC [Ste90] ist ein Programm, dass verschiedene semi-empirische Methoden, zum Beispiel AM1 [DZHS85], implementiert hat.

3 MetFrag Architektur und Implementation

Der Hauptteil der Dissertation beschreibt MetFrag, das im Rahmen der Arbeit entwickelt worden ist. Ziel des Programmes ist es, die gemessene Verbindung eines MS/MS Spektrums zu identifizieren beziehungsweise passende Kandidaten zu liefern. MetFrag ist für hochauflösende ESI-MS/MS Spektren entwickelt worden, aber kann auch mit GC/EI-MS Daten (Nominalmassen, siehe Abschnitt 4.7) genutzt werden.

3.1 Arbeitsphasen

Die grundlegende Idee von MetFrag ist, alle möglichen Fragmente eines Kandidaten zu generieren und diese mit den Peaks aus dem gemessenen Spektrum zu vergleichen. Passende Kandidaten können beispielsweise nach Masse oder Summenformel ausgewählt werden. Das Resultat eines MetFrag Laufes ist eine Liste von Molekülen, die nach einem Score geordnet sind. Außerdem werden sehr ähnliche Kandidaten zusammengefasst, um die Resultate übersichtlicher anzeigen zu können. Das Flussdiagramm in Abbildung 3.1 gibt einen Überblick über die Schritte, die in MetFrag durchgeführt werden.

3.1.1 Kandidatensuche

Um passende Kandidatenmoleküle zu finden, wird mit der exakter Masse oder Summenformel in einer Moleküldatenbank gesucht. Hierfür kann man entweder einen verfügbaren WebService nutzen oder den kompletten Datenbestand, zum Beispiel von KEGG oder PubChem, herunterladen. ChemSpider kann nur über den WebService abgefragt werden, da kein Download der kompletten Daten angeboten wird.

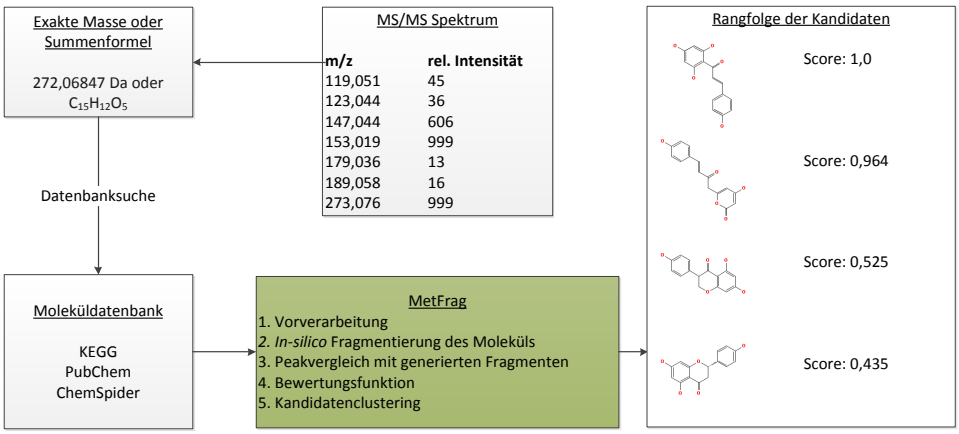


Abbildung 3.1. Aus einem MS Spektrum kann man die exakte Masse oder sogar die Summenformel der gemessenen Verbindung bestimmen. Dies wird genutzt, um eine Moleküldatenbank abzufragen. Jeder Kandidat wird daraufhin fragmentiert und die generierten Fragmente werden den Peaks zugeordnet. Das Scoring sortiert die Kandidaten nach ihrer Bewertung, wobei ähnliche Kandidaten mit gleichem Score in einem Clustering zusammengefasst werden.

Für die Evaluierung ist eine Kopie PubChem und KEGG in eine lokale Datenbank importiert worden. Das relationales Datenbankmanagementsystem (RDBMS) Postgres 9.0 mit der Chemie-Erweiterung pgchem[19] 1.3-GiST [Sch10] ist verwendet worden. Das Datenbank Schema ist in Abbildung 3.2 dargestellt und umfasst vier Tabellen: `library`, `substance`, `compound` und `name`. Hervorzuheben ist, dass `compound` nur eindeutige Strukturen enthält. Die Spalte `mol_structure` ist vom Typ `molecule`, der durch die pgchem Erweiterung bereitgestellt wird. Die Besonderheit dabei ist, dass hier die komplette Struktur eines Moleküls gespeichert wird und dadurch verschiedene Operationen wie zum Beispiel die Fingerprintgenerierung oder das Abgleichen von Substrukturen direkt als Operationen in der Datenbank zur Verfügung stehen. Die Tabelle `substance` kann zu einer Struktur verschiedene Datenbank IDs speichern. Die Namen der Strukturen sind in `name` enthalten und die verfügbaren Moleküldatenbanken in `library`.

Diese lokale Datenbank wird im Weiteren für MassStruct (Kapitel 3.3) verwendet und bildet auch die Datengrundlage für die Weboberfläche. Downloads von

[19] `http://pgfoundry.org/projects/pgchem` - Abgerufen im Mai 2011

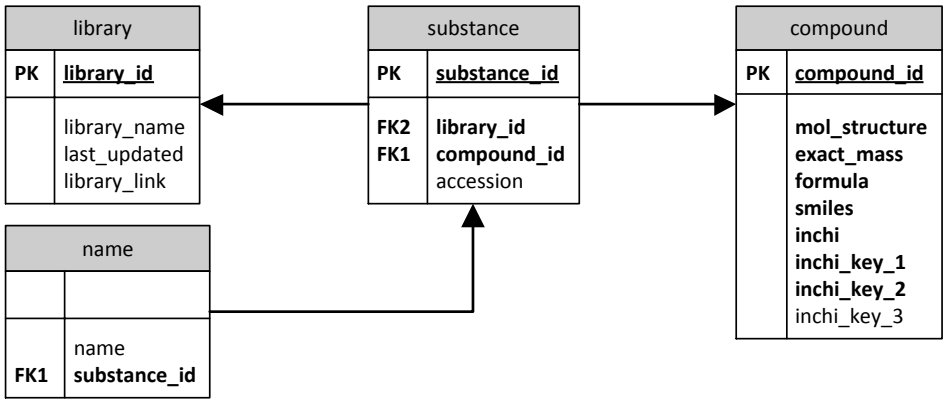


Abbildung 3.2. Datenbankschema zum Speichern der Verbindungen aus KEGG und PubChem. Die Tabelle `compound` speichert alle Verbindung (eindeutig) und unter anderem auch deren Struktur in der Spalte `molstructure`. Durch den Fremdschlüssel `compound_id` von `compound` kann in `substance` einer Struktur mehrere `accession` (Datenbank ID aus der Ursprungsdatenbank) zugeordnet werden. Dadurch wird eine redundante Speicherung der Strukturen in `compound` vermieden. Die Tabelle `library` enthält die hinzugefügten Datenbanken (z.B. Pubchem und KEGG) und `name` die möglichen vorhandenen Namen (entsprechend des Eintrags in der Ursprungsdatenbank) einer Verbindung.

PubChem und KEGG vom 4. Quartal 2010 sind eingefügt worden. Eine Ausnahme bildet ChemSpider, das nur über ein WebService abgefragt werden kann.

Eine typische Anfrage von MetFrag sucht alle Kandidaten in einem bestimmten Massenbereich, was oft hunderte bis tausende Kandidatenmoleküle zurückliefert. Diese werden, wie im folgenden Abschnitt beschrieben, vorverarbeitet.

3.1.2 Molekülvorverarbeitung

Die Molekülvorverarbeitung ist notwendig, um die Stärke der Bindungen eines Moleküls beschreiben zu können. Diese Werte werden in der Scoring Funktion (Kapitel 3.1.6) von MetFrag verwendet, damit eine genauere Rangordnung der Kandidaten berechnet werden kann.

Abbildung 3.3 zeigt die einzelnen Vorverarbeitungsschritte, die alle Kandidaten durchlaufen. Ein Kandidat mit 2D Atomkoordinaten wird mit einem Kraftfeld (OpenBabel) und einer semi-empirischen Methode (MOPAC) strukturoptimiert. Die re-

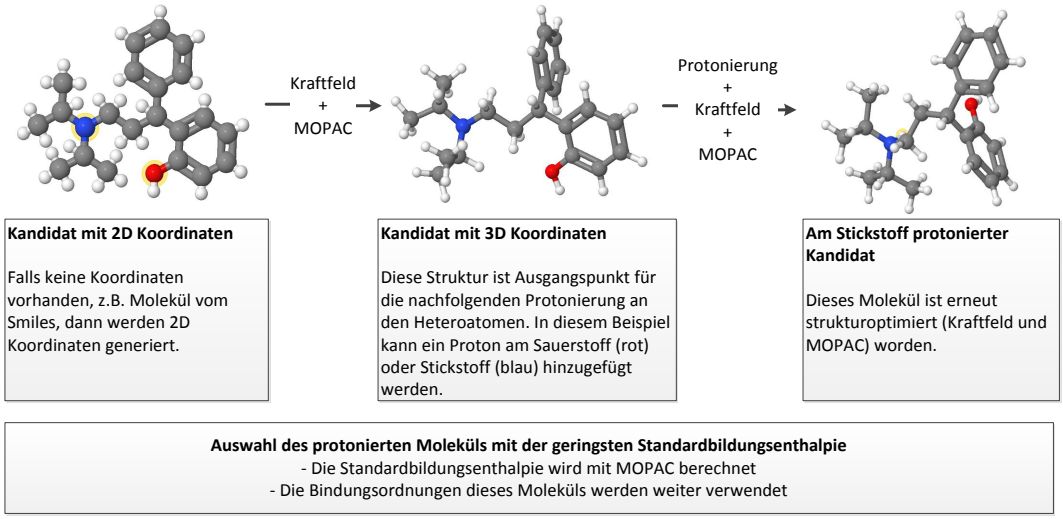


Abbildung 3.3. Exemplarisches Vorgehen zur Vorverarbeitung von CID: 20097272, um die Bindungen mit der Bindungsordnung zu annotieren. Im ersten Schritt wird ein mit 2D Koordinaten versehenes Molekül mit einer Kraftfeldmethode und MOPAC strukturoptimiert. Im nächsten Schritt wird diese Struktur einzeln an den Heteroatomen protoniert und daraufhin erneut struktoroptimiert. Am Ende wird das protonierte Molekül ausgewählt, das die geringste Standardbildungsenthalpie besitzt. Die daraus resultierenden Bindungen, die mit der Bindungsordnung annotiert sind, werden in der Scoring Funktion (Kapitel 3.1.6) verwendet.

sultierende Struktur wird einzeln an den Heteroatomen protoniert und nochmals strukturoptimiert. Das protonierte Molekül mit der kleinsten Standardbildungsenthalpie (von MOPAC berechnet) wird verwendet, um die Bindungen des Kandidaten zu annotieren. Zur Annotation werden die Bindungsordnungen aus dem Ergebnis von MOPAC verwendet. Bei der ESI Ionisierung (siehe Abschnitt 2.1.1) wird der Analyt protoniert und im Falle einer MS/MS Messung fragmentiert. Die Vorverarbeitung von MetFrag stellt eine Heuristik dar, dieses Prinzip mit ausreichend genauen, aber schnellen Verfahren nachzuahmen. Im folgenden werden die einzelnen Schritte, die nacheinander ausgeführt werden, genauer beschrieben.

Zum Spektrum passende Moleküle werden in der Regel aus einer Moleküldatenbank (siehe Abschnitt 3.1.1) heruntergeladen. Die Atomkoordinaten der Strukturen sind üblicherweise bereits vorhanden. Wird ein Molekül aus einem SMILES™ oder InChI™ generiert, so müssen zuerst 2D-Koordinaten (`--gen2D`) bestimmt werden. Implizite

Wasserstoffe werden explizit in die Molekülstruktur angefügt und die Koordinaten um den Nullpunkt (0,0,0) zentriert (`-c`):

OpenBabel: `babel --gen2D -c -i sdf input.sdf -o sdf outputGen2D.sdf`

Somit besitzen alle Moleküle den gleich Ausgangspunkt für die darauffolgende Kraftfeldoptimierung (siehe Kapitel 2.5.1), die eine erste Annäherung an die 3D-Struktur des Moleküls darstellt:

OpenBabel: `obminimize -n 4800 -sd -ff UFF outputGen2D.sdf > outputFF.pdb`.

Die Anzahl der Schritte `-n 4800`, `-sd` Gradientenabstieg und `-ff` das gewählte Kraftfeld `-ff UFF` werden durch die Parameter angegeben. `Ghemical`, `MMFF94` und `UFF` stehen in OpenBabel (Version 2.3.0) zur Verfügung. Die besten Ergebnisse sind mit dem „Universal Force Field“ (siehe Abschnitt 4.1) erzielt worden. Die resultierende Datei (`outputFF.pdb`) im „Protein Data Bank“ Format wird im nächsten Schritt in das „MOPAC Input Format“ umgewandelt und mit den folgenden MOPAC (siehe Abschnitt 2.5.2) Paramtern versehen: `AM1` - die verwendete semi-empirische Methode, `T=4800` - die maximal zur Verfügung stehende Laufzeit in Sekunden, `GEO-OK` - verhindert den Abbruch bei zu weit aneinander liegenden Atomen, `XYZ` - kartesisches Koordinatensystem, `BONDS` - Ausgabe der Bindungsordnungsmatrix:

MOPAC Parameter: `AM1 T=4800 AM1, GEO-OK, MMOK, XYZ, BONDS`

Die resultierende angenäherte 3D-Struktur des Moleküls wird einzeln an den Heteroatomen protoniert und erneut strukturoptimiert. Das protonierte Molekül mit der geringsten (von MOPAC berechneten) Standardbildungsenthalpie wird als das „Wahrscheinlichste“ angenommen. Die Bindungsordnungen dieses Moleküls werden verwendet, um die Bindungen des Kandidaten zu annotieren, und im CML Format (siehe Abbildung 2.9) abgespeichert. Bindungen zu Wasserstoffen werden nicht annotiert, weil diese nicht vom Fragmentierungsalgorithmus (siehe Abschnitt 3.1.3) betrachtet werden. Je kleiner der Wert der Bindungsordnung desto schwächer ist die Bindung zwischen den beiden Atomen. Beispielsweise besitzt die Kohlenstoffbindung von Ethan, Ethen und Ethin eine Bindungsordnung von 1,0, 2,0 bzw. 3,0

[Ste90]. Zusätzlich kann auch die Bindungslänge zwischen dem neutralen und protonierten Molekül mit der geringsten Standardbildungsenthalpie bestimmt werden. Abschnitt 4.1 beschreibt, warum letztendlich die Bindungsordnung statt der Bindungslänge in der Scoring Funktion von MetFrag verwendet wird.

3.1.3 In silico Fragmentierung

Das durch die Vorverarbeitung annotierte Molekül wird im folgenden Schritt in Fragmente zerlegt, das die Fragmentierung des Vorläuferions in Fragmentionen nachahmt (MS/MS). Dieses Vorgehen kann mit Hilfe eines Fragmentierungsbaumes (Abbildung 3.4) dargestellt werden. Durch Entfernen einer linearen Bindung zerfällt ein Molekül in zwei Fragmente. Bei Ringen und Ringsystemen müssen mindestens zwei Bindungen gebrochen werden, damit dieses auseinander bricht.

Im Folgenden werden die einzelnen Schritte der Fragmentierung (Abbildung 3.5) beschrieben, die für jeden Kandidaten einzeln durchgeführt werden. Als erstes wird das Kandidatenmolekül in eine Warteschlange eingereiht.

Im folgenden Schritt wird aus dieser Datenstruktur ein Kandidat bzw. eine Substruktur entnommen und alle Bindungen, die gebrochen werden können, in einer Liste gespeichert. Diese Liste wird nun nacheinander abgearbeitet, wobei für jede Bindung zwei Fälle unterschieden werden können: Zum einen kann eine Bindung Teil eines Ringes sein, zum anderen eine lineare Bindung. Falls der erste Fall zutrifft, dann muss eine weitere Bindung im Ring gebrochen werden. MetFrag entfernt jede Kombination der aktuellen Bindung mit einer weiteren aus dem Ring, um so alle möglichen Fragmente zu generieren. Bei einer linearen Bindung entstehen pro Iteration immer zwei Substrukturen, wobei bei einer Ringbindung maximal zwei Fragmente entstehen können, da es passieren kann, dass der Graph durch die beiden Bindungspaltungen noch immer an anderer Stelle verbunden ist.

Von diesen Strukturen wird die Masse bestimmt und überprüft, ob diese schwerer als der leichteste Peak aus dem gemessenen Spektrum sind. Falls dies nicht zutrifft, dann wird dieses Fragment nicht weiter betrachtet. Durch die Fragmentierung entstehen viele Strukturen doppelt, die im folgenden herausgefiltert werden.

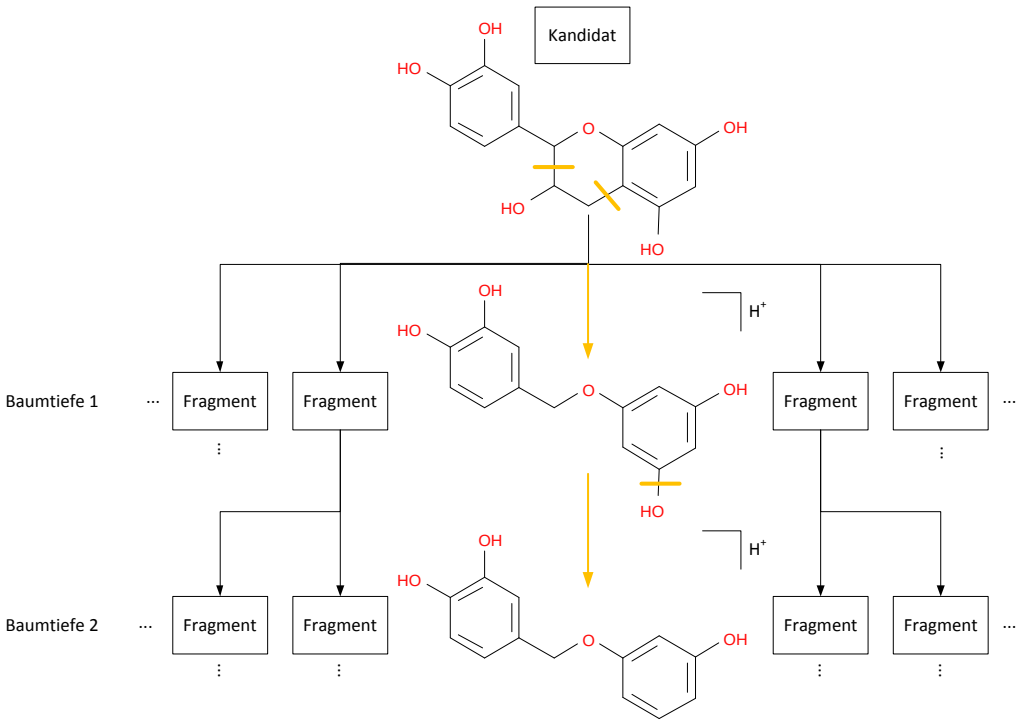


Abbildung 3.4. Ausschnitt eines Fragmentierungsbaumes des Kandidaten Epicatechin (CID: 72276). In der Mitte ist ein Ringbruch (Baumtiefe 1) dargestellt. In der darauffolgenden Fragmentierung wird eine lineare Bindung gebrochen, um das abgebildete Fragment in Baumtiefe 2 zu erhalten.

Um Fragmente, die bereits in einer vorhergehenden Iteration generiert wurden herauszufiltern, muss eine Redundanzüberprüfung durchgeführt werden. Dafür kann eine Isomorphieprüfung benutzt werden, wodurch nur eindeutige Strukturen generiert werden. Durch die immense Anzahl an generierten Fragmenten pro Kandidat, ist diese Methode nicht in angemessener Zeit durchführbar (siehe Kapitel 3.1.5).

Letztendlich verwendet MetFrag für die Filterung von doppelten Strukturen eine Summenformel basierte Redundanzüberprüfung: Für jedes generierte Fragment wird die Summenformel bestimmt und überprüft, ob diese bereits vorhanden ist. Falls dies zutrifft, wird überprüft, ob die Generierung dieses Fragmentes energetisch günstiger ist, als das bereits vorher entstandene Fragment: Die Bindungsordnungen (Kapitel 3.1.6) der gebrochenen Bindungen, um dieses Fragment zu generieren, werden aufsummiert und das Fragment mit dem größeren Ergebnis verworfen. Der Nach-

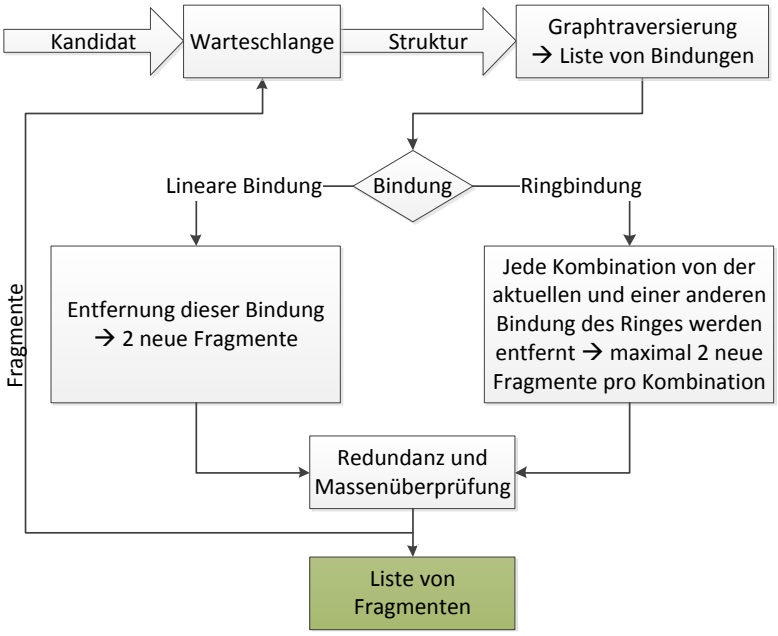


Abbildung 3.5. Fragmentierungsalgorithmus von MetFrag: Der Kandidat wird in die Warteschlange eingefügt. Je nach Bindungstyp, linear oder zyklische Bindung, wird eine oder jede Kombination von zwei Bindungen im Ring entfernt. Es resultieren maximal zwei Fragmente pro Kombination im Ring bzw. immer zwei Fragmente wenn eine lineare Bindung gebrochen wird. Die darauffolgende Massen- und Redundanzüberprüfung verhindert das Substrukturen doppelt oder zu leichte Fragmente generiert werden. Die resultierenden Fragmente werden zum einen gespeichert und zum anderen erneut in die Warteschlange eingefügt, wenn die vorher festgelegte Baumtiefe noch nicht erreicht wurde.

teil dieser Methode ist, dass nicht jedes mögliche Fragment generiert wird, aber die „energetisch“ Sinnvollsten. In Abschnitt 3.1.5 wird am Beispiel von Naringenin gezeigt, dass eine Isomorphiüberprüfung sehr zeitaufwendig und nicht unbedingt notwendig ist. Außerdem wird in Abschnitt 4.6.1 gezeigt, dass mit Baumtiefe 1 die besten Ergebnisse erzielt werden. Dadurch ist es wenig sinnvoll eine langsame Isomorphieüberprüfung durchzuführen, da durch die Summenformelüberprüfung die gleichen Massen (Baumtiefe 1) abgedeckt werden. Erst mit steigender Baumtiefe kann die genaue Überprüfung der Isomorphie Sinn machen, da aus den Fragmenten wieder neue generiert werden.

Nicht redundante Fragmente, die schwerer sind als der leichteste Peak, werden wieder in die Warteschlange eingefügt und können im nächsten Durchlauf weiter fragmen-

	$[M+H]^+$	$[M-H]^-$	M^+	M^-
Resultierende Fragmentmasse:	FM + WM - EM	FM - WM + EM	FM - EM	FM + EM

Tabelle 3.1. Unterschiedliche Massenspektrometer und Modi erfordern unterschiedliche MetFrag Einstellungen. Je nach Messmethode werden die neutralen Fragmentstrukturen nach diesen Regeln modifiziert. (M - Molekül, H - Wasserstoff, FM - Fragmentmasse, WM - Wasserstoffmasse, EM - Elektronenmasse)

tiert werden. Alle entstandenen Fragmente werden nachverarbeitet, da zum Beispiel Neutralverluste auftreten können und sich dadurch die Anzahl der Wasserstoffe des Fragmentions ändern kann. Der folgende Abschnitt gibt einen genauen Einblick, wie die (neutralen) Fragmentstrukturen den Peaks (Ionen) zugeordnet werden können.

3.1.4 Peak-Fragment Vergleich

Im Massenspektrometer können nur geladene Verbindungen (Ionen) gemessen werden. Die im vorherigen Schritt generierten Fragmente sind in der Regel ungeladen und müssen daher noch modifiziert werden, um eine Übereinstimmung zwischen Peak und Fragmentmasse zu erreichen. Diese Massendifferenz wird je nach Messmethode und Ionisierungsquelle des Instrumentes ($[M+H]^+$, $[M-H]^-$, M^+, M^-) ausgeglichen (siehe Abschnitt 2.1.1), indem die Masse von Wasserstoff und einem Elektron zu der neutralen Fragmentmasse addiert oder subtrahiert wird. Tabelle 3.1 zeigt die Vorgehensweise von MetFrag für die gebräuchlichsten Ionisierungsmethoden.

Da jedes Instrument einen mehr oder weniger großen Messfehler besitzt, wird MetFrag die gerätespezifische Abweichung als Parameter übergeben. Hierfür können zwei Parameter angegeben werden: Ein absoluter (mzabs) und relativer (mzppm) Wert. Beispielsweise beträgt der relative Fehler bei einer Masse von 800 Da und 10 mzppm 0,008 Da und 0,001 Da bei 100 Da. Diese Abweichung ist bei kleinen Massen sehr gering, sodass zusätzlich eine absolute verwendet wird. Beide Werte, mzabs und mzppm, werden addiert und bilden zusammen die erlaubte Abweichung, um einen Peak einer Fragmentstruktur zuzuordnen.

Weiterhin ist zu beachten, dass der Fragmentierungsalgorithmus von MetFrag keine Wasserstoffe abspaltet. Diese können zum Beispiel auch in Form von Neutralverlusten (Massendifferenz des Ions vor und nach der Fragmentierung) mit anderen

Atomen abgespalten werden. Ein Beispiel eines Neutralverlustes in Form von Wasser ist in Abbildung 3.6 dargestellt.

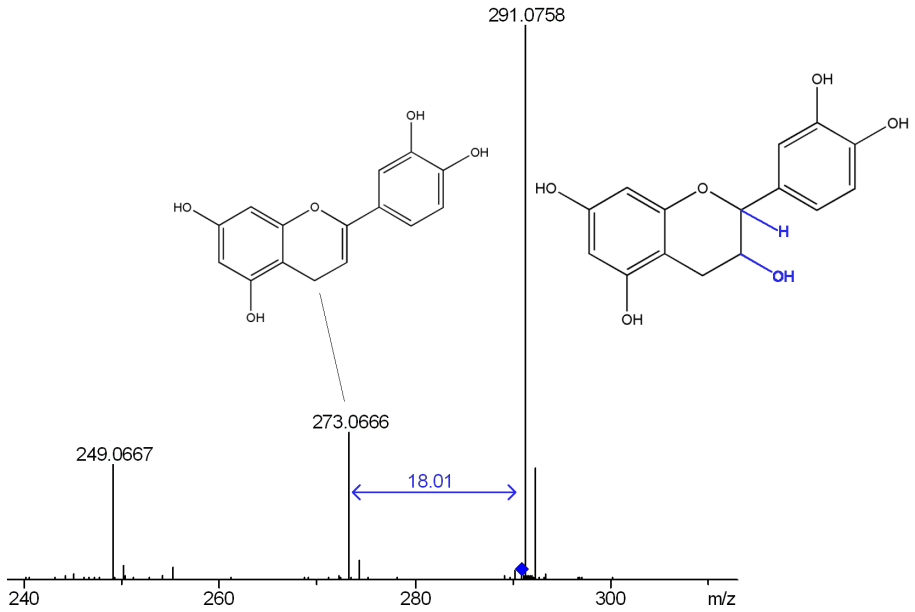


Abbildung 3.6. Ausschnitt des MS/MS Spektrums von Epicatechin: H_2O (Masse: 18,0105 Da) wird als Neutralverlust abgespalten.

Der in Abbildung 3.6 dargestellte Neutralverlust von Wasser, kann nicht durch Fragmentierung von MetFrag erklärt werden, da keine Wasserstoffe abgespalten werden. MetFrag nutzt daher Regeln, die es dennoch ermöglichen sollen, dass Fragment zuordnen zu können. Tabelle 3.2 zeigt die verwendete Regelmenge, die häufig auftretende Neutralverluste abdeckt. Die Tabelle ist einfach erweiterbar und benutzt SMARTS für das Suchen von typischen Strukturen des Neutralverlustes. Es können mehrere SMARTS pro Zeile eingetragen werden, wobei nur ein SMARTS zutreffen muss, damit die Regel angewendet wird.

Abbildung 3.7 zeigt ein Beispiel für alle möglichen Wasserverluste mit dem SMARTS: `[H][$([OH1][#6])][#6][#6][H]`. Gesucht wird eine Hydroxylgruppe, die mit einem Kohlenstoffatom verbunden ist. Dessen Nachbar ist mindestens ein CH, damit dieser SMARTS den Subgraphen zugeordnet werden kann.

Die Tabelle 3.2 deckt nur häufig vorkommende Neutralverluste ab, aber MetFrag besitzt auch einen Mechanismus, um Fragmente mit abweichender Anzahl von Wasserstoffen zuordnen zu können: Für jede Substruktur wird überprüft, ob durch die

Masse	Summenformel	SMARTS
18,011	H2O	[H][$([OH1][#6])][#6][H]
		[H][$([OH1][#6])][#6][#6][H]
		[H][$([OH1][#6])][#6][#6][#6][H]
27,011	HCN	[N][#6][H]
17,027	NH3	[$([NH2][#6])]([H])([H])[#6][H]
		[$([NH2][#6])]([H])([H])[#6][#6][H]
		[$([NH2][#6])]([H])([H])[#6][#6][#6][H]
46,005	HCOOH	[H][O][#6]([#6][H])=[O]
		[H][O][#6]([#6][#6][H])=[O]
		[H][O][#6]([#6][#6][#6][H])=[O]
31,042	CH3NH2	[#6H3][NR][#6H2]

Tabelle 3.2. Die von MetFrag verwendeten Neutralverlustregeln. Bei mehreren SMARTS muss nur eine der Regeln zutreffen. Es sind nur die Neutralverluste von Bedeutung, bei denen Wasserstoffe mit abgespalten werden.

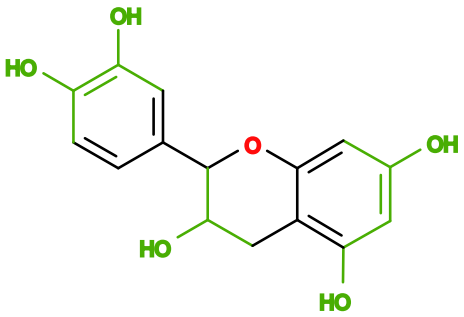


Abbildung 3.7. Epicatechin mit grün hervorgehobenen Substrukturen, die durch den SMARTS String `[H][$([OH1][#6])][#6][#6][H]` zugeordnet werden können.

Addition oder Subtraktion der Masse eines oder mehrerer Wasserstoffe (1,0078 Da) die Zuordnung des Fragmentes zu einem Peak ermöglicht. Die Anzahl der maximal zu addierenden Wasserstoffe richtet sich nach der Baumtiefe, in welcher das Fragment generiert worden ist. Das heißt, für ein Fragment, das sich im Fragmentierungsbaum in Tiefe 2 befindet, werden maximal zwei variable Wasserstoffe erlaubt. In [WD11] sind einige Beispiele von ESI-MS/MS Spektren mit Massendifferenzen von $\pm 1,0078$ Da gezeigt worden.

Außerdem kann es vorkommen, dass von einem Fragment mehrere Neutralverluste abgespalten werden. Daher erlaubt MetFrag verschiedene Kombinationen von Regeln, die pro Fragment angewendet werden. Standardmäßig wird die Maximalzahl

von Regelkombinationen mit 3 angegeben, d.h. eine Regel kann maximal drei mal kombiniert werden, um dem Fragment ein Peak zuzuordnen zu können. Dies hat sich als guter Kompromiss zwischen Geschwindigkeit und Leistung der Identifizierung herausgestellt.

Eine Möglichkeit, um die exponentiell wachsende Laufzeit (siehe Kapitel 4.4) einzuschränken, ist die Baumtiefe zu reglementieren, die vorher beschriebenen Redundanzüberprüfung, sowie zu leichte Fragmente zu ignorieren.

3.1.5 Beispiel einer Fragmentierung

Um einen Überblick der bisher angewendeten Schritte zu gewinnen, wird im folgenden am Beispiel eines MS/MS Spektrums von Naringenin (MassBank ID: PB000123) die bisherige Vorgehensweise erläutert. Einziger Kandidat ist die gemessene Verbindung (CID: 932), die mit MetFrag prozessiert wird.

Im ersten Schritt werden alle möglichen Fragmente des Kandidaten generiert (Abschnitt 3.1.3). In Baumtiefe 1 generiert MetFrag mit der Summenformel als Redundanzüberprüfung 25 Fragmente in 1,7 Sekunden. Erhöht man die Baumtiefe auf 2 werden in 2,1 Sekunden 57 Fragmente generiert. Wird anstelle der Summenformel eine richtige Isomorphieüberprüfung durchgeführt, dann werden 39 (Baumtiefe 1: 2,6s) bzw. 151 (Baumtiefe 2: 5,7s) Fragmente generiert. Naringenin hat eine monoisotopische Masse von 272,068 Da und ist nicht besonders komplex aufgebaut. Schon bei diesem kleinen Molekül braucht eine echte Isomorphieüberprüfung mehr als doppelt so lang (Baumtiefe 2), wie die Redundanzüberprüfung mit Summenformel. Auch bei dem Test auf Isomorphie von zwei Graphen ist vorher die Summenformel verglichen worden, sodass nur in Frage kommende Strukturen überprüft werden. Im nächsten Schritt werden den generierten Fragmenten Peaks zugeordnet (Abschnitt 3.1.4). Es macht in diesem Beispiel keinen Unterschied, welche Redundanzüberprüfung oder Baumtiefe verwendet wird, da zu jeden Peak ein Fragment zugeordnet werden kann. Abbildung 3.8 zeigt die wichtigsten Fragmente des MS/MS Spektrums von Naringenin (MassBank ID: PB000123), die von einem Experten annotiert worden sind.

Abbildung 3.9 zeigt die Fragmente, die durch MetFrag zugeordnet werden konnten. Die abgebildeten Strukturen mit impliziten Wasserstoffen entsprechen den un-

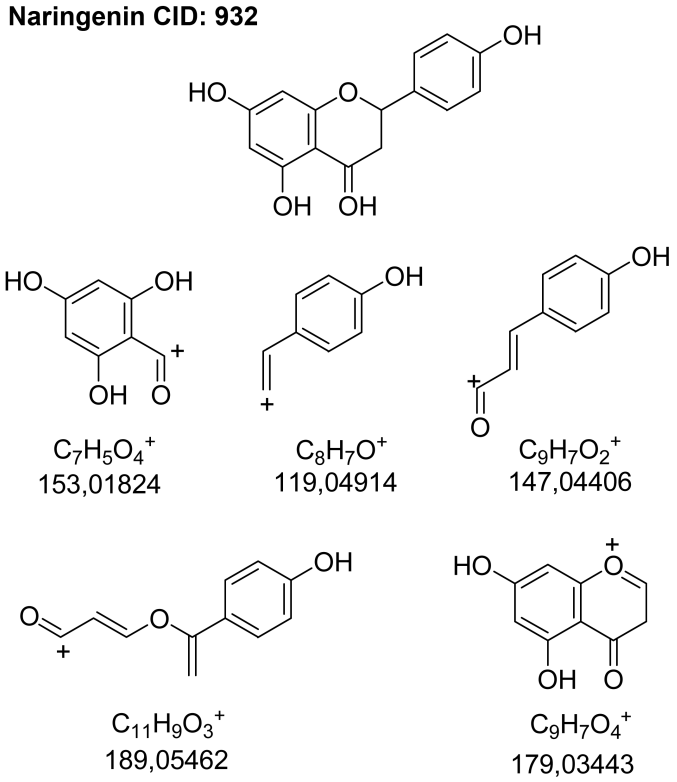


Abbildung 3.8. Annotierte Fragmente zu den wichtigsten Peaks des MS/MS Spektrums von Naringenin (MassBank ID: PB000123).

geladen Strukturen und dienen nur zur Verdeutlichung, wie der Algorithmus von MetFrag funktioniert. Diese Strukturen entsprechen nicht den als Peaks sichtbaren Fragmentionen aus dem Spektrum, können aber daraus abgeleitet werden. MetFrag ist in der Lage, die richtige Fragmentstruktur für die Peaks 119, 147, 153 und 179 zuzuordnen. Die Neutralverluste sind durch SMARTS (Tabelle 3.2) identifiziert und deren Masse von der des Fragmentes (FM) subtrahiert worden. Die Struktur von Peak 189 ist erklärt worden, jedoch hält ein Experte eine andere für wahrscheinlicher (Abbildung 3.8).

3.1.6 Bewertungsfunktion

Nachdem die Peaks zu den Fragmenten zugeordnet worden sind, wird durch eine Scoring Funktion die Rangfolge der Kandidaten bestimmt. Kandidaten, deren Fragmente schwere Peaks mit hoher Intensität erklären und bei denen die Bindungsord-

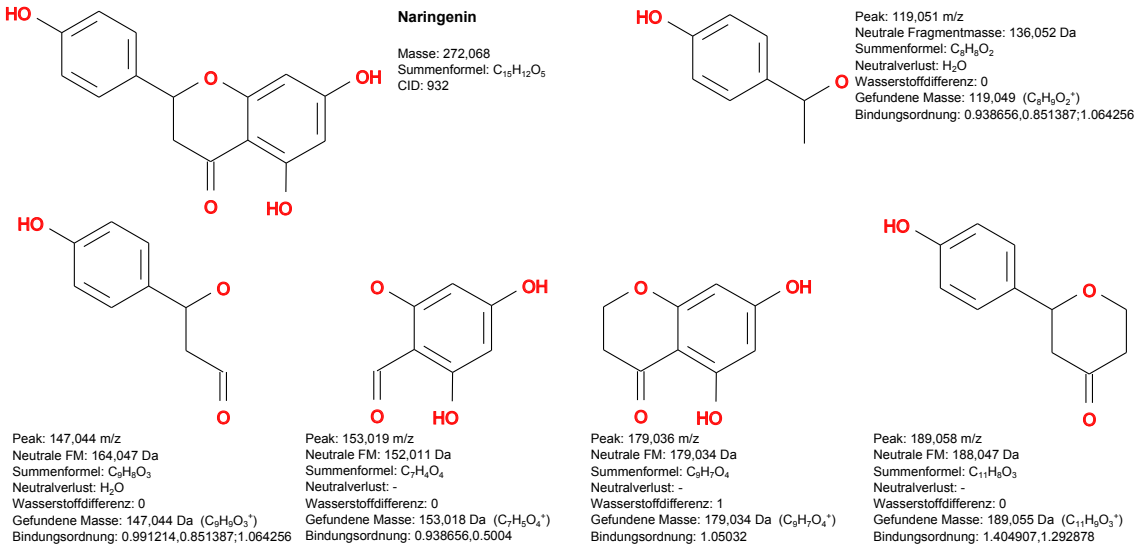


Abbildung 3.9. Von MetFrag zugeordnete Fragmente zum MS/MS Spektrum von Naringenin (MassBank ID: PB000123). Die abgebildeten Fragmentstrukturen dienen nur zur Veranschaulichung der Vorgehensweise von MetFrag und stellen bewusst keine Ionen dar. Abbildung 3.8 zeigt die korrekt annotierten Fragmentionen. Durch Komma getrennte Werte der Bindungsordnung verdeutlichen zwei gebrochene Bindungen (Ring) und ein „;“ zeigt einen weiteren Bruch an. Im Falle des Peaks 119 liegt ein Ringbruch vor und die Abspaltung von Wasser als Neutralverlust (Bindungsordnung des verbundenen Sauerstoffs) mit einbezogen wird.

nung der fragmentierten Bindungen besonders klein ist, sollen eine möglichst gute Bewertung bekommen. Da schwere Fragmentionen in der Regel charakteristischer sind als leichte [SS94], bekommen diese Kandidaten einen besseren Score. Analog werden erklärte Peaks mit hoher Intensität besser bewertet als mit niedriger, da wenig intensive Peaks auch Rauschen darstellen können. Je kleiner die Bindungsordnung einer Bindung, desto leichter bricht diese (Abbildung 4.4). Deshalb werden kleine Bindungsordnungen der gebrochenen Bindung belohnt. Gleichung 3.1 zeigt die verwendete Scoring Funktion von MetFrag, die neben der Masse und Intensität auch die Bindungsordnung des zugeordneten Fragmentes berücksichtigt.

$$S_i = \sum_{n=1}^{N_i} \left(\frac{mass_n}{max\,(mass)} \right)^a \left(\frac{int_n}{max\,(int)} \right)^b \prod_{k=1}^{K_n} \left(1 - \frac{f(bo_{n,k})}{2} \right)^c \tag{3.1}$$

$$f(bo_{n,k}) \quad = \quad \begin{cases} bo_{n,k} & , bo_{n,k} < 2 \\ 2 & , bo_{n,k} \geq 2 \end{cases}$$

Der endgültige Score S_i eines Kandidaten i ergibt sich aus der Summe der Produkte der drei Terme, die mit den Parametern a, b und c potenziert werden. Die Terme $\left(\frac{mass_n}{max(mass)}\right)^a$ und $\left(\frac{int_n}{max(int)}\right)^b$ berechnen die gewichtete Masse und gewichtete Intensität des erklärten Peaks n. Beide Werte werden über alle Kandidaten normiert und liegen jeweils zwischen 0 und 1. Es wird die schwerste Masse bzw. größte Intensität, die durch ein Fragment eines Kandidaten erklärt wird, auf 1 normiert. Je größer die jeweiligen Werte ausfallen, desto größer ist deren Anteil am Score. Im letzten Teil der Scoring Funktion wird das Produkt der Bindungsordnungen $\prod_{k=1}^{K_n}\left(1-\frac{f(bo_{n,k})}{2}\right)^c$ der entfernten Bindungen k eines Fragmentes n berechnet. Jeder Wert der Bindungsordnung $bo_{n,k}$ wird durch die Funktion f auf maximal 2 begrenzt und auf 0 bis 1 normiert. Es wird als Heuristik angenommen, dass je größer die Bindungsordnung, desto stärker ist die Bindung und desto schwerer fragmentiert diese. Daher wird in der Scoring Funktion $1-\frac{f(bo_{n,k})}{2}$ verwendet. Fragmente, die durch das Entfernen einer Bindungsordnung ≥ 2 einem Peak zugeordnet werden können, haben keinen Einfluss auf den Score S_i des Kandidaten, da das Produkt 0 wird. Für jeden Term der Gleichung sind die Parameter a, b und c trainiert worden, um die optimale Gewichtung der Teilgleichungen zu finden. Abschnitt 4.5 beschreibt die durchgeführte Parameteroptimierung, mit $a = 0,99$, $b = 0,87$ und $c = 0,17$ als Ergebnis.

3.1.7 Strukturclustering

Viele Kandidaten besitzen ähnliche Strukturen und bekommen deshalb auch gleiche Scores zugeordnet. Abbildung 3.10 zeigt die neun besten Kandidaten mit der Anzahl von MetFrag erklärten Peaks. Hier ist zu sehen, dass die blau und grün hinterlegten Strukturen Stereoisomere und dadurch nicht durch MS/MS unterscheidbar sind. Deshalb werden diese Moleküle nach der Vorgehensweise von [But99] zusammengefasst. Es werden Fingerprints von allen Strukturen erstellt, wobei solche zusammengefasst werden, die eine Tanimoto Ähnlichkeit (siehe Kapitel 2.4.3) $> 0,95$ und denselben Score besitzen. Der Repräsentant eines Clusters ist der Kandidat, der in

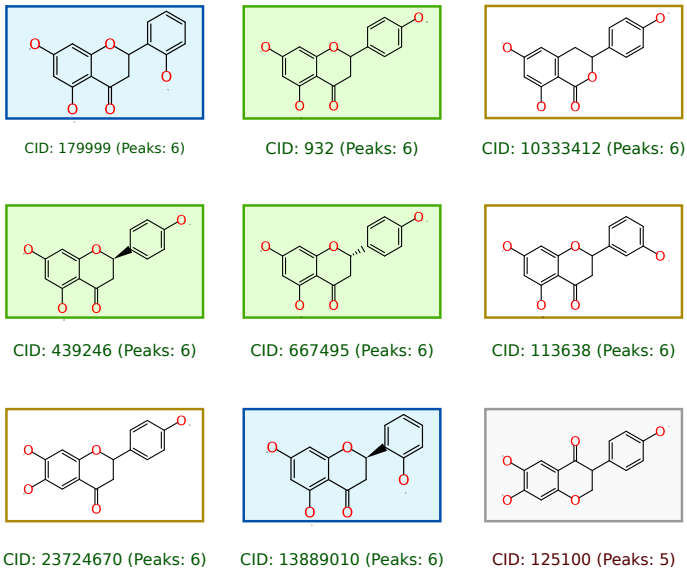


Abbildung 3.10. Ergebnis von MetFrag mit den neun besten Kandidaten aus [WSMHN10]. Die besten acht Kandidaten erklären alle sechs Peaks. Die grün und blau hinterlegten Strukturen sind Stereoisomere, die durch MS/MS nicht unterscheidbar sind und daher zusammengefasst werden können. Die Tanimoto Ähnlichkeit der restlichen drei Strukturen (CID: 10333412, CID: 113638 und CID: 23724670) ist kleiner als 0,95 und somit nicht geclustert. (aus [WSMHN10])

der Ergebnisliste als erstes aufgetreten ist. Dies erhöht vor allem die Übersichtlichkeit der Ergebnisse, da weniger Kandidaten angeschaut werden müssen. Es werden trotzdem alle Kandidaten gespeichert, sodass immer eine vollzählige Ergebnisliste vorhanden ist.

3.2 Weboberfläche und API

Dieser Abschnitt beschreibt, wie MetFrag über einen Web Browser und in eigener Software verwendet werden kann. Zum einen kann MetFrag Spektren im MassBank Format automatisiert als Kommandozeilenversion verarbeiten. Außerdem steht ein BioMoby [Con08] WebService bereit, der es beispielsweise ermöglicht MetFrag auf einfache Weise in Workflows zu benutzen. Biomoby stellt neben einer zentralen Registrierung auch Ontologien bereit, um beispielsweise eine Zeichenkette vom Typ KEGG ID übergeben zu können.

Zum anderen kann MetFrag (ohne Vorverabeitung [WSMHN10]) komfortabel über eine Weboberfläche (Abbildung 3.11) bedient werden. Diese basiert auf Java Server Faces[20] mit der JSF RI 1.2[21] und benutzt ICEFaces 1.8.2[22] als Komponentenbibliothek. Tomcat 6[23] wird als Servlet Container verwendet. Dadurch ist es möglich die in Java geschriebenen Klassen von MetFrag auch für die Weboberfläche zu verwenden. Außerdem kann auf einfache Weise Feedback gesammelt werden, um den Algorithmus weiter zu verbessern.

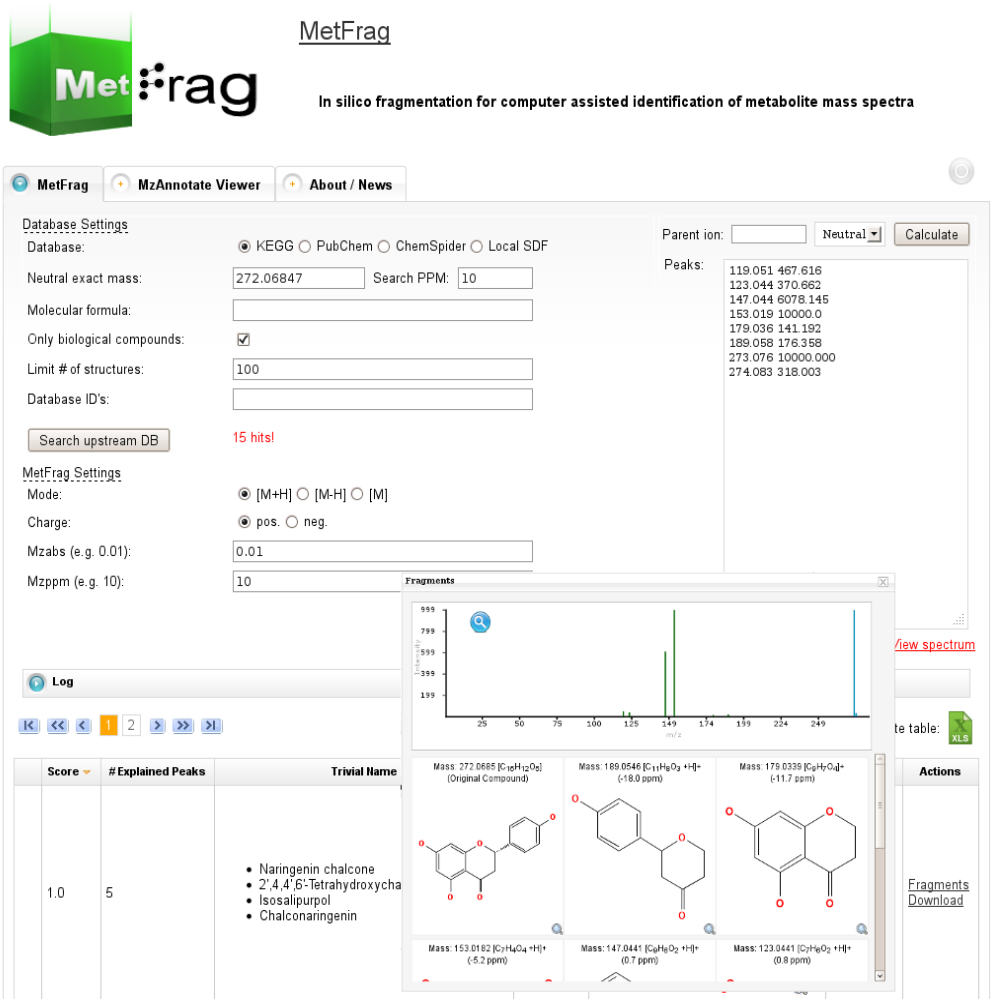


Abbildung 3.11. Beispielspektrum von Naringenin (PubChem CID: 932) mit der nach Score geordneten Rangfolge. Das Detailfenster zeigt durch MetFrag annotierte Fragmente.

[20] `http://java.sun.com/javaee/javaserverfaces/` - Abgerufen Mai 2011
[21] `http://javaserverfaces.java.net/` - Abgerufen Mai 2011
[22] `http://www.icefaces.org/` - Abgerufen Mai 2011
[23] `http://tomcat.apache.org/` - Abgerufen Mai 2011

Die Weboberfläche besitzt eine einfache API, die es externer Software erlaubt, direkt Daten an diese zu übergeben. GET-Variablen[24] übermitteln die Peakliste, exakte Masse und die Summenformel des Analyten direkt in der URL.

Durch eine begrenzte Rechenzeit auf dem WebServer ist standardmäßig eine Limitierung auf 100 Kandidaten pro Anfrage eingestellt, die vom Nutzer auch beliebig erhöht werden kann. Damit nicht unnötig viele Kandidaten bearbeitet werden, wird im folgenden Abschnitt eine Möglichkeit vorgestellt, um die wahrscheinlichsten Strukturen als erstes zu prozessieren.

3.3 Intelligente Kandidatensuche - MassStruct

Die Suche in Molekül datenbanken (Abschnitt 3.1.1) liefert die Kandidaten in ungeordneter Reihenfolge zurück. Das hat den Nachteil, dass sich unter Umständen bei 12 000 Kandidaten der Richtige erst an hinterer Position befindet. Um dieses Szenario zu vermeiden, wird im folgenden MassStruct [HWN11] vorgestellt, das die Kandidaten in eine bestimmte Reihenfolge bringt.

Typischerweise besitzt Naringenin (Abbildung 3.11) einen Peak mit der Masse 147 Da und 153 Da. Das Auftreten beider Fragmentionen lässt Rückschlüsse auf die Stoffklasse der Flavonoide zu. Dieses Wissen kann genutzt werden, um Kandidaten als erste zurückzuliefern, die Substrukturen der Flavonoide enthalten. Daher lernt MassStruct in einem Vorverarbeitungsschritt eine Zuordnung von Masse und Fragment. Diese gespeicherten Informationen können für ein Anfragespektrum genutzt werden, um die Rangfolge der Kandidaten aus der Datenbank zu bestimmen.

Als Trainingsgrundlage dienen MassBank Spektren, deren Strukturen mit Hilfe von MetFrag fragmentiert werden. Die entstandenen Fragmente werden den Peaks zugeordnet und in einer Postgres Datenbank (Abbildung 3.13) mit pgchem[25] 1.3-GiST [Sch10] Erweiterung gespeichert. Abbildung 3.12 zeigt das beschriebene Training und das Prinzip, wie die Ergebnisse an MetFrag zurückgegeben werden. MetFrag

[24] `http://msbi.ipb-halle.de/MetFrag/LandingPage.jspx?mass=272.06847&peaks=119.051%20467.616;147.044%206078.145&formula=C15H12O5&database=pubchem` - Beispiel einer URL - Abgerufen im Dezember 2011

[25] `http://pgfoundry.org/projects/pgchem`

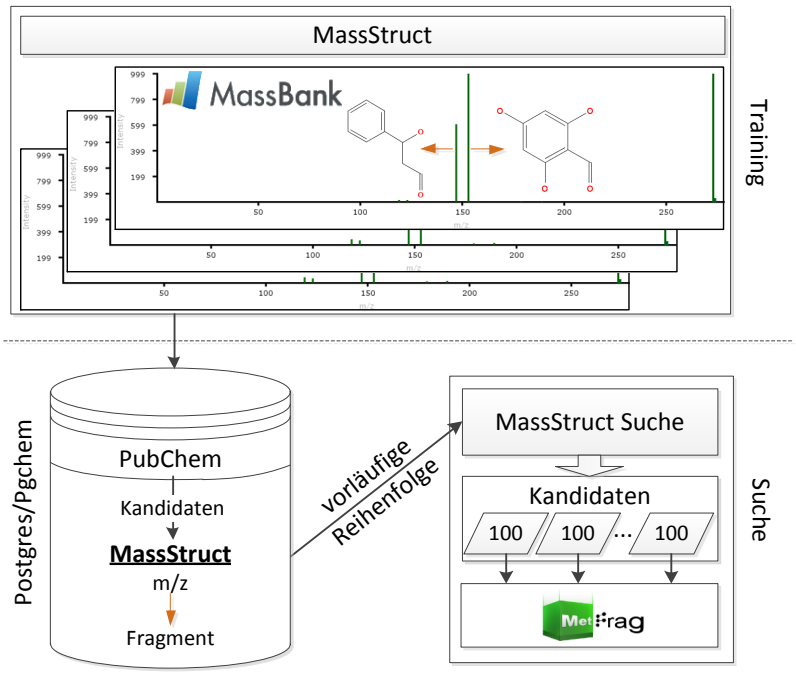


Abbildung 3.12. Workflow von MassStruct: In der oberen Abbildung ist dargestellt, wie Assoziationen gelernt und in einer Datenbank gespeichert werden. Im unteren Teil ist die Arbeitsphase zu sehen, die die vorher gelernten Informationen nutzt, um die Reihenfolge der zurückgelieferten Kandidaten bestimmen zu können. (nach [HWN11])

ist üblicherweise nicht in der Lage, jeden Peak mit einer Substruktur zu annotieren und es kann vorkommen, dass unterschiedliche Fragmente einem Peak zugeordnet worden sind. In letzteren Fall werden beide Substrukturen gespeichert.

Die MassStruct Datenbank wird bei der Suche nach passenden Kandidaten wie folgt benutzt: Peaks der gemessenen unbekannten Verbindung werden verwendet, um bereits gelernte Substrukturen aus der Trainingsdatenbank zu erhalten. Für alle Treffer aus einer Molekül datenbank, z.B. PubChem, wird dann getestet, ob die Substruktur in diesem Kandidaten enthalten ist und ggf. der Score für diesen Kandidaten um eins erhöht. Abbildung 3.14 zeigt dieses Vorgehen für ein Spektrum von Epicatechin für zwei bereits gelernte Substrukturen. Die orange (165,0483 m/z), grün (249,0667 m/z) und blau (207,0577 Da) markierten Peaks sind bereits im Vorverarbeitungsschritt gelernt worden und können somit für eine Substruktursuche in der Molekül datenbank verwendet werden. Kandidat 1 wird als erstes mit einem Score

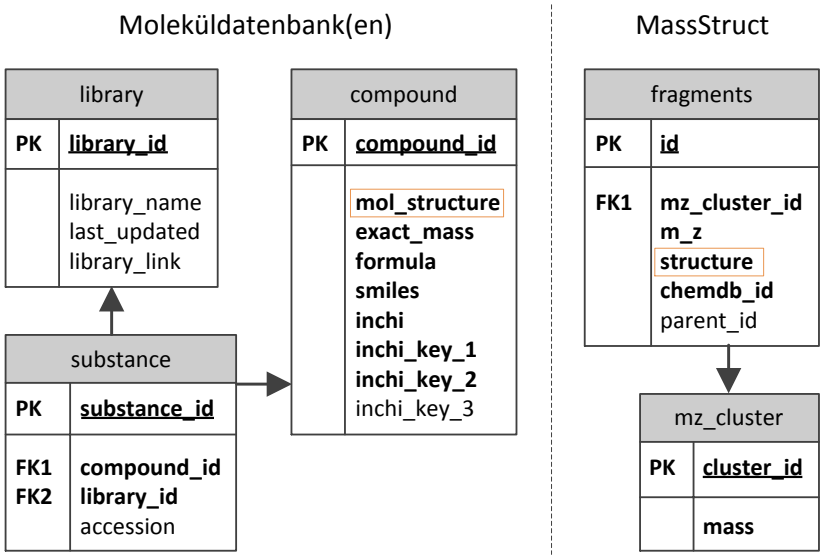


Abbildung 3.13. Datenbankschema von MassStruct: Im linken Teil sind die Tabellen abgebildet, die Strukturen aus Moleküldatenbanken gespeichert haben. MassStruct benutzt zwei Tabellen zum Speichern der Peak-Fragment Information, die während des Trainings gelernt werden. Die hervorgehobenen Tabellenspalten sind vom Typ `molecule` und speichern somit die Strukturinformationen der Moleküle bzw. Fragmente. (nach [HWN11])

von 2 zurückgeliefert, da dieser die Substrukturen von den Fragmenten mit Masse 165,0483 Da und 207,0577 Da enthält.

Die SQL Abfrage in Abbildung 3.15 zeigt, wie MassStruct die Kandidaten bewertet und nach Score geordnet zurückgibt. Die Kandidatensuche selektiert alle Verbindungen, die sich innerhalb einer erlaubten Abweichung, die abhängig vom verwendeten Massenspektrometer ist, befinden. Das von pgchem bereitgestellte Prädikat `<=` ist ein Substruktur Operator der überprüft, ob `fragment.structure` ein Subgraph von `compound.structure` ist. Der resultierende `score` eines Kandidaten (`accession`) berechnet sich aus der Anzahl der gefundenen Substrukturen. Eine solche Substruktursuche vergleicht in einem ersten Schritt mit Hilfe eines „Generalized Search Tree index“ (GiST) [HNP95] die chemischen Fingerprints zwischen Fragment- und Kandidatenstruktur. Erst Anschließend wird auf den ausgewählten Strukturen eine zeitaufwendige Substruktursuche durchgeführt.

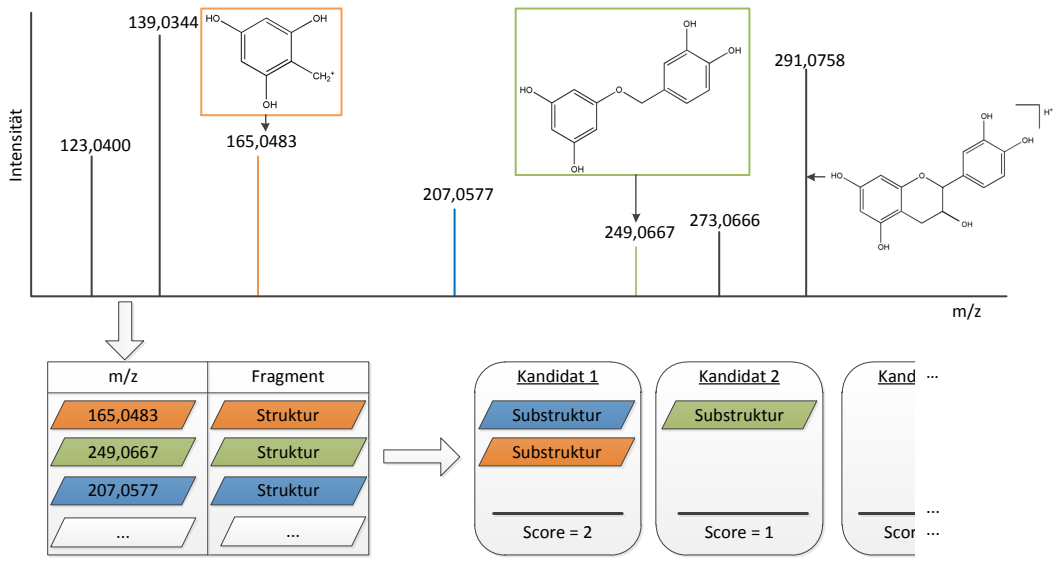


Abbildung 3.14. MassStruct Scoring am Beispiel des Spektrums von Epicatechin (Abbildung 2.4). Zu den Peaks passende Fragmentstrukturen, die in der Datenbank gespeichert sind, werden den gemessenen Peaks zugeordnet. Eine nachfolgende Substruktursuche liefert die Kandidaten aus PubChem oder KEGG als erstes zurück, deren Struktur die meisten Substrukturen beinhaltet. (nach [HWN11])

3.4 Zusammenfassung

MetFrag beinhaltet einen kombinatorischen Fragmentierungsalgorithmus, der alle möglichen Fragmente der Kandidatenstrukturen, die aus einer Molekülдatenbank mit Summenformel oder passender Masse gefunden worden sind, generieren kann. In einem Vorverarbeitungsschritt werden die Bindungen entsprechend ihrer Stärke annotiert. Durch die Begrenzung der Baumtiefe und der Verwendung einer Summenformel Redundanzüberprüfung kann die Laufzeit von MetFrag niedrig gehalten werden, sodass tausende Kandidaten prozessiert werden können. Die Fragmentionen eines MS/MS Spektrums von Naringenin sind mit Hilfe von MetFrag identifiziert und mit annotierten Fragmenten verglichen worden. Die entwickelte Scoring Funktion verwendet die Bindungsordnung aus den vorverarbeiteten Kandidaten und erstellt eine Rangfolge dieser. Kandidaten mit dem gleichen Score und einer Tanimoto Ähnlichkeit $> 0,95$ werden zusammengefasst. Neben der Weboberfläche ist auch die intelligente Kandidatensuche MassStruct vorgestellt worden. Diese ist in der Lage

```
SELECT accession, count(fragment.id) AS score
FROM compound, fragment
WHERE compound.mass BETWEEN 290.2 AND 290.3
AND (  fragment.mass BETWEEN 123.0 AND 123.1
    OR fragment.mass BETWEEN 139.0 AND 139.1
    OR fragment.mass BETWEEN 165.0 AND 165.1
    OR fragment.mass BETWEEN 207.0 AND 207.1
    OR fragment.mass BETWEEN 249.0 AND 249.1
    OR fragment.mass BETWEEN 273.0 AND 273.1)
AND fragment.structure <= compound.structure
GROUP BY accession
ORDER BY score;
```

Abbildung 3.15. Vereinfachte SQL Abfrage von MassStruct an die Postgres Datenbank. Es werden Kandidaten selektiert, die eine Molekülmasse (`compound.mass`) zwischen `290.2` Da und `290.3` Da besitzen. Außerdem werden `fragment` Strukturen gesucht, die den Peakmassen (mit gewisser Abweichung) entsprechen, und auch Substruktur (`<=`) des Kandidaten `compound.structure` sind. Letztendlich wird das Ergebnis nach dem `score`, die Anzahl der enthaltenen Substrukturen in den Kandidaten, absteigend sortiert. (nach [HWN11])

aus gelernten Peak Fragment Assoziationen, geeignete Kandidaten als erstes zurückzuliefern, die am wahrscheinlichsten die Peaks des Anfragespektrums erklären.

4 Evaluierung und Optimierung von MetFrag und MassStruct

Im folgenden Kapitel werden die Experimente vorgestellt, die im Rahmen der Arbeit durchgeführt worden sind. Als erstes wird ein geeignetes Maß zur Angabe der Bindungsstärke und zur Auswertung der Ergebnisse von Software zur Fragmentvorhersage ausgewählt. Es wird beschrieben, wie die Scoring Funktion optimiert worden ist. Im darauffolgenden Abschnitt wird MetFrag mit hochauflösenden MS/MS Spektren von [HKF+08] evaluiert und mit der kommerziellen Software MassFrontier verglichen. Weiterhin werden GC/EI-MS Daten (Nominalmassen) von [SMB09] ausgewertet und mit ähnlicher Software verglichen. Am Ende des Kapitels wird die Evaluierung von MassStruct und die weiteren Anwendungen von MetFrag vorgestellt.

4.1 Methodenauswahl zur Vorverarbeitung

Eine geeignete Methode zur Annotation der Bindungsgewichte wird im folgenden Abschnitt beschrieben. Es ist von [AHP+09] gezeigt worden, dass durch den Vergleich der Bindungslänge zwischen neutralen und protonierten Molekül (jeweils mit DFT strukturoptimiert) die Fragmente von ESI-MS/MS Spektren vorhersagt werden können.

4.1.1 Auswahl eines geeigneten Kraftfeldes

Abschnitt 3.1.2 beschreibt die Verwendung des „Universal Force Field“ (UFF) zur Bestimmung einer ersten Näherung der 3D-Struktur eines Kandidaten. Ghemical

und MMFF94 sind neben dem UFF auch in OpenBabel 2.3.0 implementiert und somit auch als Kraftfelder in Frage gekommen. Um die Zuverlässigkeit aller drei Varianten zu evaluieren, sind zum einen die Strukturen von 1000 zufällig gezogenen Spektren [KMR06] mit den jeweiligen Kraftfeldern und danach mit MOPAC (verwendete Parameter aus Abschnitt 3.1.2) strukturoptimiert worden. Zum anderen sind alle drei Kraftfelder mit dem MMFF94 Validierungsdatensatz[26] (698 Moleküle) [Hal99] in gleicher Weise ausgewertet worden.

Tabelle 4.1. Zuverlässigkeit der in OpenBabel 2.3.0 implementierten Kraftfelder UFF, Ghemical und MMFF94. Zur Evaluierung sind die Strukturen von 1000 zufällig gezogenen NIST Spektren [KMR06] und der MMF94 Validierungsdatensatz [Hal99] mit dem jeweiligen Kraftfeld strukturoptimiert worden. Die Tabelle zeigt die fehlgeschlagenen (Programmabsturz bzw. Standardbildungsenthalpie $< -10\,000$ oder $> 100\,000$) Optimierungen des jeweiligen Kraftfeldes in Kombination mit MOPAC. ([SGK+])

	UFF + MOPAC	Ghemical + MOPAC	MMFF94 + MOPAC
1000 Strukturen	5	0	0
MMFF94 Validierungsdaten	0	51	50

Tabelle 4.1 zeigt das Ergebnis dieses Experimentes. Die Kombination aus UFF und MOPAC hat insgesamt die wenigsten Fehler produziert und wird daher für die Vorverarbeitung der Moleküle verwendet. Die Zuverlässigkeit der Vorverarbeitung ist für den Workflow von MetFrag von Bedeutung, da alle weiteren Schritte darauf aufbauen.

4.1.2 Auswahl eines Maßes zur Bestimmung der Bindungsstärke

Neben MOPAC kann auch die Dichtefunktionaltheorie (DFT) (ab-initio Methode - siehe Abschnitt 2.5.2) zur Strukturoptimierung von Molekülen verwendet werden. Die (ab-initio) DFT Berechnungen von [AHP+09] werden als Referenzwerte für die folgende Evaluation verwendet. Dieser Abschnitt untersucht, ob die Bindungsordnung oder Bindungslängenänderung als ein Maß für die Bindungsstärke geeignet sind.

[26] `http://www.ccl.net/cca/data/MMFF94/` - Abgerufen im Dezember 2011

Die Bindungen der drei Teststrukturen aus [AHP+09] (PubChem CID: 20097272, 3365 und 5231054) sind mit der in Abschnitt 3.1.2 beschriebenen Vorverarbeitung annotiert worden. Abbildung 4.1 zeigt einen Vergleich der MetFrag Vorverarbeitung mit MOPAC und der in [AHP+09] verwendeten DFT am Beispiel des Moleküls mit PubChem CID: 20097272. Sowohl bei der Verwendung von DFT, als auch MOPAC hat die Bindungslänge bei den drei annotierten Bindungen am meisten zugenommen. [AHP+09] hat festgestellt, dass diese leicht im Massenspektrometer fragmentieren und dadurch entsprechende Fragmentionen im Spektrum als Peak sichtbar sind. Das intensivste Fragment des gemessenen Spektrums (Abbildung 4.1 unten) entsteht durch Abspaltung einer C_3 Gruppe, welche durch die Bindungslängenänderung vorhergesagt worden ist.

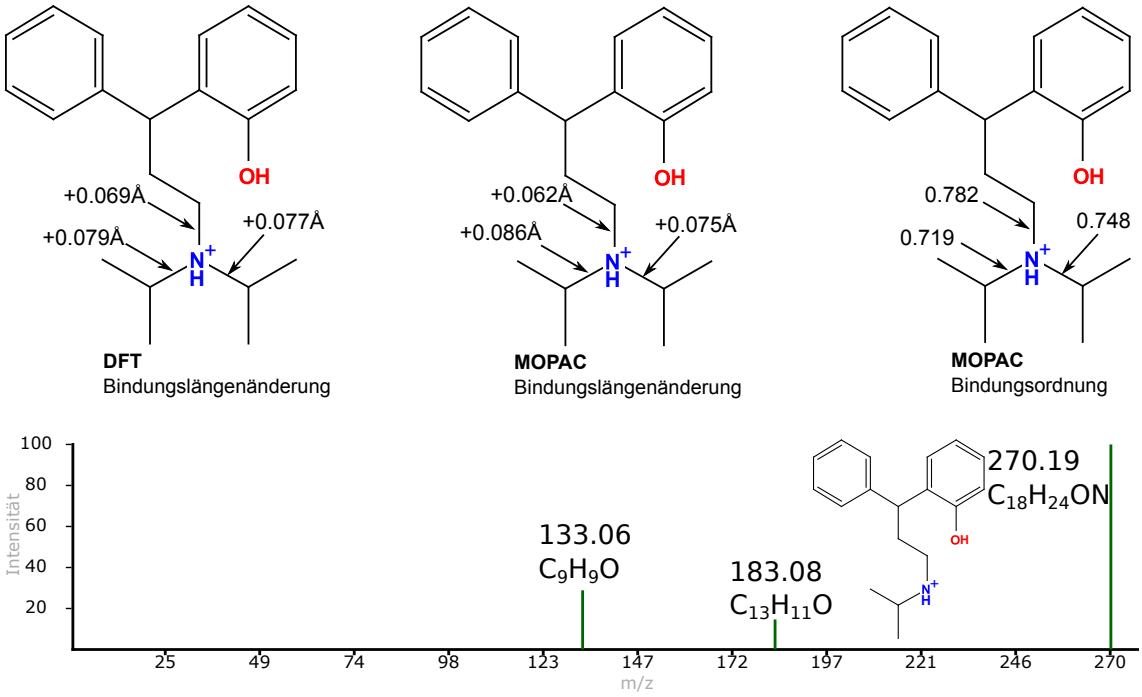


Abbildung 4.1. Vergleich der größten Bindungslängenänderung bzw. kleinsten Bindungsordnung des neutralen und mit dem am Stickstoff protonierten Molekül. Die DFT Berechnung von [AHP+09] (und anschließender Messung im Massenspektrometer) sind vergleichbar mit denen von MOPAC, da alle drei für die Abspaltung von drei Kohlenstoffe sprechen, das auch das intensivste Fragment (270,19 Da) darstellt.

MetFrag verwendet im Gegensatz zu [AHP+09] neben der Bindungslängenänderung auch die Bindungsordnungen der stabilsten Protonierung. Diese wird für jede Bindung mit Hilfe von MOPAC im Molekül bestimmt. Je geringer die Bindungsordnung,

desto wahrscheinlicher bricht diese Bindung. Abbildung 4.1 (rechts) zeigt die kleinsten Bindungsordnungen des Moleküls. Auch aus diesen Daten kann geschlussfolgert werden, dass die Kohlenstoffbindung mit der geringsten Bindungsordnung bricht. MetFrag verwendet kein DFT, da diese (ab initio) Methode für tausende Moleküle mit mehreren Protonierungsstellen zu zeitaufwendig ist.

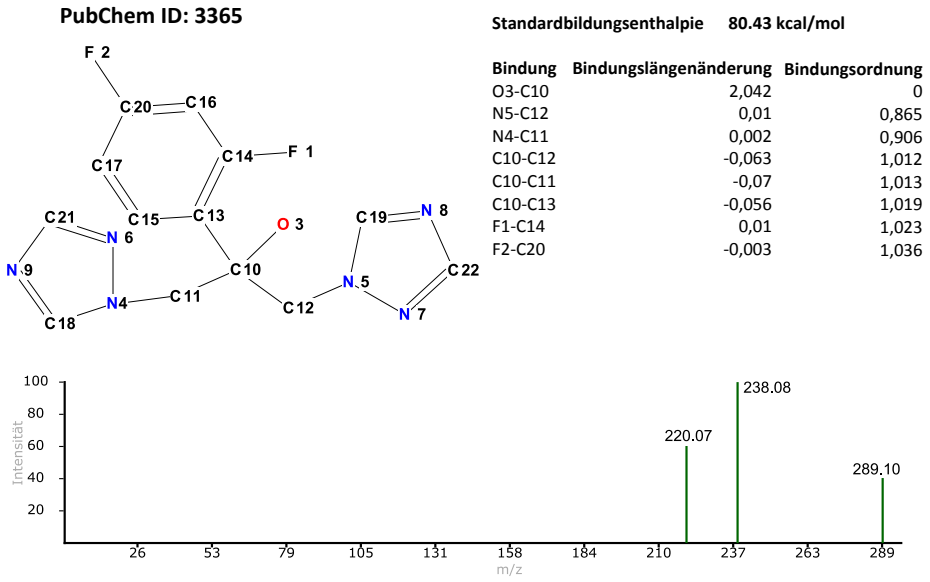


Abbildung 4.2. Spektren und Fragmentionen von [AHP$^+$09] - **ID 3365:** Geringste Standardbildungsenthalpie (80,43 kcal/mol) bei der Protonierung am Sauerstoff. Die Fragmente lassen sich durch die Spaltung der Bindung O3-C10 (Peak: 289,1004 m/z), N5-C12 oder N4-C11 (Peak: 238,0782 m/z) und beide Abspaltungen zusammen (Peak: 220,0677) erklären.

Im folgenden sind die beiden anderen Verbindungen aus [AHP$^+$09] mit PubChem CID: 3365 und 5231054 unter Benutzung der Bindungsordnung und Bindungslängenänderung untersucht worden. Auch in diesem Fall sind die beiden Moleküle vorverarbeitet worden (Abschnitt 3.1.2). Im Gegensatz zu [AHP$^+$09] verwendet MetFrag nur die stabilste Protonierung, um die Bindungsordnungen zu bestimmen. Abbildung 4.2 zeigt die Struktur, sowie dessen annotierten Bindungen. Die Verbindung mit PubChem CID: 3365 besitzt die stabilste Protonierung am O3 mit einer Standardbildungsenthalpie von 80,43 kcal/mol. Die im Massenspektrometer intensivsten Peaks lassen sich durch die Spaltung der Bindung O3-C10 (289,1004 m/z), N5-C12

oder N4-C11 (238,0782 m/z) und beiden Bindungsbrüchen zusammen (220,0677 m/z) erklären. Alle drei Bindungen besitzen die mit Abstand kleinste Bindungsordnung, wobei auch die Bindungslänge zwischen neutralem und protonierten Molekül zugenommen hat. Jedoch verlängert sich die Bindung von F1-C14 genau wie N5-C12 um 0,01 Å. Dadurch ist keine genaue Vorhersage der Bindungsbrüche mit der Bindungslängenänderung möglich.

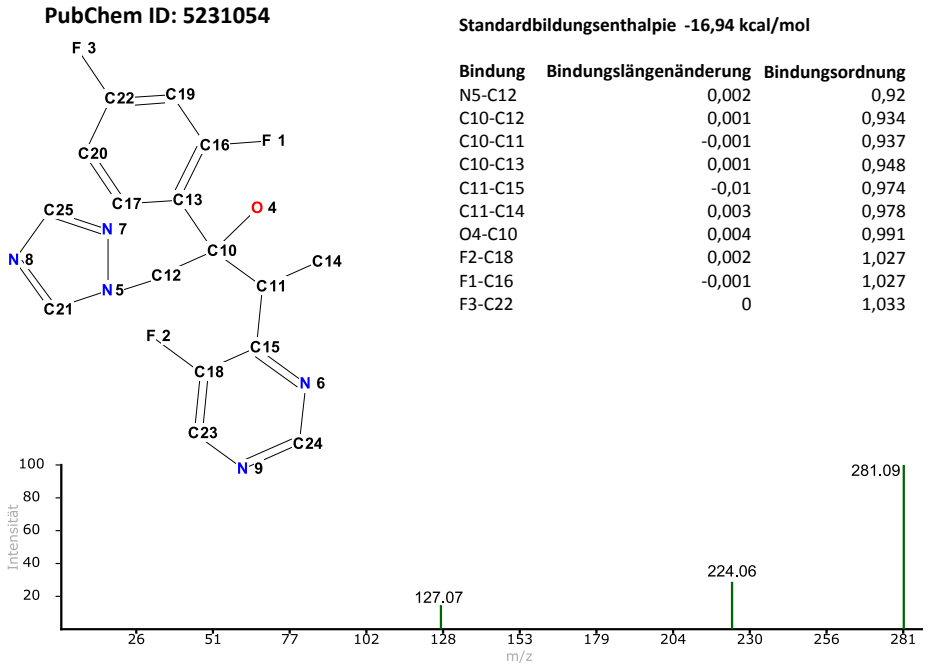


Abbildung 4.3. Spektren und Fragmentionen von [AHP$^+$09] - **ID 5231054:** Geringste Standardbildungsenthalpie bei der Protonierung von N6. Der Peak 281,09 m/z kann erklärt werden durch Auftrennen der Verbindung N5-C12. Die Peaks 224,06 m/z und 127,067 m/z können durch die Spaltung der Bindung C10-C12 erklärt werden.

Das gleiche Prinzip ist auch bei der Verbindung mit PubChem CID: 5231054 angewendet worden (Abbildung 4.3). Die stabilste Protonierung, mit einer Standardbildungsenthalpie von -16,94 kcal/mol, ist am Atom N6. Auch in diesem Beispiel werden die intensivsten Peaks durch die Spaltung der Bindungen mit der geringsten Bindungsordnung erklärt: Durch die Aufspaltung der Bindung N5-C12 kann der Peak 281,09 m/z erklärt werden. Die beiden anderen Peaks (224,0632 m/z und 127,0666 m/z) können durch die Aufspaltung der Bindung C10-C12 zugeordnet werden, die

die Verbindung in zwei Fragmente teilt. Auch in diesem Beispiel gelingt es nicht mit der MetFrag Vorverarbeitung und unter Benutzung der Bindungslängenänderung eindeutig vorherzusagen welche Bindung bricht, da die Änderung mit 0,004 Å bei O4-C10 am größten ist.

Um herauszufinden, ob die Bindungsordnung oder die Bindungslängenänderung am besten für die Scoring Funktion geeignet ist, ist folgendes Experiment durchgeführt worden: Für alle 102 Verbindungen aus [HKF+08] sind die Strukturen der richtigen Kandidaten vorverarbeitet und mit MetFrag (Baumtiefe 1) fragmentiert (Abschnitt 3.1.3) und den Peaks zugeordnet (Abschnitt 3.1.4) worden. Für jedes Fragment, das einen Peak erklärt, ist die Bindungslängenänderung bzw. Bindungsordnung der gebrochenen Bindung gespeichert worden. Zusätzlich werden alle Fragmente, die kein Peak erklären, mit den entsprechenden Werten für die Bindungslängenänderung bzw. Bindungsordnung gespeichert.

Abbildung 4.4 zeigt das kumulative Histogramm über die Bindungslängenänderung und die Bindungsordnung. Die rote Kurve zeigt die Werte für Fragmente, die keinen Peak erklärt haben. Werte der Bindungsordnung bzw. Bindungslängenänderung für Fragmentionen, die durch MetFrag erklärt werden konnten, sind in der grünen Kurve dargestellt.

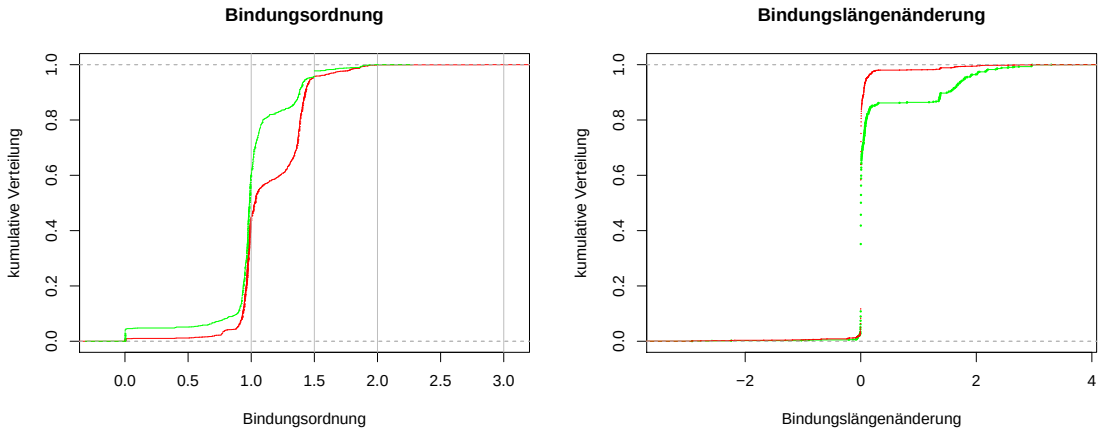


Abbildung 4.4. Die Bindungsordnung und Bindungslängenänderung der Bindung(en) die entfernt worden sind, um ein Fragment zu erklären (grüne Kurve) bzw. kein Fragment erklärt haben (rote Kurve) sind dargestellt. Die Fragmentierung und Zuordnung der Fragmente ist mit MetFrag und den Daten von [HKF+08] durchgeführt worden. Dabei ist nur der richtige Kandidat prozessiert worden.

Für die Bindungsordnung ist ein Unterschied im Anstieg zwischen der grünen und roten Kurve in einem Bereich von 0 und 0,5 sowie zwischen 0,8 und 1,2 erkennbar. Bindungen, die eine solche Bindungsordnung besitzen erklären häufiger einen gemessenen Peak. Die grauen Linien zeigen die Bindungsordnung einer Einfachbindung ($\approx 1,0$), einer aromatischen Bindung ($\approx 1,5$) bzw. Doppelbindung ($\approx 2,0$) an. Ein sprunghafter Anstieg ist bei Bindungsordnung 1,5 zu erkennen, das durch die Festlegung der aromatischen Bindung auf diesen Wert zustande kommt. Abbildung 4.4 (rechts) zeigt die entsprechenden Werte für die Bindungslängenänderung. Hier ist ein Unterschied zwischen 1,5 und 2,5 Å zu erkennen. Diese Werte für die Bindungslängenänderung erscheinen allerdings recht hoch und deuten auf eine nicht ausreichende Optimierung des Moleküls hin. Außerdem ist bei diesem Experiment festgestellt worden, dass die Bindungsordnungen wesentlich weniger schwanken. Daher wird in der Scoring Funktion von MetFrag (siehe Abschnitt 3.1.6) die Bindungsordnung verwendet.

4.2 Maße zur Bestimmung der Rangordnung

Die Ergebnisse von *in silico* Fragmentierungssoftware können nicht alleine durch den Rang der korrekten Lösung in der sortierten Ergebnisliste interpretiert werden, da dieser stark von der Anzahl der (ähnlichen) Kandidaten abhängt. Dies kann bei dem Vergleich der Ergebnisse von MetFrag unter der Verwendung von PubChem aus dem Jahr 2006 und 2009 beobachtet werden, da sich hier der Rang der richtigen Verbindung verschlechtert, je mehr (ähnliche) Kandidaten verwendet werden (siehe Abschnitt 4.6.1 und 4.6.2). Weiterhin macht es einen Unterschied, ob die richtige Verbindung auf Platz 50 gefunden wurde und 100 oder 10000 Kandidaten bearbeitet worden sind. Daher wurde von [KMR06] die „relative ranking position“ (RRP) eingeführt, die das Verhältnis zwischen Rang und Anzahl der Kandidaten beschreibt:

$$RRP = 0,5 \left(1 + \frac{BC - WC}{TC - 1}\right) \ . \tag{4.1}$$

BC („better candidates“) steht für die Anzahl an besseren Kandidaten, WC („worse candidates“) - Anzahl von schlechteren Kandidaten und TC („total candidates“) für die Gesamtzahl der prozessierten Kandidaten. Ein RRP von 0 bedeutet, dass die richtige Struktur Rang 1 besitzt und RRP 1 ist der schlechteste Fall. In der folgenden Auswertung wird der RRP immer über den Rang berechnet und nicht über den Tanimoto Cluster Rang, da in die Berechnung die Anzahl der Kandidaten mit einbezogen wird und durch das Clustern diese verzerrt werden würde. Außerdem kann MetFrag dadurch mit ähnlicher Software verglichen werden, die kein Zusammenfassen der Strukturen ermöglicht.

Der RRP sagt trotzdem nicht viel darüber aus, wie brauchbar das Ergebnis für einen Anwender wirklich ist. Der RRP $0,02$ wird für eine Verbindung auf Rang 200 zugeordnet, wenn insgesamt 10000 Kandidaten bearbeitet worden sind. Obwohl der RRP sehr niedrig ist, ist es fraglich, dass wirklich alle Strukturen bis zum Rang 200 angeschaut werden. Daher wird in den folgenden Experimenten neben dem RRP auch der Median des Ranges der korrekten Lösung in der sortierten Ergebnisliste mit angegeben. Obwohl es sich bei einem Rang um ein ordinalskaliertes Merkmal handelt, wird das arithmetische Mittel angegeben, damit veranschaulicht werden kann, auf welchen Platz sich der richtige Kandidat im Durchschnitt befindet.

4.3 Test- und Trainingsdaten

Um verschiedene Tests durchzuführen, Parameter zu schätzen und Vergleiche mit ähnlicher Software zu erstellen, sind verschiedene Datensätze verwendet worden: 510 hochauflösende Spektren von [HKF+08] sind mit einem MicroMass II MS/MS (positiver Modus) mit fünf unterschiedlichen Kollisionsenergien (10, 20, 30, 40 und 50 eV) gemessen worden. Die gemessenen Verbindungen haben eine durchschnittliche Masse von 372,48 Da (Abbildung 4.5) und decken einen Bereich zwischen 137 Da und 609 Da ab. Das Massenspektrometer hat eine gute Genauigkeit, weshalb mzppm auf 10 und mzabs auf 0,0 festgelegt worden sind, wie es auch von [HKF+08] vorgenommen wurde. Da alle Spektren im positiven Modus von einem $[M+H]^+$ Vorläuferion aufgenommen worden sind, ist auch in MetFrag die entsprechende Option gewählt worden (siehe Tabelle 3.1). In der folgenden Evaluierung werden diese Parameter in

MetFrag verwendet, sofern es nicht explizit anders angegeben wird. Im Gegensatz zu [HKF+08] sind alle Messungen einer Verbindung mit unterschiedlichen Kollisionsenergien zu einem Spektrum zusammengefasst worden. Dadurch werden alle zur Verfügung stehenden Peaks genutzt. Tabelle A.1 zeigt alle 102 Verbindungen mit den dazugehörigen MassBank IDs.

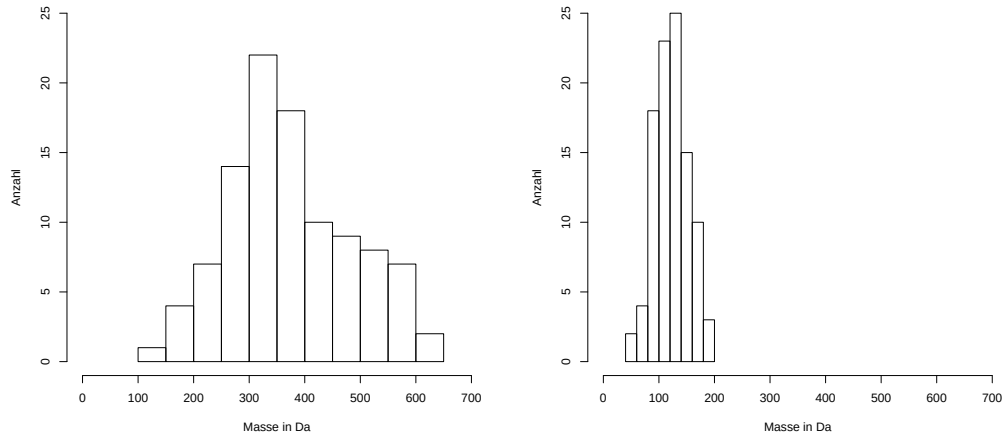


Abbildung 4.5. Histogramm über die Masse der Analyten aus dem MS/MS Datensatz von [HKF+08] (links) und den GC/MS Daten von [SMB09] (rechts).

MetFrag ist auch mit GC/EI-MS Daten von [SMB09] evaluiert worden. Dieser Testdatensatz enthält 100 Spektren (Nominalmassen) aus der NIST '05, wobei die Kandidaten mit MOLGEN 3,5 generiert worden sind. Die gemessenen Analyten haben eine durchschnittliche Masse von 123,19 Da (Abbildung 4.5 rechts) und reichen von 56 Da bis 188 Da. Dieser Datensatz enthält somit wesentlich kleinere und leichtere Moleküle als die ESI-MS/MS Daten.

Neben den beiden beschriebenen Datensätzen zur Evaluierung von MetFrag, ist ein dritter Datensatz verwendet worden. Dieser enthält 240 Spektren vom Riken Plant Science Center, die mit einem ESI-MS/MS Massenspektrometer gemessen wurden. Es handelt sich um „Ramp" Spektren, d.h. die Kollisionsenergie wurde während der Messung erhöht, um möglichst viele Fragmente zu messen. Tabelle A.8 enthält alle MassBank IDs der verwendeten Spektren. Die Daten decken einen Bereich zwischen 73,05 Da und 837,16 Da ab und die durchschnittliche Masse beträgt 289,22 Da (Abbildung 4.6).

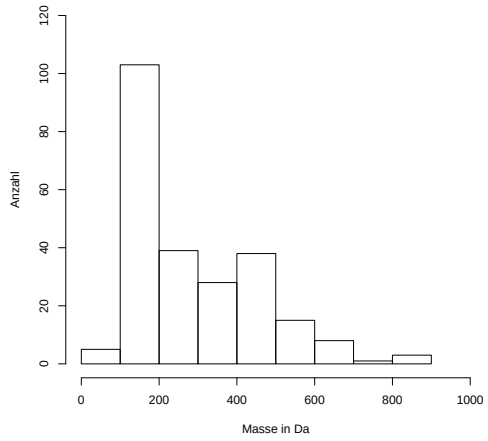


Abbildung 4.6. Histogramm über die Masse der Analyten der MS/MS Daten vom RIKEN Plant Science Center.

4.4 Theoretische und empirische Laufzeitanalyse

Im folgenden Abschnitt wird die theoretische Laufzeit, die von MetFrag zum Fragmentieren eines Moleküls gebraucht wird, abgeschätzt. Die Zeit zum Zuordnen der Peaks hängt sehr stark von dem untersuchten Spektrum und dessen Peaks (Anzahl und Masse) ab und wird daher nicht betrachtet. Ebenso ist die Laufzeit für die Vorverarbeitung, die hauptsächlich von MOPAC abhängt, nicht Bestandteil dieser Arbeit. Jedoch ist die Gesamtlaufzeit von MetFrag für den Datensatz von [HKF$^+$08] mit den Kandidaten aus PubChem 2009 in Tabelle A.2 angegeben. Für die Evaluierung ist ein virtualisierter Linux Rechner mit 8 Kernen und 16 GB RAM verwendet worden. Die virtuellen Maschinen laufen auf einen Server mit zwei 12 Kern AMD Opteron (2300 MHz) und 98GB RAM.

Für die Fragmentierung eines einzelnen Kandidaten ist die Anzahl der Ringe c, die Anzahl von Bindungen im größten Ring l, die Anzahl von Bindungen im Ring r und die Gesamtzahl der Bindungen im Molekül b von Bedeutung. Insgesamt können maximal 2^b Fragmente aus einem Molekül generiert werden. Außerdem wird der Molekülgraphen traversiert, was durch $|b|$ beschränkt ist. Die Gesamtlaufzeit, um alle Fragmente möglichen Fragmente zu generieren, ist in Formel 4.2 dargestellt.

$$\mathcal{O}\left(b \cdot 2^b\right) \tag{4.2}$$

Da MetFrag üblicherweise alle Fragmente bis zu einer gewissen Baumtiefe k berechnet, wird diese im Folgenden betrachtet. Hierbei muss zwischen linearen und zyklischen Bindungen unterschieden werden. Zwei neue Fragmente entstehen, wenn eine lineare Bindung gebrochen wird und somit können maximal $2b$ Fragmente pro Baumtiefe generiert werden, was insgesamt $2b^k$ Fragmente ergibt. Wenn eine Bindung Teil eines Ringsystems ist, wird jede Kombination der aktuellen Bindung und einer weiteren entfernt, sodass pro Kombination maximal zwei neue Fragmente entstehen. Somit werden $\frac{r \cdot (r-1)}{2}$ Bindungen gebrochen und maximal $r \cdot (r-1)$ Fragmente generiert. Die Laufzeit in Abhängigkeit der Baumtiefe k ist in Formel 4.3 angegeben.

$$\mathcal{O}\left(b \cdot ((l \cdot (l-1))^c \cdot 2b)^k\right) \tag{4.3}$$

Zusätzlich zur theoretischen ist auch die praktische Laufzeit von MetFrag untersucht worden. Die folgenden Experimente sind auf einen PC mit Intel Q9400 CPU (2,66Ghz) und 8Gb RAM unter Ubuntu 10.04 (x64) durchgeführt worden. Zur Bestimmung der Laufzeit sind zufällig Moleküle gezogen und fragmentiert worden. Die Masse der Moleküle ist in PubChem nicht gleichverteilt. Daher ist der Bereich zwischen 100 und 1000 Da in 90 gleichgroße Teile geteilt worden. Zum Beispiel sind zwischen 100 Da und 200 Da ein Abschnitt pro 10 Da erstellt worden, der jeweils 500 Moleküle enthält. Dadurch wird gewährleistet, das pro Massenbereich gleich viele Verbindungen bearbeitet werden. Insgesamt sind auf diese Weise 45 000 Moleküle zwischen 100 und 1000 Da mit Baumtiefe 1 und 2 prozessiert worden. Abbildung 4.7 zeigt die Laufzeit von MetFrag, die für die Fragmentierung der Moleküle benötigt wurde.

Zur besseren Übersicht ist die Ordinate logarithmiert, wodurch der exponentielle Anstieg der Laufzeit nicht direkt sichtbar ist. Rote bzw. schwarze Punkte beschreiben die Laufzeit der Moleküle, die durch MetFrag benötigt wurde, um alle Fragmente bis Baumtiefe 1 bzw. 2 zu berechnen. Die dunkelblaue und hellblaue Linie gibt die nichtlineare Regression der jeweiligen Baumtiefe an. Je schwerer, dadurch in der Regel auch größer und verzweigter, das Molekül ist, desto mehr Zeit wird für die Fragmentierung benötigt. In der theoretischen Betrachtung ist bereits festgestellt worden, dass Strukturen mit großen Ringsystemen die Laufzeit verschlechtern. Ab-

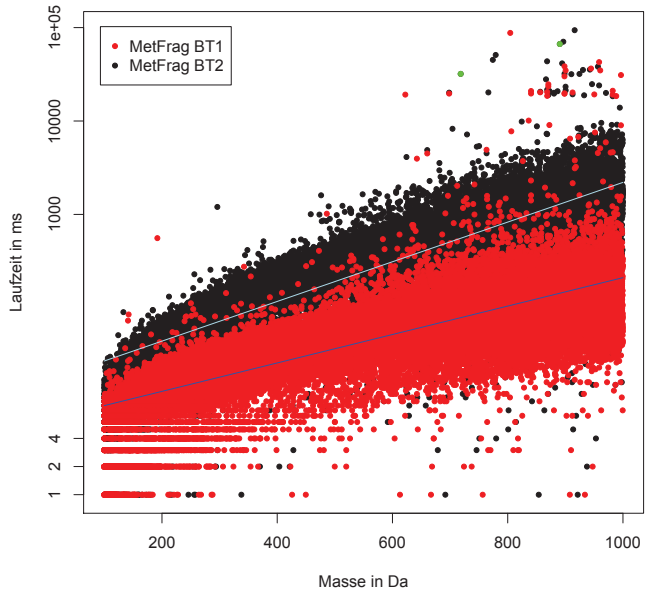


Abbildung 4.7. Laufzeit mit zufällig gezogenen Molekülen mit einer Masse zwischen 100 Da bis 1000 Da mit einer Baumtiefe (BT) 1 und 2. Pro 100 Da sind 10 Teile mit je 500 Molekülen erstellt worden, sodass der gesamte Bereich mit gleich vielen Molekülen abgedeckt ist. Abbildung 4.7 zeigt die Strukturen der beiden grün markierten Punkte, die besonders lange für die Fragmentierung gebraucht haben.

bildung 4.8 zeigt die Strukturen der grün markierten Ausreißer, die aus vielen Ringen aufgebaut sind. Es gibt eine ganze Reihe von Molekülen bzw. Elemente die nur 1 ms bzw. 2 ms zur Fragmentierung benötigten. Dies sind vor allem einzelne Elemente, wie zum Beispiel Cd oder Xe, bei denen keine Bindung gebrochen werden muss, aber alleine schwerer als 100 Da sind.

4.5 Parameteroptimierung der Scoring Funktion

Für die Parameter a, b und c der Scoring Funktion (Gleichung 3.1), ist eine Parameteroptimierung durchgeführt worden. Als Grundlage der Optimierung dient der 510 Spektren umfassende Datensatz von [HKF+08] mit den PubChem Kandidaten aus Juni 2009. Für diese zeitaufwendige Optimierung wird der Bergsteigeralgorithmus [NR03] verwendet, da dieser schnell das Abbruchkriterium erreicht und somit ein lokales Optimum gefunden hat. Konkret minimiert der Bergsteigeralgo-

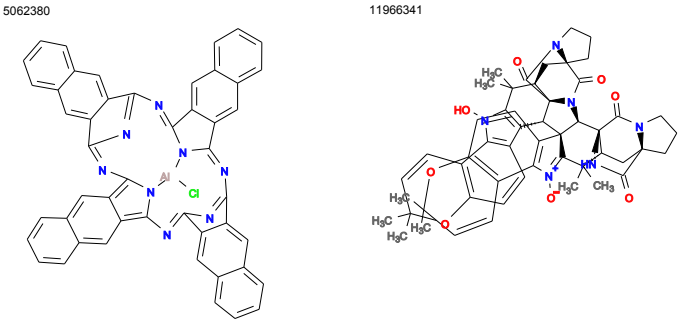


Abbildung 4.8. CID: 5062380 und 11966341 sind Beispiele für Moleküle, die mit Baumtiefe 2 besonders lang für die Berechnung der Fragmente gebraucht haben. CID: 5062380 (Summenformel: $C_{48}H_{24}AlClN_8$) besitzt eine monoisotopische Masse von 774,163 Da und braucht für die Fragmentierung 45 Sekunden. 67s ist für CID: 11966341 (Summenformel: $C_{52}H_{54}N_6O_8$) mit einer monoisotopischen Masse von 890,400 Da benötigt worden.

rithmus die Zielfunktion $f(\mathbf{x})$, mit $\mathbf{x} = (a, b, c)$ und f berechnet pro Iteration den durchschnittlichen RRP über alle Testspektren entsprechend der Scoring Funktion in Gleichung 3.1. Das heißt pro Iteration werden 102 zusammengefügte Spektren mit durchschnittlich 2544 Kandidaten und den jeweiligen Parametersatz evaluiert. Nach der Initialisierung werden die Nachbarn berechnet. Diese werden durch Addition und Subtraktion mit einer randomisierten Zahl zwischen 0 und 1 zu den vorherigen Parametern berechnet. Danach bildet der beste Nachbar, d.h. der mit dem geringsten durchschnittlichen RRP über alle Testspektren, den Ausgangspunkt für eine erneute Berechnung aller Nachbarn. Das Abbruchkriterium ist dann erreicht, wenn keine Verbesserung mehr gefunden werden kann und somit ein lokales Optimum erreicht wurde. Damit bessere Lösungen gefunden werden, wird der Bergsteigeralgorithmus mehrfach mit zufälligen Initialisierungen gestartet und auch die Schrittweite, mit der die Parameter variiert werden, ist randomisiert zwischen 0 und 1 gewählt.

Es besteht die Gefahr die Trainingsdaten auswendig gelernt zu haben („Overfitting“). Dies soll vermieden werden, da sonst auf anderen Datensätzen kein gutes Ergebnis erreicht werden kann. Zur Überprüfung ist eine zehnfache Kreuzvalidierung durchgeführt worden. Dabei wird der Datensatz zufällig in Trainings- (90%) und Testdaten (10%) aufgeteilt. Auf diese Weise werden 10 verschiedene (randomisierte) Partitionen, die jeweils aus einem Trainings- und Testdatensatz bestehen, erstellt. Auf jeder dieser Partitionen wird ein (lokaler) optimaler Parametersatz auf den Trainingsdaten mit Hilfe des Bergsteigeralgorithmus bestimmt. Mit den opti-

Tabelle 4.2. Testfehler der zehnfachen Kreuzvalidierung auf dem Datensatz von [HKF+08]. Es sind 90% zum Training und 10% zum Testen verwendet worden.

Partition	RRP Testdaten	Testfehler
1	0,0956	0,0002
2	0,1200	0,0016
3	0,0995	0,0003
4	0,0191	0,0052
5	0,0634	0,0005
6	0,0992	0,0003
7	0,0788	0,0000
8	0,0539	0,0011
9	0,0964	0,0002
10	0,1140	0,0011
Durchschnitt:	**0,0840**	**0,0011**

mierten Werten für a,b und c wird auf dem Testdatensatz der durchschnittliche RRP bestimmt. Tabelle 4.2 zeigt die zehn Partitionen und den durchschnittlichen RRP auf den Testdaten. Der Testfehler Err (Gleichung 4.4) wird wie folgt berechnet:

$$
\begin{aligned}
x_i & \quad \text{durchschnittlicher RRP der Testpartitionen i} \\
N &= 10 \text{ (zehnfache Kreuzvalidierung)} \\
x &= (x_1, \ldots, x_N) \\
Err(x_i) &= (x_i - x_{\bar{i}})^2 \\
x_{\bar{i}} &= \frac{1}{N-1} \sum_{j=1,\ j \neq i}^{N} x_j \\
Err\,(x) &= \frac{1}{N} \sum_{i=1}^{N} Err\,(x_i)\ . \qquad (4.4)
\end{aligned}
$$

Der durchschnittliche Testfehler (Gleichung 4.4) ist mit $0,0011$ sehr gering und daher kann davon ausgegangen werden, dass keine Überanpassung stattgefunden hat. Die endgültigen Parameter für a, b und c sind anschließend auf allen Daten trainiert

worden [HTF08]: $a = 0,99$, $b = 0,87$ und $c = 0,17$ und für die Evaluierung mit den GC/EI-MS (Abschnitt 4.7) verwendet worden. Zur Auswertung der MS/MS Daten werden die entsprechend trainierten Parameter der einzelnen Partitionen für die Scoring Funktion verwendet, sodass die Testdaten nicht zum Trainieren verwendet wurden.

4.6 Evaluierung von MS/MS Daten

Die Evaluierung von MetFrag mit hochauflösenden MS/MS Spektren wird im Folgenden beschrieben. Außerdem wird ein Vergleich mit MassFrontier angestellt. Diese Tests verwenden die Spektren von [HKF$^+$08] mit unterschiedlichen PubChem Versionen aus den Jahren 2006 und 2009. Für die Evaluierung wird im Folgenden neben dem Rang der korrekten Verbindung in der sortierten Ergebnisliste auch der Tanimoto Cluster Rang mit angegeben. Dieser fasst strukturell sehr ähnliche Kandidaten mit gleichem Score zusammen.

4.6.1 MetFrag - Hill Daten mit PubChem 2009

Das erste Experiment umfasst die Evaluierung von MetFrag auf den Daten von [HKF$^+$08] unter der Verwendung von PubChem 06/2009. Die Parameter der Scoring Funktion sind entsprechend der durchgeführten zehnfachen Kreuzvalidierung (Abschnitt 4.5) gewählt worden, sodass kein auswendig lernen stattgefunden hat. Es sind nur Kandidaten, die aus den Elementen CHONPS aufgebaut sind und mindestens ein Heteroatom enthalten, verwendet worden. Abbildung 4.9 veranschaulicht im linken Diagramm die Anzahl an Kandidaten, die pro Spektrum bearbeitet worden sind. Durchschnittlich sind, mit einer erlaubten Abweichung von 10ppm von der berechneten Masse, 2 544 Strukturen als Treffer in der Datenbank gefunden worden. Wie in Abschnitt 3.1.2 beschrieben, sind alle Kandidaten vorverabeitet worden, sodass jede Bindung mit der Bindungsordnung annotiert ist. In der Mitte von Abbildung 4.9 (rechts logarithmiert) ist der Tanimoto Cluster Rang der richtigen Verbindung dargestellt. Im rechten Diagramm kann abgelesen werden (gestrichelte

Linien), dass der Rang des richtigen Kandidaten in der Hälfte der Fälle besser als 9,5 ist. Der durchschnittliche RRP beträgt 0, 084.

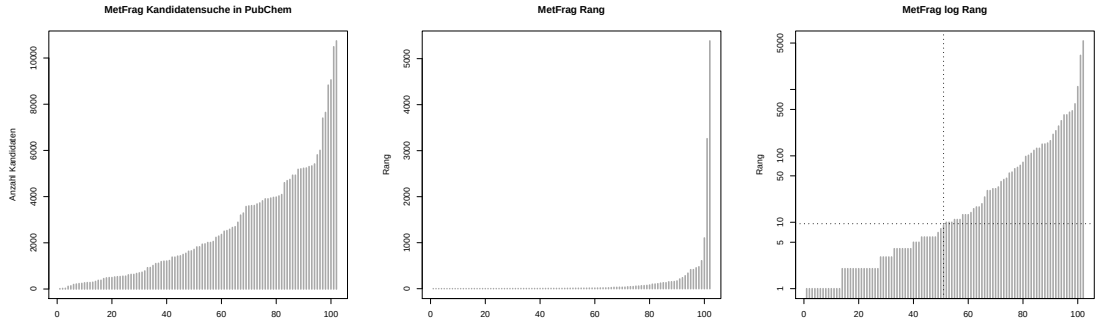


Abbildung 4.9. Kandidatensuche (10 ppm erlaubte Abweichung) in PubChem mit exakter Masse des Analyten (links). MetFrag (Baumtiefe 1) Tanimoto Cluster Rang des richtigen Moleküls (mitte) und mit logarithmierter Ordinate (rechts).

Im folgenden Abschnitt wird ein Vergleich zwischen der aktuellen und der MetFrag Version von [WSMHN10] durchgeführt. Diese ältere Version ist im März 2010 publiziert worden und verwendet als Bindungsgewichte statische Bindungsdissoziationsenergien (BDE) aus [Luo03]. Dieser Wert beschreibt, wieviel Energie benötigt wird, um die Bindung zwischen zwei Atomen zu brechen. Beispiele sind die Einfachbindung zwischen zwei Kohlenstoffatomen `C-C` mit 348 $\frac{\text{kJ}}{\text{mol}}$ oder eine Kohlenstoffdoppelbindung `C=C`: 612 $\frac{\text{kJ}}{\text{mol}}$ [Luo03]. Die BDE berücksichtigt somit weder, ob sich die Bindung in einem Ring befindet, noch die nähere Umgebung der verbundenen Atome. Weiterhin ist in [WSMHN10] eine andere Scoring Funktion verwendet worden, die statt der Bindungsordnung, die BDE berücksichtigt. Neutralverluste werden, im Gegensatz zu Kapitel 3.1.4, am Anfang direkt im Molekülgraph abgespalten und bilden somit die erste Ebene (Baumtiefe 1) im Fragmentierungsbaum. Baumtiefe 2 (BT2) aus [WSMHN10] entspricht somit BT1 der aktuellen Version.

Abbildung 4.10 zeigt die Verteilung der Ränge, die jeweils individuell nach Rang des Korrekten in der Ergebnisliste sortiert sind. Die grüne Linie beschreibt die Tanimoto Cluster Ränge von MetFrag (BT 1) mit Bindungsordnung, das in 50% der Fälle den Richtigen auf Rang 9,5 (gestrichelte Linie) oder besser eingeordnet hat. Erhöht man die Baumtiefe auf 2, dann verschlechtert sich der Median auf 14,5. Die Massensuche in PubChem mit einer erlaubten Abweichung von 10ppm hat durchschnittlich 2544 Kandidaten zurückgeliefert. Der durchschnittliche RRP beträgt 0, 084 bzw. 0, 104

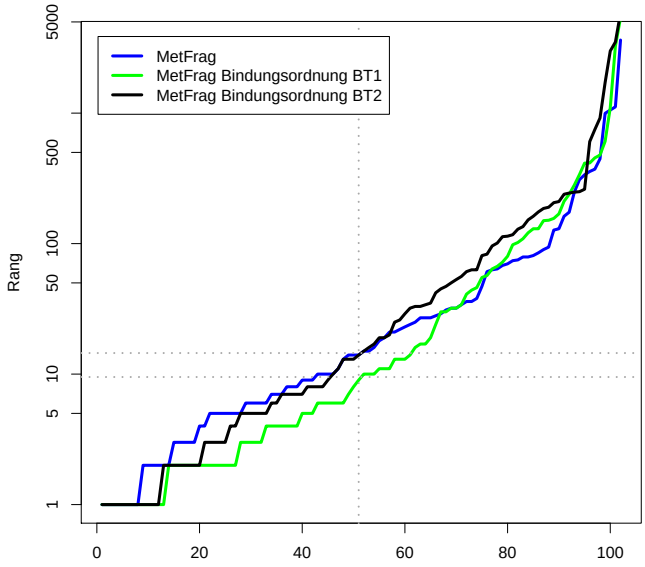


Abbildung 4.10. Vergleich der Tanimoto Cluster Ränge zwischen MetFrag aus [WSMHN10] und der aktuellen Version (Baumtiefe 1 und 2) mit der Bindungsordnung als Maß für die Stärke einer Bindung. Die Ränge sind jeweils individuell aufsteigend sortiert worden.

unter Verwendung von Baumtiefe 2. Die blaue Linie in Abbildung 4.10 zeigt die Ergebnisse aus [WSMHN10] mit einem Median von 14,5 und einem durchschnittlichen RRP von $0,089$. Im Mittel sind 2509 Kandidaten bearbeitet worden.

Vergleicht man den RRP zwischen der aktuellen (BT 1) und der MetFrag Version aus [WSMHN10] zeigt sich, dass durch die Verwendung der Bindungsordnung ein besseres Ergebnis erzielt wird. Der Median ist von 14,5 auf 9,5 reduziert worden. Abbildung 4.10 unterstreicht diesen Eindruck, da vor allem die Anzahl der Top 10 Ränge von 46 auf 54 zugenommen hat. Es sind wenigen Verbindungen schlechtere Ränge zugewiesen worden. Vor allem Strychnin und Strychnin N-oxid (Abbildung 4.11) mit einem Tanimoto Cluster Rang von 3261 bzw. 5386 haben einen schlechteren Rang als in [WSMHN10], denen Rang 997 und 3632 zugeordnet worden ist. Beide Ergebnisse sind trotz allem keine Hilfe bei der Identifizierung der unbekannten gemessenen Verbindung. Durch Verwendung einer größeren Baumtiefe erhält man im Mittel ein schlechteres Ergebnis. Zum Beispiel ist Terbutaline (CID: 5403) durch Erhöhung der Baumtiefe auf Cluster Rang 8, statt 102 (BT 1), eingeordnet worden. Auf der an-

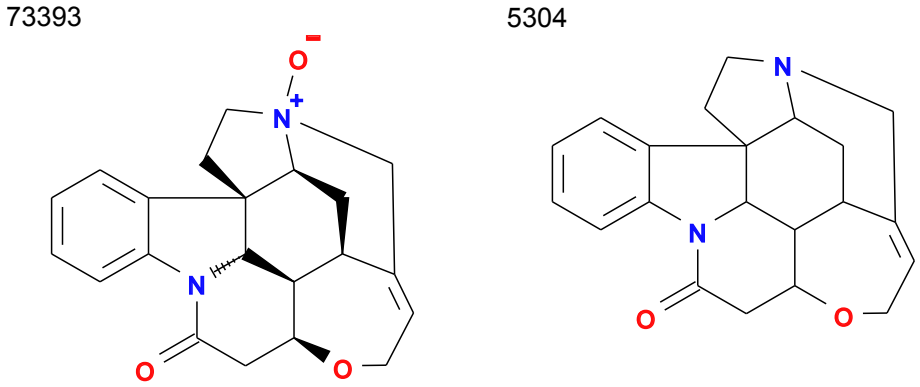


Abbildung 4.11. Strychnine (CID: 5304) und Strychnine N-oxide (CID: 73393) werden von MetFrag auf einen schlechten Rang eingeordnet.

deren Seite verschlechtert sich zum Beispiel der Rang von Anileridine (CID: 8944) von Rang 3 auf 917. Die leicht unterschiedliche Anzahl der Kandidaten zwischen der aktuellen und der MetFrag Version aus [WSMHN10] ist durch die verschiedenen CDK Versionen zu erklären, das im Bereich der Typisierung von Atomen deutlich verbessert wurde.

Letztendlich erzielt MetFrag mit den aus der Bindungsordnung abgeleiteten Bindungsgewichten ein besseres Ergebnis, als das mit den Bindungsdissoziationsenergien aus [WSMHN10] möglich gewesen ist. Jedoch ist die in Kapitel 3.1.2 vorgestellte Vorgehensweise zeitaufwendig, da die Struktur der Kandidaten mindestens einmal protoniert und optimiert werden muss. Insgesamt sind 259 507 Kandidaten vorverarbeitet worden, wobei manche Kandidaten, zum Beispiel Dobutamine (CID: 36811) und Isoxsuprine (CID: 3783) identische Summenformeln und dadurch die selben Kandidaten haben. Die detaillierten Ergebnisse der MetFrag Versionen sind in Tabelle A.3 (Bindungsordnung) bzw. Tabelle A.4 (BDE) zu finden.

4.6.2 Vergleich mit MassFrontier - PubChem 2006

Um einen Vergleich zwischen MassFrontier, MetFrag (BDE) [WSMHN10] und MetFrag mit Bindungsordnungen durchzuführen, ist im Folgenden auf die Spektren und Kandidaten von [HKF$^+$08] zurückgegriffen worden. In dieser Veröffentlichung ist ein Verfahren vorgestellt worden, dass es ermöglichen soll Kandidatenmoleküle entsprechend einer Rangfolge zu ordnen. Dafür ist von den Kandidaten (10 ppm erlaubte

Abweichung) aus PubChem (Februar 2006) mit Hilfe von MassFrontier 4 ein *in silico* Spektrum generiert worden. Dieses wurde mit dem experimentell aufgenommenen Spektrum verglichen und die Anzahl der übereinstimmenden Peaks bilden die Grundlage der resultierenden Rangfolge. Für die im Hauptteil von [HKF+08] vorgestellten Ergebnisse, ist nachträglich das Spektrum einer bestimmten Kollisionsenergie ausgewählt worden, dass das beste Ergebnis liefert. Dieser Schritt benötigt jedoch Vorwissen, sodass ein unverfälschter Vergleich nicht möglich ist. Deshalb ist MetFrag mit den Ergebnisse aus dem Anhang von [HKF+08] verglichen worden. Die Autoren haben hierfür das Spektrum mit der kleinsten Kollisionsenergie ausgewählt, dessen Intensität des Molekülions kleiner als 22% ist. MetFrag benutzt wieder über alle Kollisionsenergien zusammengefasste Spektren.

Als Grundlage für die Kandidatensuche dient PubChem vom Februar 2006 mit $6 \cdot 10^6$ Einträgen. Diese Abfrage liefert im Durchschnitt 341 Kandidaten zurück, wobei [HKF+08] im Mittel 272 Strukturen pro Spektrum verarbeitet hat. Es sind nur Verbindungen verwendet worden, die aus den Elementen CHONPS aufgebaut sind und die nicht nur aus Kohlenstoff und Wasserstoff bestehen. Eine mögliche Ursache für die Diskrepanz zwischen der Anzahl an Kandidaten sind unterschiedliche PubChem Daten oder möglicherweise nachträglich geänderten Einträge. Obwohl MetFrag den Nachteil hat, mehr Kandidaten pro Spektrum zu prozessieren, wird es im Folgenden mit den Ergebnissen aus [HKF+08] verglichen.

Abbildung 4.12 zeigt die Verteilung (logarithmierter Ordinate) der Ränge von MassFrontier (rot) [HKF+08], MetFrag mit BDE [WSMHN10] (blau - Tanimoto Cluster Rang) und der aktuellen MetFrag Version (grün - Tanimoto Cluster Rang). Der Median beträgt 4 für MassFrontier und MetFrag (BDE) bzw. 2 mit der aktuellen MetFrag Version. Tabelle 4.3 zeigt die Ergebnisse für eine Auswahl der 102 Spektren (alle Ergebnisse in Tabelle A.5).

Der durchschnittliche RRP von MassFrontier ist $0,1180$. MetFrag (BDE) erreicht im Mittel einen RRP von $0,1102$. Das arithmetische Mittel des RRP der aktuellen MetFrag Version beträgt $0,0952$.

Da unterschiedlich viele Kandidaten verwendet worden sind, kann man objektiv nur den RRP vergleichen, der auch die Cluster Ränge nicht mit in die Berechnung einbezieht. MassFrontier und MetFrag (BDE) liegen sehr dicht beieinander

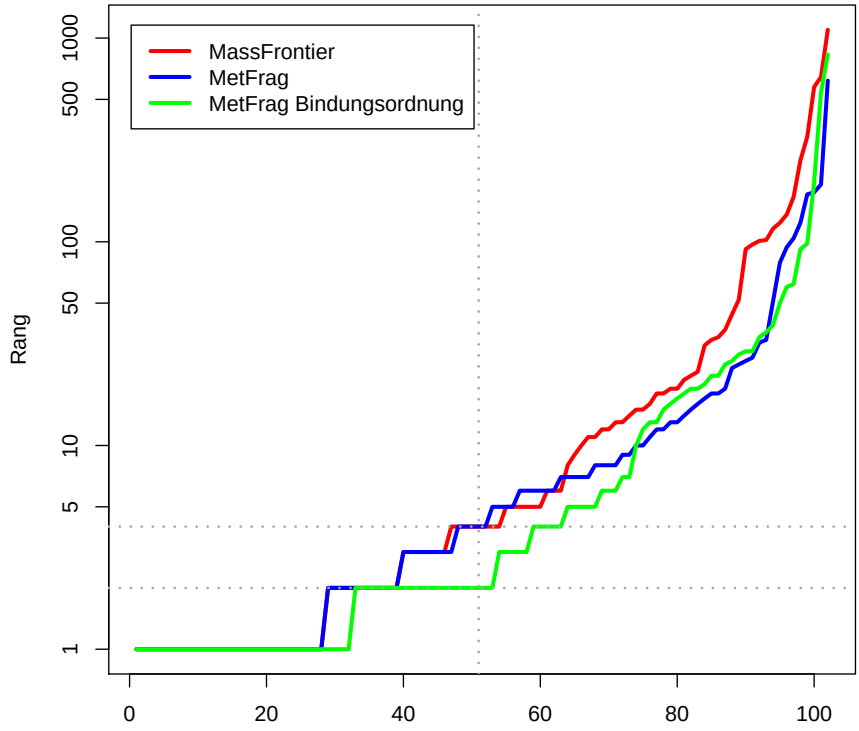


Abbildung 4.12. Vergleich der Ränge zwischen MetFrag (Tanimoto Cluster Rang) aus [WSMHN10], der aktuellen Version mit Bindungsordnung (Tanimoto Cluster Rang) und MassFrontier. Es wurden nur Kandidaten verwendet, die im oder vor Februar 2006 zu PubChem hinzugefügt worden sind.

wobei letzteres besser abschneidet, wie dies auch in [WSMHN10] gezeigt wurde. Eine bessere Rangordnung ermöglicht die aktuelle MetFrag Version, die den geringsten RRP aufweisen kann. Ausreißer sind wiederum Strychnin und Strychnin N-oxide, die auch MassFrontier Probleme bereiten. Zusammenfassend erreicht die aktuelle MetFrag Version die beste Leistung, wobei auch mit der BDE als Bindungsgewicht [WSMHN10] bessere Ergebnisse erzielt werden als mit MassFrontier 4.0.

Tabelle 4.3. Ausschnitt aus dem Vergleich zwischen den Ergebnissen von MassFrontier 4 [HKF+08] und MetFrag. Die MS/MS Spektren stammen aus [HKF+08] und die Kandidaten sind mit exakter Masse (10 ppm erlaubte Abweichung) in PubChem (Februar 2006) gesucht worden. Der durchschnittliche RRP von MetFrag ist mit 0,0952 besser als der von MassFrontier mit 0,1180. Die kompletten Ergebnisse sind in Tabelle A.5 zu finden.

	MassFrontier		**MetFrag Bindungsordnung**		
Verbindung	**Kand.**	**Rang**	**Kand.**	**Rang**	**Tan. Rang**
Adenosine Diphosphate	32	3	46	6	1
Alfentanil	134	1	162	1	1
Ampicillin	615	1	780	5	1
Apramycin	54	1	58	3	1
Dihydroergotamine	35	1	38	2	1
Diphenoxylate	333	4	369	1	1
Ergocristine	16	1	27	6	1
Ergoloid Mesylate	7	1	10	1	1
Etodolac	420	1	579	1	1
Fenoterol	370	5	519	5	1
Gallamine	10	1	8	1	1
Leucine Enkephalin	53	2	60	3	1
Methionine Enkephalin	66	1	68	2	1
Methylergonovine	515	1	629	6	1
Morphine-3-Glucuronide	179	2	170	2	1
Oxybutynin	114	6	155	3	1
Oxytetracycline	483	4	617	11	1
Piperacetazine	494	1	625	1	1
...	...	...	...	...	...
Naltrexone	1035	34	1421	359	196
Strychnine	664	575	882	824	540
Strychnine N-oxide	1185	1098	1667	1338	830
Durchschnitt:	**272,2±24,2**	**44,2±14,1**	**340,6±31,8**	**39,8±15,6**	**25,1±9,8**
Standardabweichung:	**244,17**	**142,47**	**320,67**	**157,95**	**99,25**
Median:	**183,5**	**4**	**248**	**5**	**2**
75% Quantil:	**431,25**	**17,5**	**513,25**	**21**	**13**

4.6.3 Einfluss der Massengenauigkeit auf die Leistung von MetFrag

Nominalmassen allein machen es viel schwieriger, das richtige Fragment einem Peak zuzuordnen. Um dies zu verdeutlichen, ist die aktuelle MetFrag Version zum einen mit einer relativen Abweichung von 10 ppm auf den Daten von [HKF+08] evaluiert

worden und zum anderen mit einer erlaubten absoluten Abweichung von 0,5 Da, um Nominalmassen simulieren zu können.

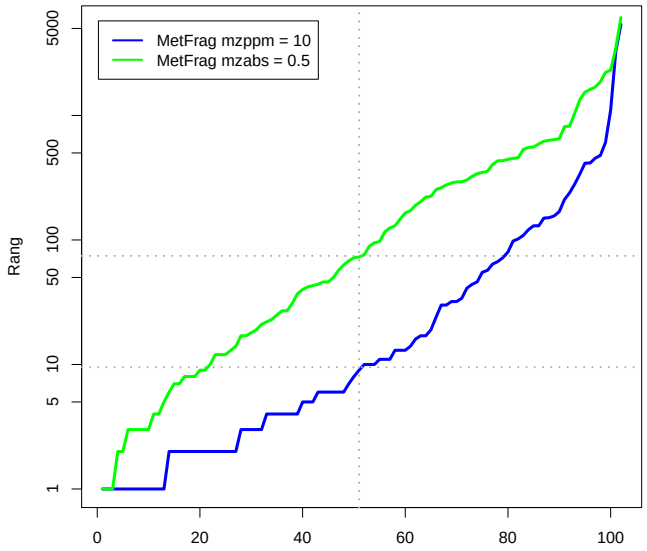


Abbildung 4.13. Vergleich der von MetFrag zurückgelieferten Ränge der korrekten Verbindung mit einer erlaubten Abweichung von 10 ppm (blau) und mit einer absoluten Abweichung von 0,5 Da (grün).

Abbildung 4.13 zeigt die Verteilung der Ränge mit logarithmierter Ordinate und Tabelle A.6 die detaillierten Ergebnisse. Der Median erhöht sich von 9,5 auf 74,5 und der RRP von 0,084 auf 0,238. Trotzdem ist im folgenden Abschnitt versucht worden, MetFrag mit GC/MS Daten gegen die kommerzielle Software MassFrontier 5, ACD/MS Fragmenter und MOLGEN-MS zu vergleichen.

4.7 Grenzen von MetFrag mit GC/EI-MS Daten

Neben hochauflösenden ESI-MS Geräten, werden häufig GC/MS Geräte eingesetzt, die meist über kein Flugzeitanalysator (TOF) verfügen und somit nur Nominalmassengenauigkeit erreichen. MetFrag wird im folgenden Kapitel auf GC/EI-MS Daten (Elektronenstoßionisation) evaluiert und mit kommerzieller Software verglichen. EI-MS ist besser untersucht und verstanden als ESI-MS, weshalb viele Regeln

zur Fragmentierung bekannt und alleine in NIST '11 243 893 Spektren von 212 961 verschiedenen Verbindungen verfügbar sind. Diese GC/MS (Nominalmassen) Spektren (siehe Kapitel 4.3) aus der älteren NIST '05 wurden von [SMB09] verwendet, um einen Vergleich zwischen bestehender Software zur Fragmentvorhersage durchzuführen.

4.7.1 Vergleich mit ähnlicher Software

MetFrag ist das einzige Programm in diesem Vergleich, dass keine regelbasierten Algorithmen einsetzt, sondern nur kombinatorisch alle möglichen Fragmente erzeugt und mit der Scoring Funktion eine Reihenfolge der Kandidaten erstellt. Insgesamt sind 183 872 Kandidaten, wie in Kapitel 3.1.2 beschrieben, vorverarbeitet worden. Da nicht jeder Kandidat ein Heteroatom enthält, sind auch Kohlenstoffe protoniert worden. Die verwendeten MetFrag Einstellungen sind im Kapitel 4.3 beschrieben.

Tabelle 4.4. GC/MS Resultate von [SGK$^+$] mit 100 bzw. 27 GC/EI-MS Spektren aus NIST '05. MetFrag (BDE) [WSMHN10] ist mit Baumtiefe (BT) 2 und 3 verwendet worden. Außerdem ist die aktuelle MetFrag Version mit Bindungsordnungen (BT 1 und 2) mit MassFrontier (3 bzw. 5 Schritt Fragmentierung und mit/ohne Bibliothek), ACD Fragmenter (3 bzw. 5 Schritt Fragmentierung) und MOLGEN-MS verglichen worden. Die Ergebnisse von MassFrontier, ACD Fragmenter und MOLGEN-MS stammen aus [SMB09].

	RRP 100 Spektren	RRP 27 Spektren
MetFrag BT 1	0,3430	0,4850
MetFrag BT 2	0,3803	0,4146
MetFrag (BDE) BT 2	0,3587	0,4302
MetFrag (BDE) BT 3	0,4806	0,4510
MassFrontier 3 Schritt	0,2685	0,3754
MassFrontier 5 Schritt	0,3527	0,3926
MOLGEN-MS	0,2734	0,3519
ACD Fragmenter 3 Schritt		0,5197
ACD Fragmenter 5 Schritt		0,5349
MassFrontier 3 Schritt mit Bibliothek		0,3887
MassFrontier 5 Schritt mit Bibliothek		0,3815

Es wird zwischen 100, der komplette Datensatz, und 27, eine Auswahl mit nur wenigen Kandidatenstrukturen, Spektren unterschieden. Diese Einteilung ist von [SMB09] vorgenommen worden, da die Laufzeit von ACD Fragmenter und MassFrontier mit Benutzerbibliothek zu lang gewesen ist. Tabelle 4.4 zeigt das arithmetische Mittel der RRPs der einzelnen Programme.

Die 27 Spektren haben durchschnittlich 55 Kandidaten (maximal 200). MOLGEN-MS erreicht mit einem durchschnittlichen RRP von 0,3519 das beste Ergebnis vor MassFrontier (3 Schritt) 0,3754 und 0,3815 (3 Schritt mit Bibliothek) und MetFrag (BT 2) mit 0,4146 bzw. 0,4850 (BT 1). Die MetFrag Version aus [WSMHN10] hat im Mittel einen RRP von 0,4302 mit Baumtiefe 2. Bei Verwendung von Baumtiefe 3 verschlechtert sich das Ergebnis auf 0,4510. Die Ergebnisse von ACD Fragmenter sind mit einem durchschnittlichen RRP von 0,5197 (3 Schritt) bzw. 0,5349 (5 Schritt) wesentlich schlechter. Insgesamt hat dieser Test mit nur 27 Spektren wenig Aussagekraft, da die Datenbasis zu klein gewählt ist. Zudem besitzen 10 Spektren weniger als 20 Kandidaten und 6 davon 8 oder weniger, sodass dadurch jede Verschlechterung des Ranges einen großen Einfluss auf den mittleren RRP hat. Insgesamt erreicht die regelbasierte Software (ohne ACD Fragmenter) bessere Ergebnisse als MetFrag. Interessanterweise hat MetFrag mit BT 2 den besten RRP unter den MetFrag Versionen, das durch die kleine Datenbasis mit sehr wenigen Kandidaten erklärt werden kann: Schwankungen vom Rang der richtigen Verbindung haben einen sehr großen Einfluss auf den durchschnittlichen RRP. Normalerweise werden mit steigender Baumtiefe mehr Peaks erklärt, welches aber nicht unbedingt den Rang des Richtigen in der Ergebnisliste verbessert (siehe Abschnitt 4.6.1).

Um ein repräsentativeres Ergebnis zu erhalten, ist außerdem der gesamte Datensatz mit 100 Spektren evaluiert worden. Den besten durchschnittlichen RRP erreicht MassFrontier (3 Schritt) mit 0,2685 vor MOLGEN-MS (0,2734) und MetFrag (BT 1) mit einem mittleren RRP von 0,3430. MassFrontier mit 5 Schritt Fragmentierung hat einen durchschnittlichen RRP von 0,3527, MetFrag (BT 2, BDE) 0,3587. Die größeren Baumtiefen führen zu einem größeren RRP mit 0,3803 (MetFrag BT 2) und 0,4806 (MetFrag BDE BT 3). Im Gegensatz zu dem kleinen Datensatz mit 27 Spektren wird hier bestätigt, dass geringere Baumtiefen besser für das Scoring sind.

Insgesamt ist MOLGEN-MS und MassFrontier (3 Schritt Fragmentierung) das beste Programm für (diese) GC/MS Spektren. Da Umlagerungen in EI-MS Spektren wesentlich häufiger auftreten als in ESI-MS, haben die regelbasierten Programme (mit Ausnahme von ACD Fragmenter) bessere Ergebnisse. Trotz eines anderen Ionisierungsverfahrens, dass mit keiner Protonierung des Analyten einhergeht, erreicht das aktuelle MetFrag mit Bindungsordnungen bessere Ergebnisse als die ältere Version aus [WSMHN10].

Tabelle A.7 zeigt die kompletten Ergebnisse, wobei fett gedruckte NIST IDs Teil des kleineren Datensatzes sind.

4.8 MassStruct Evaluation

Im folgenden Abschnitt werden die Resultate von MassStruct [HWN11] genauer betrachtet. Zum einen wird der Rang der korrekten Verbindung mit unterschiedlich großen Trainingdatensätzen ausgewertet, zum anderen wird die Laufzeit mit PubChem (Q4 2010) evaluiert.

Ergebnisse zur Vorsortierung der Kandidaten

Es sind 240 MS/MS Spektren mit unterschiedlichen Kollisionsenergien von 218 Verbindungen (siehe komplette Ergebnistabelle A.8 und Abschnitt 4.3) verwendet worden. Insgesamt enthalten die Spektren 2 083 Peaks, von denen MetFrag 1 280 Peaks annotieren konnte. Der PubChem Snapshot (Q4 2010) enthält 28 838 421 Verbindungen und belegt in der PostgreSQL Datenbank mit pgchem Erweiterung inklusive aller Indizes 150 GB an Speicherplatz.

Es ist eine zehnfache Kreuzvalidierung durchgeführt worden, um zu untersuchen ob übertrainiert wurde. Hierfür ist auf 216 (90%) zufällig gezogenen Spektren trainiert und auf 24 (10%) getestet worden. Das Training umfasst die Speicherung der mit MetFrag annotierten m/z→Struktur Paare in der Datenbank. Dieser Vorgang ist zehnmal wiederholt worden, sodass insgesamt 10 Partitionen vorhanden sind. Tabelle 4.5 zeigt die durchschnittlichen RRPs auf den jeweiligen Testdaten. Durch einen

geringen Testfehler (Gleichung 4.4) von 0,00260 kann davon ausgegangen werden, dass kein Overfitting stattgefunden hat.

Tabelle 4.5. MassStruct Testfehler der zehnfachen Kreuzvalidierung mit den 240 MS/MS Spektren. Es sind 90% zum Training und 10% zum Testen verwendet worden.

Partition	RRP Testdaten	Testfehler
1	0,24116	0,00025
2	0,29130	0,00159
3	0,22100	0,00146
4	0,29397	0,00183
5	0,25591	0,00000
6	0,27391	0,00042
7	0,26313	0,00007
8	0,30908	0,00355
9	0,26566	0,00013
10	0,13915	0,01669
Durchschnitt:		**0,00260**

Tabelle 4.6. MassStruct RRPs mit unterschiedlichen Partitionierungsgrößen, um eine größere Abdeckung zu simulieren.

	RRP	
Partitionierung	Median	ø
1:1	0,50	0,35
2:1	0,30	0,29
3:1	0,31	0,30
4:1	0,25	0,29
9:1	0,16	0,28

Zur Auswertung sind zufällig Spektren gezogen und die mit MetFrag annotierten m/z→Struktur Paare in der Trainingsdatenbank gespeichert worden. Die Aufteilung von Trainings- und Testspektren ist in verschiedenen Verhältnissen (1:1, 2:1, 3:1, 4:1 und 9:1) vorgenommen worden, um den Einfluss der Anzahl von Trainingsspektren zu überprüfen. Tabelle 4.6 zeigt die Ergebnisse dieses Experimentes.

Tabelle 4.7. Auszug aus dem Ergebnis der 1:1 Partitionierung. Alle Ergebnisse sind in Tabelle A.8 zu finden. Die Trainingsdaten sind zufällig in zwei gleichgroße Partitionen aufgeteilt und jeweils auf dem einen trainiert und dem anderen getestet worden. Im zweiten Schritt sind die Datensätze zum Trainieren und Testen getauscht worden. Die angegebene Laufzeit beinhaltet nur die Zeit der SQL-Anfrage. (nach [HWN11])

MassBank ID	CID	RRP	Kand.	Laufzeit in s
PR100121 PR100122	13804	0,000	50910	477
PR100113 PR100114	834	0,000	34491	283
PR100317	13804	0,001	50910	448
PR100239	5319853	0,001	72127	867
PR100390	165627	0,001	8743	100
PR100329	717531	0,001	29828	252
PR100198	439155	0,001	105691	950
PR101031	5274585	0,001	46649	526
PR100359	6441269	0,001	72127	793
PR100296	92136	0,002	8743	84
PR100076	34755	0,002	108138	932
PR100363	442456	0,002	9745	200
...	...	...	...	...
PR100119 PR100120	6132	0,698	48529	561
PR100414	4687	0,699	18873	186
PR100185	637540	0,703	13098	191
PR101040	6433206	0,714	9765	57
PR100394	649	0,739	4175	42
PR100413	70346	0,750	22378	209

Durch eine bessere Abdeckung der Trainingsspektren verbessert sich der durchschnittliche RRP von 0,35 auf 0,28 und der Median verbessert sich von 0,50 auf 0,16. Ein Teil der Ergebnisse der 1:1 Partitionierung ist in Tabelle 4.7 dargestellt. Insgesamt liefert MassStruct, verglichen mit einer randomisierten Reihenfolge, den korrekten Kandidaten im Durchschnitt eher zurück. MassStruct ändert nichts an der Kandidatenauswahl, da lediglich die Reihenfolge beeinflusst wird. Somit sollte MassStruct verwendet werden, auch wenn keine große Trainingsmenge vorhanden ist, da der richtige Kandidat im Durchschnitt deutlich früher prozessiert wird.

Laufzeit

Im folgenden Abschnitt wird die Zeit, die MassStruct für eine Datenbankanfrage auf dem kompletten Datensatz benötigt, genauer betrachtet. Im Durchschnitt werden 31 700 Kandidaten (grüner Kreis in Abbildung 4.14), 16 360 im Median, und bis zu 100 000 bearbeitet. Eine Anfrage dauert durchschnittlich 330s, oder 10ms pro Kandidat.

Der (virtualisierte) Datenbankserver verfügte über 2 CPUs, 2 GB RAM und lief auf einem VMWare ESX Cluster mit 2,6 GHz Intel Xeon CPUs. Die Datenpartition befand sich auf einen FC-SAN Speichersystem.

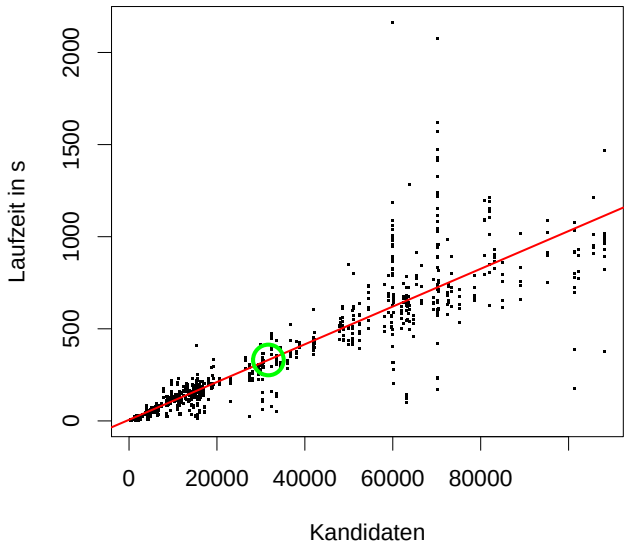


Abbildung 4.14. Laufzeiten der SQL Abfragen für den kompletten Datensatz. Der Anstieg der Regressionsgerade beträgt 10ms/Kandidat, wobei der Durchschnitt (32 000 Kandidaten in ≈5 min) markiert ist. (nach [HWN11])

Die Laufzeit ist abhängig von der Anzahl der Kandidaten, was bedeutet je genauer das Massenspektrometer ist, desto weniger Kandidaten und desto schneller ist die Abfrage. Normalerweise sind bei großen Datenbanken die Ausführungszeiten abhängig von der Festplattenleistung, in diesem Fall wird die meiste Zeit für die Substruktursuche benötigt.

4.9 Anwendungen von MetFrag in der Massenspektrometrie Community

Im folgenden Abschnitt wird die Nutzung von MetFrag[27], weitere Anwendungen und Feedback von Benutzern vorgestellt.

Die Webseite ermöglicht es, direkt Feedback abzusenden, was im Laufe der Zeit zu einer stetigen Verbesserung von MetFrag und der Weboberfläche geführt hat. Abbildung 4.15 zeigt die Besucherzahlen von März 2010, dem Veröffentlichungsdatum von MetFrag [WSMHN10], bis zum Dezember 2011. Von anfänglich ca. 400-800 Besuchern pro Monat im Jahr 2010 ist die Anzahl der Besucher im Dezember 2011 auf 1 295 angestiegen. In dem beobachteten Zeitraum sind insgesamt 180 Feedback Anfragen beantwortet worden.

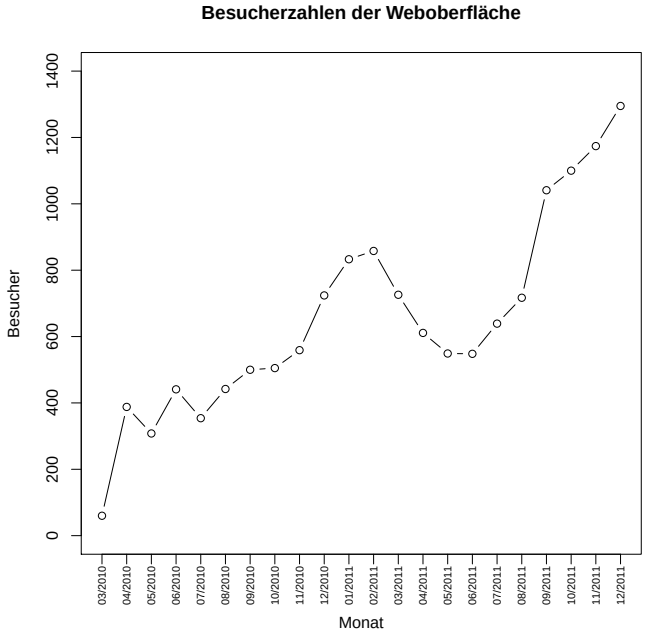


Abbildung 4.15. Benutzerzahlen der MetFrag Weboberfläche von März 2010 bis Dezember 2011.

Weiterhin wird der Fragmentierungsalgorithmus von MetFrag für die Vorhersage von Fragmentstrukturen in METLIN (siehe Abschnitt 2.2.2) verwendet. Das „National

[27] `http://msbi.ipb-halle.de/MetFrag/` - Abgerufen im Dezember 2011

Physical Laboratory“ für „Surface and NanoAnalysis“ (Middlesex, Großbritannien) hat MetFrag erfolgreich mit G-SIMS („Gentle Secondary Ion Mass Spectrometry“) [GS00] Spektren getestet. Tabelle 4.8 zeigt die Ergebnisse mit vier Beispielmolekülen.

Tabelle 4.8. MetFrag Ergebnisse für 4 mit G-SIMS gemessenen Verbindungen und passenden Kandidaten aus PubChem.

Molekül	Kand.	Peaks	erklärte Peaks	Rang
Irganox 1010	30	8	6	1
Folsäure	414	4	3	2
Koffein	1705	7	2	5
Cholesterol	499	14	3	19

Des Weiteren wurde der MetFrag Score, neben weiteren (experimentellen) Daten, verwendet um zwei unbekannte Verbindungen (GC/EI-MS Spektren) zu identifizieren [SGK+].

MetFrag ist an die Herstellersoftware Smart Formula 3D (Bruker) angebunden. Durch Übergabe von Peakliste, Masse, Summenformel und Moleküldatenbank (mittels GET Parameter, Abschnitt 3.2) kann die Weboberfläche direkt aufgerufen werden. Abbildung 4.16 zeigt die Übermittlung der Messdaten in die entsprechenden Eingabefelder.

Neben Smart Formula 3D kann auch SIRIUS [BLLP09] die Daten direkt an die Webseite übergeben.

4.10 Zusammenfassung

In diesem Kapitel ist die Wahl der Vorverarbeitung mit verschiedenen Experimenten erläutert worden und wieso letztendlich die Bindungsordnung statt der Bindungslängenänderung verwendet wird. Die exponentielle theoretische Laufzeit ist analysiert und praktisch untersucht worden. Die durchgeführte Parameteroptimierung resultiert in a = 0,99, b = 0,87 und c = 0,17, als verwendete Parameter der Scoring Funktion. Die Kreuzvalidierung bestätigt, dass kein Overfitting durchgeführt worden ist. Der RRP, als ein Maß zur Bestimmung der Güte von Software zur

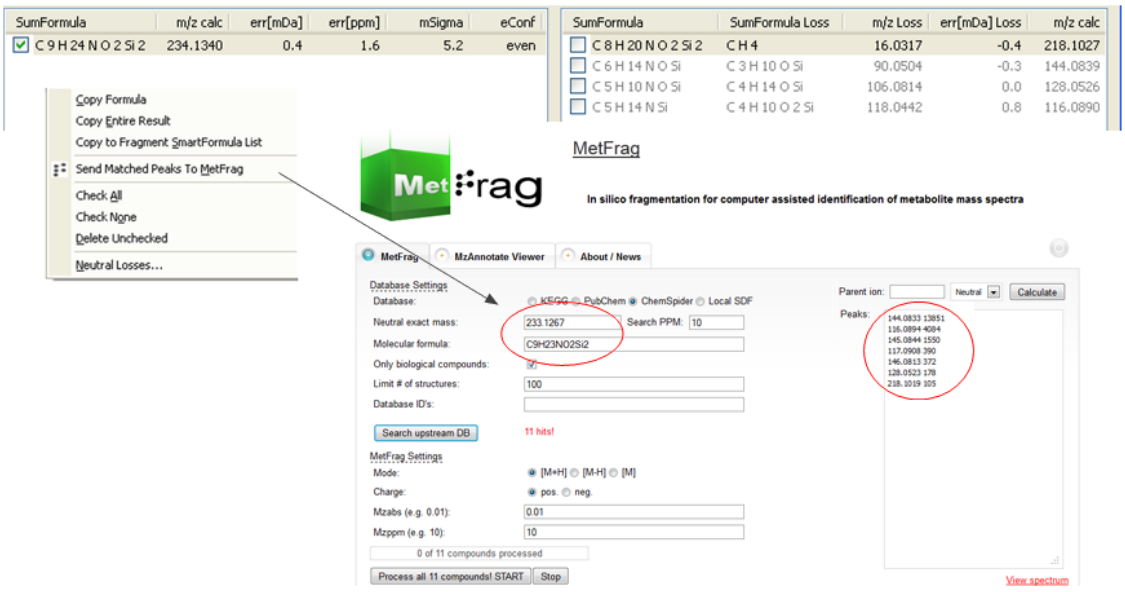


Abbildung 4.16. Übertragung der Peakliste, Masse und Summenformel des Analyten von Smart Formula 3D (Bruker) zu der MetFrag Weboberfläche.

Fragmentvorhersage, und der Median werden benutzt, um die Ergebnisse von MetFrag und ähnlicher Programme auszuwerten. Als erstes ist ein 510 Spektren umfassender ESI-MS/MS Datensatz von [HKF+08] evaluiert worden. Die beste Leistung erreicht MetFrag mit Bindungsordnungen mit einem durchschnittlichen RRP von 0,084 (BT 1) bzw. 0,104 (BT 2) verglichen mit 0,089 der MetFrag Version mit BDE aus [WSMHN10]. Unter Verwendung der Kandidaten aus Pubchem vom Februar 2006 ist MetFrag mit Bindungsordnungen (RRP 0,095) besser als MetFrag mit BDE (RRP 0,110) und MassFrontier 4 (RRP 0,118). Es ist festgestellt worden, dass eine Nominalmassenauflösung den durchschnittlichen RRP von 0,084 auf 0,2384 erhöht. MetFrag ist auf 100 GC/MS Spektren aus der NIST '05 [SMB09] evaluiert und mit ähnlicher regelbasierter Software verglichen worden. Den besten durchschnittlichen RRP erreicht MassFrontier (RRP 0,2685) vor MOLGEN-MS (RRP 0,2734) und MetFrag mit Bindungsordnungen (RRP 0,3430) bzw. mit Bindungsdissoziationsenergien (RRP 0,3587). Die MassStruct Auswertung zeigt, dass durch eine erhöhte Abdeckung Trainingsspektren der durchschnittliche RRP des richtigen Kandidaten gesenkt wird und dadurch weniger Kandidaten prozessiert werden müssen. Die Besucherzahlen der MetFrag Weboberfläche zeigen einen Anstieg auf 1295 Besucher im Dezember 2011 von ehemals $\approx$ 400 Besuchern in den ersten Monaten nach der Veröffentlichung von [WSMHN10].

5 Zusammenfassung und Ausblick

Die Metabolomik ist ein wichtiger Zweig zur Erforschung der Funktion von biologischen Systemen. Zur Identifizierung von *Metaboliten* ist die Massenspektrometrie auch heute noch die Methode der Wahl. Vor allem hochauflösende ESI-MS/MS Instrumente ermöglichen die Messung von Analyten mit einer sehr hohen Massengenauigkeit. Die benötigte Zeit zur Durchführung einer Messung ist durch die Kopplung mit einer UPLC deutlich verringert. Dadurch ist immer noch die Identifizierung von Analyten der größte Flaschenhals im Metabolomik Workflow.

Die Identifizierung von Verbindungen anhand von GC/MS und vor allem MS/MS stellt immer noch ein Problem dar, da bis heute die Spektreninterpretation nicht komplett automatisiert werden konnte. Für GC/MS Spektren stehen umfangreicherere Spektrenbibliotheken zur Verfügung, die die Identifizierung einer Verbindung wesentlich erleichtern. ESI-MS/MS Massenspektrometer erlauben es die Probe mit unterschiedlichen Modi und Kollisionsenergien zu messen, die jeweils in unterschiedlichen Spektren resultieren und den Aufbau von einer umfangreichen Spektrenbibliothek zusätzlich erschwert.

MetFrag, die im Rahmen dieser Arbeit entwickelte Software, setzt an diesen Punkt an und ermöglicht eine Vielzahl von Kandidaten zu annotieren bzw. die gemessene Verbindung sogar zu identifizieren.

Software zur Fragmentvorhersage können in regelbasierende und kombinatorische Algorithmen eingeteilt werden. Erstere extrahieren Regeln aus Publikationen, die in einer Datenbank abgespeichert werden. Das hat den Nachteil, dass unbekannte Verbindungen nicht problemlos identifiziert werden können. Kombinatorische Algorithmen hingegen funktionieren mit allen Stoffklassen, beanspruchten aber bisher viel Zeit oder waren nicht öffentlich verfügbar.

MetFrag, eine kombinatorische Software zur Fragmentvorhersage, ist im Rahmen dieser Arbeit entwickelt worden. Der Quellcode steht unter einer Open Source Lizenz (GPL) und ist somit das bis dato einzige quelloffene Programm zur Fragmentvorhersage.

Das MetFrag System nutzt die Masse oder Summenformel einer gemessenen Verbindung, um passende Kandidaten in einer Moleküldatenbank zu finden. Jedes dieser Moleküle wird vorverarbeitet, damit die Bindungen mit der Bindungsordnung annotiert werden können. Danach findet die eigentliche Fragmentierung der Kandidaten statt, die durch die Verwendung geringer Baumtiefen und einer heuristischen Redundanzüberprüfung auch für tausende Kandidaten schnell realisiert werden kann. Durch die Verwendung „beweglicher“ Wasserstoffe und weniger Neutralverlustregeln können die Fragmente den Peaks zugeordnet werden. Die auf MS/MS Daten optimierte Scoring Funktion berechnet für jeden Kandidaten einen Score und stellt somit eine Rangfolge dieser her. Ein zusätzliches Clustering kann Kandidaten mit gleichem Score zusammenfassen, sofern diese eine hohe chemische Ähnlichkeit besitzen.

Zur Evaluierung ist ein 510 Spektren umfassender Datensatz verwendet worden, der Messungen von insgesamt 102 Verbindungen (102 zusammengefügte Spektren) enthält. Neben der theoretisch bestimmten exponentiellen Laufzeit ist diese auch praktisch an 45000 Kandidaten aus PubChem evaluiert worden. Außerdem wurde die Gesamtlaufzeit von MetFrag für das Prozessieren aller 102 Spektren mit PubChem 2009 betrachtet. Es wurden Experimente für die Auswahl eines geeigneten Maßes zur Angabe der Bindungsstärke durchgeführt. Die mit MOPAC berechneten Bindungsordnungen sind letztendlich verwendet worden, um das Scoring der einzelnen Kandidaten zu verbessern. Um eine optimale Gewichtung zwischen Masse, Intensität und Bindungsordnung zu erreichen, wurde eine Parameteroptimierung auf Basis der 102 Spektren durchgeführt. Der Vergleich zwischen den einzelnen MetFrag Versionen (BDE und Bindungsordnung), unter Verwendung von Pubchem Daten aus 2009, fällt zugunsten der aktuellen Version aus. Weiterhin wurde festgestellt, dass durch die Erhöhung der Baumtiefe kein besseres Ergebnis erzielt werden kann. Die kommerzielle, regelbasierte Software MassFrontier 4, die ebenfalls auf diesen Daten evaluiert wurde (PubChem 2006), lieferte schlechtere Ergebnisse als MetFrag. Durch

eine simulierte Nominalmassenauflösung verschlechtern sich erwartungsgemäß die Ergebnisse von MetFrag enorm.

Neben dem hochauflösenden ESI-MS/MS Datensatz wurde auch ein GC/MS Datensatz (Nominalmassen) verwendet, um die Grenzen von MetFrag auszutesten. Dabei reichen MassFrontier, MetFrag und ACD Fragmenter nicht an die Ergebnisse der speziell für GC/EI-MS Spektren entwickelten Software MOLGEN-MS heran.

Verbesserungspotential von MetFrag ist im Bereich des Parametertrainings vorstellbar. Es könnten unterschiedliche Parametersätze für verschiedene Ionisierungstechniken bzw. Massenspektrometer unterschiedlicher Hersteller trainiert werden, um bessere Ergebnisse zu erzielen. Hierfür müssten jedoch ausreichend viele Trainingsspektren vorhanden sein. Weiterhin könnten Summenformeln anstelle von Peakmassen verwendet werden. Smart Formula 3D und SIRIUS [BLLP09] sind Beispiele für solche Software, die Summenformeln zu Peaks mit Isotopeninformation zuordnen kann. Diese Information kann direkt im Fragmentierungsalgorithmus genutzt werden, um zu entscheiden, ob eine weitere Fragmentierung der Substruktur für die Annotation von Peaks von Nutzen ist. Desweiteren kann dadurch die Massendifferenz auf 0 gesenkt werden, sodass nur passende Summenformeln den Peaks zugeordnet werden.

Neben MetFrag wurde auch eine Methode zur intelligenten Kandidatensuche (MassStruct) vorgestellt. Dieser Ansatz lernt Peak $\rightarrow$ Fragment Assoziationen, die durch MetFrag bereitgestellt werden. Bei einer Massensuche mit einem Anfragespektrum kann die aufgebaute Trainingsdatenbank genutzt werden, um bereits bekannte Strukturen zu den gemessenen Peaks zu finden. Eine Substruktursuche sucht schließlich passende Kandidaten mit den zuvor gefunden Strukturen. Das Ergebnis ist eine sortierte Liste von Molekülen, die nach der Anzahl der gefundenen Substrukturen geordnet ist.

Zur Auswertung wurde ein 240 MS/MS Spektren umfassender Datensatz in verschieden große Partitionen aufgeteilt. Auf jeweils einem Teil wurde trainiert und auf dem anderen getestet. Je mehr Trainingsdaten zur Verfügung stehen, desto besser ist die Position des richtigen Kandidaten. Die durchschnittliche Laufzeit von MassStruct beträgt 330s für eine typische Anfrage in PubChem (10ms pro Kandidat). Durch

MassStruct kann die Laufzeit von MetFrag verringert werden, da das korrekte Molekül häufig eher aus der Datenbank zurückgeliefert wird und dadurch abgebrochen werden kann, sobald ein gutes Ergebnis vorliegt. Es können trotzdem alle Kandidaten prozessiert werden, sodass auch wenn keine Trainingsdaten vorhanden sind, keine Verschlechterung eintritt.

Eine weitere Verbesserung von MassStruct könnte durch ein verbessertes Training erzielt werden. Vorstellbar wäre, dass Fragmente die Peaks mit der gleichen Masse erklären, mit einem hierarchischen Clusterverfahren über die Tanimoto Ähnlichkeit der Fingerprints geclustert werden. Der resultierende Baum wird in einer bestimmten Höhe abgeschnitten und aus den immer noch zusammenhängenden Knoten kann mit Hilfe der „Maximum Common Substructure“ ein Repräsentant berechnet werden. Dadurch wäre es möglich weniger spezialisierte Fragmentstrukturen zu lernen und dadurch womöglich bessere Ergebnisse, trotz weniger Trainingsdaten, zu erzielen.

Für die auf Java Server Faces basierenden Weboberfläche von MetFrag wurden die steigenden Benutzerzahlen vorgestellt. Die Herstellersoftware Bruker Smart Formula 3D und SIRIUS besitzen eine Schnittstelle zu MetFrag. Diese kann direkt aus dem Programm angesprochen werden. Sobald das noch im Beta Stadium befindliche Smart Formula 3D an die Kunden ausgeliefert wird, ist mit voraussichtlich weiter ansteigenden Besucherzahlen zu rechnen.

Diese Arbeit liefert einen Beitrag zur Beschleunigung der Strukturaufklärung von massenspektrometrischen Daten. Die fortschreitende Entwicklung von Massenspektrometern ermöglicht eine bessere automatisierte Auswertung, die durch den erhöhten Durchsatz an Proben auch notwendig ist. Letztendlich ist die Metabolomik ein wichtiger Bestandteil der Systembiologie und hilft, in Kombination mit der Proteomik und Genomik, biologische Organismen besser zu verstehen.

6 Glossar

MS
Massenspektrometrie

MS/MS
Tandem-Massenspektrometrie

ESI
Elektrospray Ionisierung

EI
Elektronenstoßionisation

GC/MS
Gaschromatographie gekoppelt mit einem Massenspektrometer

GC/EI-MS
Gaschromatographie gekoppelt mit einem Massenspektrometer mit der Elektronenstoßionisation als Ionisierungsmethode

GC/TOF-MS
Gaschromatographie gekoppelt mit einem Massenspektrometer und einem Flugzeitanalysator

LC/MS
Flüssigchromatographie gekoppelt mit einem Massenspektrometer

HPLC
Hochleistungsflüssigchromatographie

UPLC
„Ultra Performance Liquid Chromatography“

ESI-MS/MS
Elektrospray Ionisierung in Verbindung mit einem MS/MS Instrument

ESI-MS
Elektrospray Ionisierung in Verbindung mit einem Massenspektrometer

CDK
„Chemistry Development Kit“

RDBMS
Relationales Datenbankmanagementsystem

A Anhang

Beispiel einer SD Datei

Die SD Datei von Ethanol (Beispielcode A.1) beinhaltet den Header (Zeilen 1-3), die Atome (Zeilen 4-13) und die Konnektivitätsinformationen (Zeilen 14-22), das zusammen die Moldatei ergibt. Zusätzlich zu diesen Informationen kann die SD Datei dazugehörige Daten, wie zum Beispiel eine PubChem ID (Zeilen 23-24), abspeichern.

```
1  702
2    -OEChem-10081105562D
3
4    9   8   0      0   0   0   0   0   0999 V2000
5       2.5369   -0.2500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
6       3.4030    0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
7       4.2690   -0.2500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8       3.8015    0.7249    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
9       3.0044    0.7249    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
10      3.9590   -0.7869    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
11      4.8059   -0.5600    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
12      4.5790    0.2869    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
13      2.0000    0.0600    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
14   1  2  1  0  0  0  0
15   1  9  1  0  0  0  0
16   2  3  1  0  0  0  0
17   2  4  1  0  0  0  0
18   2  5  1  0  0  0  0
19   3  6  1  0  0  0  0
20   3  7  1  0  0  0  0
21   3  8  1  0  0  0  0
22 M END
23 > <PUBCHEM_COMPOUND_CID>
24 702
```

Beispielcode A.1. Ausschnitt einer SD Datei von Ethanol (`http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=702&disopt=DisplaySDF` - Abgerufen im Oktober 2011) mit der dazugehörigen PubChem ID.

Hill Daten

Tabelle A.1. Testspektra von [HKF+08] mit dazugehörigen MassBank ID's und exakter Masse der Verbindung.

Verbindung	MassBank ID's	Masse
4-amino-Antipyrine	CO000001 CO000002 CO000003 CO000004 CO000005	203,106
6a-Methylprednisolone	CO000006 CO000007 CO000008 CO000009 CO000010	374,209
Acepromazine	CO000011 CO000012 CO000013 CO000014 CO000015	326,145
Acetophenazine	CO000016 CO000017 CO000018 CO000019 CO000020	411,198
Adenosine Diphosphate	CO000021 CO000022 CO000023 CO000024 CO000025	427,029
Adiphenine	CO000026 CO000027 CO000028 CO000029 CO000030	311,189
Albuterol	CO000031 CO000032 CO000033 CO000034 CO000035	239,152
Alfentanil	CO000036 CO000037 CO000038 CO000039 CO000040	416,254
Amfenac	CO000041 CO000042 CO000043 CO000044 CO000045	255,090
Aminophylline	CO000046 CO000047 CO000048 CO000049 CO000050	180,065
Ampicillin	CO000051 CO000052 CO000053 CO000054 CO000055	349,110
Anileridine	CO000056 CO000057 CO000058 CO000059 CO000060	352,215
Antipyrine	CO000061 CO000062 CO000063 CO000064 CO000065	188,095
Apomorphine	CO000066 CO000067 CO000068 CO000069 CO000070	267,126
Apramycin	CO000071 CO000072 CO000073 CO000074 CO000075	539,280
Betaxolol	CO000076 CO000077 CO000078 CO000079 CO000080	307,215
Boldenone Undecylenate	CO000081 CO000082 CO000083 CO000084 CO000085	452,329
Bumetanide	CO000086 CO000087 CO000088 CO000089 CO000090	364,109
Buprenorphine	CO000091 CO000092 CO000093 CO000094 CO000095	467,304
Buspirone	CO000096 CO000097 CO000098 CO000099 CO000100	385,248
Cholesterol	CO000101 CO000102 CO000103 CO000104 CO000105	386,355
Cromolyn	CO000106 CO000107 CO000108 CO000109 CO000110	468,069
Cymarin	CO000111 CO000112 CO000113 CO000114 CO000115	548,299
Daunorubicin	CO000116 CO000117 CO000118 CO000119 CO000120	527,179
Dextromethorphan	CO000121 CO000122 CO000123 CO000124 CO000125	271,194
Dihydroergotamine	CO000126 CO000127 CO000128 CO000129 CO000130	583,279
Dimefline	CO000131 CO000132 CO000133 CO000134 CO000135	323,153
Diphenoxylate	CO000136 CO000137 CO000138 CO000139 CO000140	452,246
Dobutamine	CO000141 CO000142 CO000143 CO000144 CO000145	301,168
Doxorubicin	CO000146 CO000147 CO000148 CO000149 CO000150	543,174
Drofenine	CO000151 CO000152 CO000153 CO000154 CO000155	317,236
Enalapril	CO000156 CO000157 CO000158 CO000159 CO000160	376,200
Enalaprilat	CO000161 CO000162 CO000163 CO000164 CO000165	348,169
Ephedrine	CO000166 CO000167 CO000168 CO000169 CO000170	165,115
Ergocristine	CO000171 CO000172 CO000173 CO000174 CO000175	609,295
Ergoloid Mesylate	CO000176 CO000177 CO000178 CO000179 CO000180	591,342
Etamiphylline	CO000181 CO000182 CO000183 CO000184 CO000185	279,170
Etodolac	CO000186 CO000187 CO000188 CO000189 CO000190	287,152
Fenbendazole	CO000191 CO000192 CO000193 CO000194 CO000195	299,073
Fenoterol	CO000196 CO000197 CO000198 CO000199 CO000200	303,147
Folic Acid	CO000201 CO000202 CO000203 CO000204 CO000205	441,140
Gallamine	CO000206 CO000207 CO000208 CO000209 CO000210	510,464
Gingerol	CO000211 CO000212 CO000213 CO000214 CO000215	294,183
Hematoporphyrin I	CO000216 CO000217 CO000218 CO000219 CO000220	598,279
Hydrocortisone	CO000221 CO000222 CO000223 CO000224 CO000225	362,209

Hydroxybutorphanol	CO000226	CO000227	CO000228	CO000229	CO000230	343,215
Hydroxyphenethylamine	CO000231	CO000232	CO000233	CO000234	CO000235	137,084
Isoxsuprine	CO000236	CO000237	CO000238	CO000239	CO000240	301,168
Ketorolac	CO000241	CO000242	CO000243	CO000244	CO000245	255,090
Leucine Enkephalin	CO000246	CO000247	CO000248	CO000249	CO000250	555,269
Mebeverine	CO000251	CO000252	CO000253	CO000254	CO000255	429,252
Mefenamic Acid	CO000256	CO000257	CO000258	CO000259	CO000260	241,110
Meprobamate	CO000261	CO000262	CO000263	CO000264	CO000265	218,127
Methionine Enkephalin	CO000266	CO000267	CO000268	CO000269	CO000270	573,226
Methotrexate	CO000271	CO000272	CO000273	CO000274	CO000275	454,171
Methylergonovine	CO000276	CO000277	CO000278	CO000279	CO000280	339,195
Morphine-3-Glucuronide	CO000281	CO000282	CO000283	CO000284	CO000285	461,169
Naltrexone	CO000286	CO000287	CO000288	CO000289	CO000290	341,163
Nandrolone	CO000291	CO000292	CO000293	CO000294	CO000295	274,193
Nimesulide	CO000296	CO000297	CO000298	CO000299	CO000300	308,047
Norpropoxyphene	CO000301	CO000302	CO000303	CO000304	CO000305	325,204
Noscapine	CO000306	CO000307	CO000308	CO000309	CO000310	413,147
Ormetoprim	CO000311	CO000312	CO000313	CO000314	CO000315	274,143
Oxaprozin	CO000316	CO000317	CO000318	CO000319	CO000320	293,105
Oxybutynin	CO000321	CO000322	CO000323	CO000324	CO000325	357,230
Oxycodone	CO000326	CO000327	CO000328	CO000329	CO000330	315,147
Oxytetracycline	CO000331	CO000332	CO000333	CO000334	CO000335	460,148
Perindopril	CO000336	CO000337	CO000338	CO000339	CO000340	368,231
Piperacetazine	CO000341	CO000342	CO000343	CO000344	CO000345	410,203
Poldine	CO000346	CO000347	CO000348	CO000349	CO000350	340,191
Prazosin	CO000351	CO000352	CO000353	CO000354	CO000355	383,159
Prednisolone	CO000356	CO000357	CO000358	CO000359	CO000360	360,194
Prednisolone Tebutate	CO000361	CO000362	CO000363	CO000364	CO000365	458,267
Prednisone	CO000366	CO000367	CO000368	CO000369	CO000370	358,178
Prolintane	CO000371	CO000372	CO000373	CO000374	CO000375	217,183
Pyrilamine	CO000376	CO000377	CO000378	CO000379	CO000380	285,184
Remifentanil	CO000381	CO000382	CO000383	CO000384	CO000385	376,200
Reserpine	CO000386	CO000387	CO000388	CO000389	CO000390	608,273
Rolitetracycline	CO000391	CO000392	CO000393	CO000394	CO000395	527,227
Salmeterol	CO000396	CO000397	CO000398	CO000399	CO000400	415,272
Spectinomycin	CO000401	CO000402	CO000403	CO000404	CO000405	332,158
Streptomycin	CO000406	CO000407	CO000408	CO000409	CO000410	581,266
Strychnin	CO000411	CO000412	CO000413	CO000414	CO000415	334,168
Strychnin N-oxide	CO000416	CO000417	CO000418	CO000419	CO000420	350,163
Sufentanil	CO000421	CO000422	CO000423	CO000424	CO000425	386,203
Sulfadimethoxine	CO000426	CO000427	CO000428	CO000429	CO000430	310,074
Sulfasalazine	CO000431	CO000432	CO000433	CO000434	CO000435	398,069
Taurocholate	CO000436	CO000437	CO000438	CO000439	CO000440	515,292
Tenoxicam	CO000441	CO000442	CO000443	CO000444	CO000445	337,019
Terbutaline	CO000446	CO000447	CO000448	CO000449	CO000450	225,137
Terfenadine	CO000451	CO000452	CO000453	CO000454	CO000455	471,314
Testosterone Propionate	CO000456	CO000457	CO000458	CO000459	CO000460	344,235
Tetracaine	CO000461	CO000462	CO000463	CO000464	CO000465	264,184
Tetracycline	CO000466	CO000467	CO000468	CO000469	CO000470	444,153
Tetramisole	CO000471	CO000472	CO000473	CO000474	CO000475	204,072
Theobromine	CO000476	CO000477	CO000478	CO000479	CO000480	180,065
Thiethylperazine	CO000481	CO000482	CO000483	CO000484	CO000485	399,180

Thioridazine	CO000486 CO000487 CO000488 CO000489 CO000490	370,154
Thiothixene	CO000491 CO000492 CO000493 CO000494 CO000495	443,170
Thonzide	CO000496 CO000497 CO000498 CO000499 CO000500	511,438
Tripelennamine	CO000501 CO000502 CO000503 CO000504 CO000505	255,174
Vecuronium	CO000506 CO000507 CO000508 CO000509 CO000510	557,431

Empirische Laufzeiten von MetFrag

Tabelle A.2. Die Laufzeit von MetFrag die zum Fragmentieren und Zuordnen der Fragmente zu den Peaks benötigt worden ist. Die verwendeten Spektren stammen von [HKF+08] mit den passenden Kandidaten von PubChem 2009. Die Zeit zur Vorverarbeitung wird nicht mit betrachtet.

Verbindung	**CID**	**Kand.**	**Laufzeit in s**	**Laufzeit in min**
Thonzide	5456	16	20,14	0,34
Gallamine	3450	32	27,02	0,45
Vecuronium	39765	26	58,82	0,98
Ergoloid Mesylate	592735	117	64,70	1,08
Prolintane	14592	1018	70,18	1,17
Tenoxicam	5282194	138	71,38	1,19
Tetramisole	3913	483	79,95	1,33
Hydroxyphenethylamine	5610	783	85,20	1,42
Aminophylline	2153	634	92,02	1,53
Meprobamate	4064	498	99,17	1,65
Tripelennamine	5587	614	107,37	1,79
Boldenone Undecylenate	25702	294	108,07	1,80
Theobromine	5429	634	109,04	1,82
Ergocristine	98255	217	111,52	1,86
Dihydroergotamine	3066	269	127,78	2,13
Nimesulide	4495	535	139,13	2,32
Reserpine	5052	236	139,66	2,33
Adenosine Diphosphate	197	197	142,70	2,38
Methionine Enkephalin	42785	329	149,03	2,48
Drofenine	3166	932	157,69	2,63
Ephedrine	5032	1635	160,46	2,67
Terbutaline	5403	1114	161,59	2,69
Cholesterol	304	377	162,66	2,71
Doxorubicin	1691	279	163,38	2,72
Terfenadine	5405	522	169,60	2,83
Sulfadimethoxine	5323	672	174,06	2,90
Cromolyn	2882	281	175,84	2,93
Betaxolol	2369	1102	176,65	2,94
Ketorolac	3826	1382	182,13	3,04
Hematoporphyrin I	11103	245	183,97	3,07
Dextromethorphan	3008	1187	190,79	3,18
Antipyrine	2206	1437	208,65	3,48
Daunorubicin	2958	541	214,99	3,58
Leucine Enkephalin	3903	555	216,45	3,61
Etamiphylline	28329	1234	217,52	3,63

Amfenac	2136	1382	221,35	3,69
Rolitetracycline	6420073	727	228,43	3,81
Streptomycin	19649	381	235,83	3,93
Taurocholate	8959	452	237,64	3,96
Gingerol	3473	2014	251,28	4,19
Albuterol	2083	1211	258,95	4,32
Apramycin	71428	700	271,52	4,53
Nandrolone	9904	1547	272,08	4,53
Buprenorphine	2476	557	282,28	4,70
Cymarin	539061	496	290,48	4,84
Pyrilamine	4992	2236	296,20	4,94
Mefenamic Acid	4044	2362	307,02	5,12
Mebeverine	4031	1425	318,30	5,30
Sulfasalazine	5384001	924	320,04	5,33
Buspirone	2477	1652	322,10	5,37
4-amino-Antipyrine	2151	1497	325,91	5,43
Salmeterol	5152	1206	338,23	5,64
Oxaprozin	4614	2058	338,46	5,64
Norpropoxyphene	18804	2699	338,70	5,64
Tetracaine	5411	2891	348,07	5,80
Fenbendazole	3334	1828	363,06	6,05
Apomorphine	2215	2663	364,71	6,08
Hydroxybutorphanol	3064246	2291	388,98	6,48
Oxybutynin	4634	2015	414,28	6,90
Acepromazine	6077	1823	443,71	7,40
Fenoterol	3343	3200	444,71	7,41
Testosterone Propionate	5701990	2530	453,02	7,55
Dobutamine	36811	3904	474,83	7,91
Ormetoprim	23418	2503	491,51	8,19
Adiphenine	2031	3967	535,30	8,92
Isoxsuprine	3783	3904	552,60	9,21
Alfentanil	51263	1956	555,31	9,26
Morphine-3-Glucuronide	4318740	1941	568,93	9,48
Spectinomycin	2021	2585	573,33	9,56
Enalaprilat	5362033	4603	576,36	9,61
Sufentanil	41693	3682	609,86	10,16
Anileridine	8944	5228	625,74	10,43
Etodolac	3308	4090	627,88	10,46
Poldine	11018	3981	637,02	10,62
Dimefline	3078	5329	696,82	11,61
Perindopril	107807	5203	708,93	11,82
Bumetanide	2471	5239	720,22	12,00
Enalapril	3222	4928	814,88	13,58
Prazosin	4893	3289	818,98	13,65
Prednisolone Tebutate	4898	1716	832,61	13,88
Remifentanil	60815	4928	958,32	15,97
Piperacetazine	19675	5181	967,87	16,13
Ampicillin	2174	4692	982,45	16,37
Acetophenazine	441185	5300	1067,37	17,79
Methylergonovine	4140	6004	1138,53	18,98
Diphenoxylate	13505	4029	1166,43	19,44
Oxytetracycline	5280972	3572	1303,73	21,73

Prednisolone	4894	3597	1337,35	22,29
Thiethylperazine	5440	7644	1356,78	22,61
Oxycodone	4635	4743	1368,19	22,80
Tetracycline	5353990	3818	1377,09	22,95
Hydrocortisone	3640	3938	1427,62	23,79
Prednisone	4900	3615	1445,09	24,08
Thiothixene	941651	5808	1566,97	26,12
Folic Acid	3405	7399	1579,81	26,33
Noscapine	4544	3609	1586,00	26,43
Strychnine	5304	5414	1668,30	27,80
6a-Methylprednisolone	4159	3727	1675,52	27,93
Thioridazine	5452	10487	1814,95	30,25
Methotrexate	4112	9054	1894,25	31,57
Naltrexone	4428	8827	2334,17	38,90
Strychnine N-oxide	73393	10741	4103,26	68,39
Durchschnitt:		**2544,19**	**577,80**	**9,63**

MetFrag - Hill Daten mit PubChem 2009

Tabelle A.3. MetFrag (Baumtiefe 1 und 2) mit vorverabeiteten Bindungen (siehe Abschnitt 3.1.2) und der Scoring Funktion aus Abschnitt 3.1.6. Die MS/MS Daten stammen von [HKF$^+$08]. Es wurde PubChem vom Juni 2009 verwendet, um Kandidaten mit passender exakter Masse (10 ppm Abweichung) zu finden. Der durchschnittliche RRP beträgt $0,084$ bzw. 0,104 (BT 2).

Verbindung	**CID**	**Kand.**	**Rang BT1**	**Tan. Rang BT1**	**Tan. Rang BT2**
4-amino-Antipyrine	2151	1497	339	281	186
6a-Methylprednisolone	4159	3727	660	415	248
Acepromazine	6077	1823	11	11	7
Acetophenazine	441185	5300	2	2	2
Adenosine Diphosphate	197	197	17	2	2
Adiphenine	2031	3967	15	11	45
Albuterol	2083	1211	159	98	34
Alfentanil	51263	1956	2	2	29
Amfenac	2136	1382	146	121	113
Aminophylline	2153	634	6	6	13
Ampicillin	2174	4692	20	1	6
Anileridine	8944	5228	3	3	917
Antipyrine	2206	1437	526	454	135
Apomorphine	2215	2663	14	10	19
Apramycin	71428	700	12	1	2
Betaxolol	2369	1102	51	32	13
Boldenone Undecylenate	25702	294	54	34	3
Bumetanide	2471	5239	6	6	8
Buprenorphine	2476	557	71	46	96
Buspirone	2477	1652	53	32	117
Cholesterol	304	377	241	30	25
Cromolyn	2882	281	9	9	7

Cymarin	539061	496	98	44	162
Daunorubicin	2958	541	36	5	7
Dextromethorphan	3008	1187	268	156	250
Dihydroergotamine	3066	269	17	2	7
Dimefline	3078	5329	177	130	20
Diphenoxylate	13505	4029	2	2	3
Dobutamine	36811	3904	8	2	152
Doxorubicin	1691	279	37	4	5
Drofenine	3166	932	6	4	2
Enalapril	3222	4928	40	24	5
Enalaprilat	5362033	4603	10	6	15
Ephedrine	5032	1635	138	67	244
Ergocristine	98255	217	15	1	2
Ergoloid Mesylate	592735	117	1	1	2
Etamiphylline	28329	1234	100	64	56
Etodolac	3308	4090	23	17	61
Fenbendazole	3334	1828	10	10	1
Fenoterol	3343	3200	10	4	3
Folic Acid	3405	7399	19	17	42
Gallamine	3450	32	2	2	1
Gingerol	3473	2014	18	14	32
Hematoporphyrin I	11103	245	106	72	63
Hydrocortisone	3640	3938	918	608	190
Hydroxybutorphanol	3064246	2291	11	7	175
Hydroxyphenethylamine	5610	783	28	11	63
Isoxsuprine	3783	3904	17	5	206
Ketorolac	3826	1382	9	6	3
Leucine Enkephalin	3903	555	12	2	6
Mebeverine	4031	1425	11	6	1
Mefenamic Acid	4044	2362	153	130	261
Meprobamate	4064	498	62	55	8
Methionine Enkephalin	42785	329	1	1	8
Methotrexate	4112	9054	6	2	53
Methylergonovine	4140	6004	9	1	3
Morphine-3-Glucuronide	4318740	1941	9	2	17
Naltrexone	4428	8827	1893	1103	2995
Nandrolone	9904	1547	243	169	129
Nimesulide	4495	535	3	3	1
Norpropoxyphene	18804	2699	19	13	33
Noscapine	4544	3609	339	211	605
Ormetoprim	23418	2503	205	150	4
Oxaprozin	4614	2058	276	239	101
Oxybutynin	4634	2015	4	2	1
Oxycodone	4635	4743	718	477	1716
Oxytetracycline	5280972	3572	32	4	10
Perindopril	107807	5203	28	13	47
Piperacetazine	19675	5181	1	1	1
Poldine	11018	3981	10	5	9
Prazosin	4893	3289	11	10	50
Prednisolone	4894	3597	498	338	210
Prednisolone Tebutate	4898	1716	10	4	35
Prednisone	4900	3615	608	414	749

Prolintane	14592	1018	20	13	26
Pyrilamine	4992	2236	20	8	11
Remifentanil	60815	4928	3	3	7
Reserpine	5052	236	42	16	33
Rolitetracycline	6420073	727	11	3	5
Salmeterol	5152	1206	28	19	16
Spectinomycin	2021	2585	56	30	19
Streptomycin	19649	381	17	2	5
Strychnine	5304	5414	4990	3261	3505
Strychnine N-oxide	73393	10741	8747	5386	5868
Sufentanil	41693	3682	1	1	1
Sulfadimethoxine	5323	672	4	4	2
Sulfasalazine	5384001	924	5	4	5
Taurocholate	8959	452	52	2	13
Tenoxicam	5282194	138	9	6	5
Terbutaline	5403	1114	153	102	8
Terfenadine	5405	522	4	1	4
Testosterone Propionate	5701990	2530	208	109	114
Tetracaine	5411	2891	248	151	2
Tetracycline	5353990	3818	136	41	83
Tetramisole	3913	483	101	80	239
Theobromine	5429	634	66	57	81
Thiethylperazine	5440	7644	3	3	1
Thioridazine	5452	10487	4	2	1
Thiothixene	941651	5808	3	1	1
Thonzide	5456	16	1	1	1
Tripelennamine	5587	614	1	1	14
Vecuronium	39765	26	11	1	1
Durchschnitt:		**2544,19**	**241,6 ±99,8**	**152,5 ±62,4**	**205,1 ±74,3**
Standardabweichung:		**2365,65**	**1007,97**	**630,49**	**749,92**
Median:		**1825,5**	**18,5**	**9,5**	**14,5**
75% Quantil:		**3904**	**104,75**	**62,25**	**92,75**

MetFrag (BDE) - Hill Daten mit PubChem 2009

Tabelle A.4. MetFrag mit Bindungsdissoziationsenergien [WSMHN10] und den MS/MS Daten von [HKF$^+$08]. Es wurde PubChem vom Juni 2009 verwendet, um Kandidaten mit passender exakter Masse (10 ppm Abweichung) zu finden. Der durchschnittliche RRP beträgt $0,089$.

Verbindung	**CID**	**Kand.**	**Rang**	**Tan. Rang**
4-amino-Antipyrine	2151	1496	146	130
6a-Methylprednisolone	4159	3727	58	18
Acepromazine	6077	1822	15	13
Acetophenazine	441185	5300	2	2
Adenosine Diphosphate	197	197	19	3
Adiphenine	2031	3967	7	6
Albuterol	2083	1211	325	174
Alfentanil	51263	1959	2	2

Amfenac	2136	1382	117	94
Aminophylline	2153	634	9	9
Ampicillin	2174	4692	83	38
Anileridine	8944	5228	13	8
Antipyrine	2206	1435	503	446
Apomorphine	2215	2566	97	70
Apramycin	71428	736	12	1
Betaxolol	2369	1102	7	5
Boldenone Undecylenate	25702	294	27	16
Bumetanide	2471	5242	5	5
Buprenorphine	2476	558	60	34
Buspirone	2477	1685	41	24
Cholesterol	304	377	288	47
Cromolyn	2882	281	8	8
Cymarin	539061	496	68	21
Daunorubicin	2958	541	33	5
Dextromethorphan	3008	1187	131	79
Dihydroergotamine	3066	269	18	3
Dimefline	3078	4550	1609	1056
Diphenoxylate	13505	4024	2	2
Dobutamine	36811	3916	43	21
Doxorubicin	1691	280	41	10
Drofenine	3166	933	40	32
Enalapril	3222	4927	171	85
Enalaprilat	5362033	4603	49	29
Ephedrine	5032	1636	53	22
Ergocristine	98255	217	53	28
Ergoloid Mesylate	592735	119	1	1
Etamiphylline	28329	1234	6	6
Etodolac	3308	4091	7	5
Fenbendazole	3334	1826	88	75
Fenoterol	3343	3207	10	4
Folic Acid	3405	7401	19	15
Gallamine	3450	32	1	1
Gingerol	3473	2014	28	14
Hematoporphyrin I	11103	245	47	36
Hydrocortisone	3640	3938	96	23
Hydroxybutorphanol	3064246	2294	6	3
Hydroxyphenethylamine	5610	783	80	63
Isoxsuprine	3783	3916	160	79
Ketorolac	3826	1382	37	31
Leucine Enkephalin	3903	555	13	5
Mebeverine	4031	1430	4	2
Mefenamic Acid	4044	2362	451	372
Meprobamate	4064	498	13	10
Methionine Enkephalin	42785	329	12	10
Methotrexate	4112	9053	11	7
Methylergonovine	4140	6004	9	1
Morphine-3-Glucuronide	4318740	1941	466	248
Naltrexone	4428	8827	114	68
Nandrolone	9904	1547	150	90
Nimesulide	4495	535	30	27

Norpropoxyphene	18804	2699	27	19
Noscapine	4544	3609	562	337
Ormetoprim	23418	2502	1441	1117
Oxaprozin	4614	2058	487	357
Oxybutynin	4634	2024	9	5
Oxycodone	4635	4729	146	81
Oxytetracycline	5280972	3572	37	6
Perindopril	107807	3041	18	4
Piperacetazine	19675	5181	1	1
Poldine	11018	3997	13	9
Prazosin	4893	2385	174	127
Prednisolone	4894	3598	62	27
Prednisolone Tebutate	4898	1716	12	9
Prednisone	4900	3615	138	74
Prolintane	14592	1018	106	64
Pyrilamine	4992	2236	8	6
Remifentanil	60815	4927	21	14
Reserpine	5052	237	52	27
Rolitetracycline	6420073	727	10	2
Salmeterol	5152	1206	22	15
Spectinomycin	2021	2584	17	1
Streptomycin	19649	374	43	25
Strychnine	5304	5103	1525	997
Strychnine N-oxide	73393	10782	5751	3632
Sufentanil	41693	3734	13	11
Sulfadimethoxine	5323	675	3	3
Sulfasalazine	5384001	926	11	8
Taurocholate	8959	458	46	7
Tenoxicam	5282194	138	9	6
Terbutaline	5403	1390	239	162
Terfenadine	5405	538	20	14
Testosterone Propionate	5701990	2554	65	32
Tetracaine	5411	2937	5	5
Tetracycline	5353990	3825	92	10
Tetramisole	3913	483	353	309
Theobromine	5429	634	249	61
Thiethylperazine	5440	7668	3	3
Thioridazine	5452	10504	4	2
Thiothixene	941651	5827	3	1
Thonzide	5456	16	1	1
Tripelennamine	5587	626	49	36
Vecuronium	39765	31	20	7
Durchschnitt:		**2509**	**175,3 ±61,5**	**111,5 ±39,7**
Standardabweichung:		**2339,44**	**620,77**	**400,51**
Median:		**1824**	**31,5**	**14,5**
75% Quantil:		**3893,25**	**96,75**	**62,5**

Vergleich von MetFrag mit MassFrontier - Hill Daten mit PubChem 2006

Tabelle A.5. Vergleich von MassFrontier 4 [HKF+08] mit MetFrag. Der PubChem Datenbestand vom Februar 2006 ist verwendet worden. Der durchschnittliche RRP von MetFrag ist mit 0,0952 besser als der von MassFrontier mit 0,1180.

		MassFrontier		MetFrag		
Verbindung	**CID**	**Kand.**	**Rang**	**Kand.**	**Rang**	**Tan. Rang**
4-amino-Antipyrine	2151	226	16	273	40	34
6a-Methylprednisolone	4159	226	11	295	39	28
Acepromazine	6077	281	4	338	6	6
Acetophenazine	441185	435	1	544	2	2
Adenosine Diphosphate	197	32	3	46	6	1
Adiphenine	2031	623	6	796	11	7
Albuterol	2083	143	15	205	21	13
Alfentanil	51263	134	1	162	1	1
Amfenac	2136	344	11	380	41	36
Aminophylline	2153	94	21	173	2	2
Ampicillin	2174	615	1	780	5	1
Anileridine	8944	563	251	666	2	2
Antipyrine	2206	306	97	341	123	98
Apomorphine	2215	453	12	639	4	2
Apramycin	71428	54	1	58	3	1
Betaxolol	2369	190	5	259	29	16
Boldenone Undecylenate	25702	21	2	32	3	3
Bumetanide	2471	619	10	769	2	2
Buprenorphine	2476	40	2	49	14	6
Buspirone	2477	36	1	29	4	3
Cholesterol	304	52	52	78	44	7
Cromolyn	2882	33	2	37	4	4
Cymarin	539061	61	8	83	20	5
Daunorubicin	2958	110	12	129	18	4
Dextromethorphan	3008	166	23	237	52	29
Dihydroergotamine	3066	35	1	38	2	1
Dimefline	3078	644	644	1000	32	26
Diphenoxylate	13505	333	4	369	1	1
Dobutamine	36811	447	44	642	5	2
Doxorubicin	1691	60	3	81	17	2
Drofenine	3166	117	4	148	2	2
Enalapril	3222	246	1	286	6	2
Enalaprilat	5362033	370	2	455	3	2
Ephedrine	5032	246	5	307	26	15
Ergocristine	98255	16	1	27	6	1
Ergoloid Mesylate	592735	7	1	10	1	1
Etamiphylline	28329	100	3	104	4	4
Etodolac	3308	420	1	579	1	1
Fenbendazole	3334	403	92	479	3	3
Fenoterol	3343	370	5	519	5	1
Folic Acid	3405	602	13	826	4	2

Gallamine	3450	10	1	8	1	1
Gingerol	3473	182	2	196	3	2
Hematoporphyrin I	11103	42	33	45	21	18
Hydrocortisone	3640	260	4	301	97	62
Hydroxybutorphanol	3064246	180	2	201	2	2
Hydroxyphenethylamine	5610	166	166	173	14	10
Isoxsuprine	3783	447	5	642	4	2
Ketorolac	3826	344	37	380	5	4
Leucine Enkephalin	3903	53	2	60	3	1
Mebeverine	4031	96	2	75	2	2
Mefenamic Acid	4044	579	328	631	31	29
Meprobamate	4064	85	19	84	12	12
Methionine Enkephalin	42785	66	1	68	2	1
Methotrexate	4112	644	116	763	5	2
Methylergonovine	4140	515	1	629	6	1
Morphine-3-Glucuronide	4318740	179	2	170	2	1
Naltrexone	4428	1035	34	1421	359	196
Nandrolone	9904	124	18	129	29	20
Nimesulide	4495	136	136	148	2	2
Norpropoxyphene	18804	392	15	476	6	5
Noscapine	4544	275	3	363	54	22
Ormetoprim	23418	270	124	317	25	22
Oxaprozin	4614	461	101	607	76	60
Oxybutynin	4634	114	6	155	3	1
Oxycodone	4635	776	102	997	136	92
Oxytetracycline	5280972	483	4	617	11	1
Perindopril	107807	102	2	119	5	3
Piperacetazine	19675	494	1	625	1	1
Poldine	11018	682	19	493	3	3
Prazosin	4893	185	4	352	2	2
Prednisolone	4894	269	13	362	52	39
Prednisolone Tebutate	4898	143	4	165	3	1
Prednisone	4900	344	6	418	76	50
Prolintane	14592	105	9	118	7	6
Pyrilamine	4992	268	1	296	1	1
Remifentanil	60815	246	1	286	1	1
Reserpine	5052	28	3	31	10	5
Rolitetracycline	6420073	105	1	149	6	1
Salmeterol	5152	32	1	37	2	2
Spectinomycin	2021	310	1	360	6	5
Streptomycin	19649	37	1	43	5	1
Strychnine	5304	664	575	882	824	540
Strychnine N-oxide	73393	1185	1098	1667	1338	830
Sufentanil	41693	445	1	496	1	1
Sulfadimethoxine	5323	94	18	145	4	4
Sulfasalazine	5384001	106	5	116	3	2
Taurocholate	8959	59	4	65	10	1
Tenoxicam	5282194	28	1	34	3	1
Terbutaline	5403	175	31	225	25	19
Terfenadine	5405	34	1	35	1	1
Testosterone Propionate	5701990	134	3	183	15	5
Tetracaine	5411	308	22	362	33	25

Tetracycline	5353990	529	5	675	43	13
Tetramisole	3913	120	1	122	23	19
Theobromine	5429	94	14	173	18	17
Thiethylperazine	5440	569	2	670	2	2
Thioridazine	5452	849	1	1091	1	1
Thiothixene	941651	726	1	909	3	1
Thonzide	5456	4	1	4	1	1
Tripelennamine	5587	97	3	102	1	1
Vecuronium	39765	3	1	4	3	1
Durchschnitt:		**272,2±24,2**	**44,2±14,1**	**340,6±31,8**	**39,8±15,6**	**25,1±9,8**
Standardabweichung:		**244,17**	**142,47**	**320,67**	**157,95**	**99,25**
Median:		**183,5**	**4**	**248**	**5**	**2**
75% Quantil:		**431,25**	**17,5**	**513,25**	**21**	**13**

MetFrag mit simulierten Nominalmassen - Hill Daten

Tabelle A.6. MetFrag mit Vorverarbeitung und den MS/MS Daten von [HKF+08]. Es wurde PubChem vom Juni 2009 verwendet, um Kandidaten mit passender exakter Masse (10 ppm Abweichung) zu finden. Die erlaubte Abweichung zum Zurodnen von Peak und Fragment wurde auf 0,5 Da festgelegt, um zu simulieren, dass das Massenspektrometer nur Nominalmassen messen kann.

Verbindung	CID	Kandidaten	Rang	Tan. Rang
4-amino-Antipyrine	2151	1497	786	650
6a-Methylprednisolone	4159	3722	2632	1692
Acepromazine	6077	1823	877	590
Acetophenazine	441185	5300	311	203
Adenosine Diphosphate	197	197	19	3
Adiphenine	2031	3967	680	458
Albuterol	2083	1211	280	172
Alfentanil	51263	1956	120	68
Amfenac	2136	1382	40	40
Aminophylline	2153	634	157	147
Ampicillin	2174	4691	13	2
Anileridine	8944	5228	7	6
Antipyrine	2206	1437	929	821
Apomorphine	2215	2663	2110	1540
Apramycin	71428	700	14	2
Betaxolol	2369	1102	32	22
Boldenone Undecylenate	25702	294	42	27
Bumetanide	2471	5238	754	554
Buprenorphine	2476	557	544	349
Buspirone	2477	1652	60	44
Cholesterol	304	377	107	17
Cromolyn	2882	281	10	10
Cymarin	539061	496	147	73
Daunorubicin	2958	541	191	95
Dextromethorphan	3008	1187	354	190

Dihydroergotamine	3066	269	25	9
Dimefline	3078	5328	32	17
Diphenoxylate	13505	4028	2210	1339
Dobutamine	36811	3904	24	13
Doxorubicin	1691	279	133	72
Drofenine	3166	932	41	25
Enalapril	3222	4928	115	63
Enalaprilat	5362033	4603	7	4
Ephedrine	5032	1635	60	27
Ergocristine	98255	217	41	23
Ergoloid Mesylate	592735	117	4	3
Etamiphylline	28329	1234	1029	621
Etodolac	3308	4089	413	287
Fenbendazole	3334	1828	499	404
Fenoterol	3343	3200	9	3
Folic Acid	3405	7398	945	631
Gallamine	3450	32	20	18
Gingerol	3473	2014	69	42
Hematoporphyrin I	11103	245	161	116
Hydrocortisone	3640	3935	3445	2221
Hydroxybutorphanol	3064246	2290	682	432
Hydroxyphenethylamine	5610	783	28	14
Isoxsuprine	3783	3904	28	7
Ketorolac	3826	1382	10	8
Leucine Enkephalin	3903	555	5	1
Mebeverine	4031	1425	13	8
Mefenamic Acid	4044	2362	256	221
Meprobamate	4064	498	183	130
Methionine Enkephalin	42785	329	4	3
Methotrexate	4112	9053	218	125
Methylergonovine	4140	6002	817	534
Morphine-3-Glucuronide	4318740	1941	116	89
Naltrexone	4428	8824	3172	1865
Nandrolone	9904	1547	958	640
Nimesulide	4495	535	289	263
Norpropoxyphene	18804	2699	392	254
Noscapine	4544	3605	1042	559
Ormetoprim	23418	2503	397	293
Oxaprozin	4614	2058	115	97
Oxybutynin	4634	2015	567	324
Oxycodone	4635	4738	3503	2308
Oxytetracycline	5280972	3568	286	165
Perindopril	107807	5202	11	3
Piperacetazine	19675	5181	8	5
Poldine	11018	3981	63	46
Prazosin	4893	3288	80	57
Prednisolone	4894	3597	1492	1042
Prednisolone Tebutate	4898	1716	341	224
Prednisone	4900	3598	2329	1618
Prolintane	14592	1018	106	43
Pyrilamine	4992	2236	61	46
Remifentanil	60815	4927	53	31

Reserpine	5052	236	32	12
Rolitetracycline	6420073	727	37	21
Salmeterol	5152	1206	52	37
Spectinomycin	2021	2585	487	277
Streptomycin	19649	381	27	12
Strychnin	5304	5408	5332	3472
Strychnin N-oxide	73393	10427	10395	6110
Sufentanil	41693	3681	15	8
Sulfadimethoxine	5323	672	7	7
Sulfasalazine	5384001	924	5	4
Taurocholate	8959	452	62	9
Tenoxicam	5282194	138	30	19
Terbutaline	5403	1114	476	341
Terfenadine	5405	522	4	1
Testosterone Propionate	5701990	2529	1649	812
Tetracaine	5411	2891	839	444
Tetracycline	5353990	3801	491	303
Tetramisole	3913	483	101	76
Theobromine	5429	634	499	293
Thiethylperazine	5440	7625	835	452
Thioridazine	5452	10487	689	434
Thiothixene	941651	5806	519	356
Thonzide	5456	16	1	1
Tripelennamine	5587	614	71	50
Vecuronium	39765	26	25	12
Durchschnitt:		**2540,13**	**596,1 ±130,3**	**379,7 ±80**
Standardabweichung:		**2354,03**	**1315,90**	**807,79**
Median:		**1825,5**	**118**	**74,5**
75% Quantil:		**3904**	**561,25**	**392**

MetFrag Ergebnisse mit GC/MS Daten

Tabelle A.7. Ergebnisse von MetFrag mit den GC/MS Testdatensatz von [SMB09]. Der durchschnittliche RRP über alle 100 Spektren beträgt 0,3430 (BT1) bzw. 0,3803 (BT2). Die Kandidaten sind vorverarbeitet (siehe Abschnitt 3.1.2) und mit MetFrag prozessiert worden. Fett gedruckte Nummern sind Spektren, die in dem kleineren Datensatz verwendet worden sind (siehe Tabelle 4.4). R - Rang; LZ - Laufzeit; BT - Baumtiefe

NIST ID	Summenformel	Kand.	Rang BT1	LZ BT1	Rang BT2	LZ BT2
61627	C9H16	1901	733	219,74	118	160,444
26708	C8H17N1	2258	477	436,434	32	386,689
113790	C9H20O1	405	57	120,362	182	53,307
158384	C7H14	55	43	32,201	26	14,219
38909	C10H18	5567	687	410,744	1861	346,77
61924	C10H20	851	310	167,376	644	97,734
60708	C8H12	2081	160	231,472	484	119,446
1911	C6H12O2	1313	52	326,09	31	177,689
61640	C13H28	801	59	41,7	405	90,819
4617	C1N3F5	11	5	14,629	5	13,679

194167	C4H8N2O1	6754	1168	650,25	4113	1036,27
186524	C6H9O1Br1	3703	123	309,207	1625	483,84
38120	C1H5Si1Br1	2	2	12,472	2	11,748
146109	C4H2N2F1Cl1	6393	1491	837,046	2820	700,998
73456	C5H11Br1	8	4	13,102	4	12,839
61694	C9H14	7242	1069	253,97	7140	484,514
42198	C6H11O1Br1	1115	149	66,264	519	107,419
109982	C4H7Si1Cl3	729	124	58,358	557	59,693
120	C2H3N1O1	26	17	14,604	17	14,032
154091	C8H14	653	406	45,438	315	52,091
71109	C6H14N2	2338	899	250,335	1214	378,874
162833	C10H18	5567	1036	260,16	1035	560,887
249757	C5H9N1	313	72	40,539	127	37,138
3238	C5H10O2S1	4560	633	488,356	3354	402,191
113090	C8H14	653	441	32,574	375	41,295
63698	C3H4N2O1	1371	101	151,917	1188	136,404
74975	C6H12O3	6171	2119	731,185	22	764,185
185578	C5H10O4	5841	2116	970,678	342	1251,724
61113	C10H20	851	192	81,507	457	117,911
160559	C4H13N1P2	396	209	75,678	169	91,282
46389	C5H10O3	1656	41	184,196	150	164,773
46612	C9H18O1	4745	2313	379,718	10	407,675
105465	C7H16Si1	889	8	66,032	231	67,23
61433	C11H24	158	66	25,348	21	53,769
113438	C8H16	138	129	25,529	31	51,605
215368	C6H10O1	747	541	61,993	263	63,298
20664	C9H20	34	16	13,935	14	24,02
62859	C8H14	653	515	50,147	82	54,468
69684	C11H24O1	2426	2	207,772	697	334,523
629	C5H13N1	17	3	14,917	5	14,03
152851	C4H7O2Cl1	487	100	69,394	150	81,333
114082	C6H14O1	32	15	16,317	10	15,14
196609	C5H11N1O2	6418	3691	765,784	3607	983,766
204405	C9H14	7243	3931	289,997	5262	338,488
28546	C5H12O2	69	21	29,152	34	28,063
113901	C9H16	1901	198	92,877	1869	149,034
193841	C6H16O1Si1	425	102	57,869	81	67,698
604	C4H6O2	263	122	42,674	10	39,995
73972	C9H21N1O1	7769	3269	708,978	2119	741,312
63639	C2H6O2	5	2	12,118	2	12,213
135135	C4H8N1O1Cl1	1371	435	170,049	193	228,347
63008	C5H6	39	6	13,703	2	13,282
61471	C13H28	801	316	80,961	525	127,991
60569	C8H17Cl1	89	44	20,238	7	39,784
41785	C8H16O1	1684	400	106,89	143	119,171
66064	C9H14	7243	2749	219,206	5211	239,924
160476	C6H10O1	747	292	62,359	105	69,305
73870	C8H12	2081	155	62,024	1091	60,988
108516	C4H12N2	38	14	26,402	31	26,983
4169	C3H3Cl3	8	8	13,029	4	12,454
46224	C5H13N1	17	14	13,514	10	14,064
158830	C7H9Br1	2732	531	160,807	239	234,781

61715	C8H14	653	340	44,795	428	54,694
1123	C4H4O3	1073	64	101,803	72	101,088
156613	C9H22N1P1	9660	201	883,76	2	1791,093
176	C2H7P1	2	2	11,672	2	11,813
114550	C7H14O1	596	260	60,659	166	70,603
214253	C5H13N1O1	149	25	36,525	80	35,296
70751	C7H19N3	4238	712	533,813	1184	1145,338
62909	C6H12O1	211	98	39,709	19	37,697
37206	C7H13N1	3809	2957	280,484	2897	360,779
229049	C4H11N1O1	56	26	18,081	7	17,503
19272	C6H10	76	38	15,302	45	17,069
831	C2N1F3	5	1	12,31	1	12,072
114407	C7H12	221	80	22,385	93	20,413
5393	C4H6O2Cl2	1131	94	120,04	171	224,584
30409	C5H18Si3	521	324	58,638	225	91,653
60785	C9H20O1	405	44	41,639	277	93,971
72642	C9H22N2	4994	3973	412,893	1593	492,881
118272	C3H7N1O1	84	53	19,913	12	29,668
108346	C3H7O2Br1	38	3	15,683	13	26,819
26687	C8H14	653	517	45,764	337	58,074
113772	C7H14O1	596	49	60,81	1	83,299
1614	C8H16	138	31	20,092	45	27,213
107506	C9H19F1	211	1	33,869	11	52,699
98625	C6H14Si1	314	35	31,445	107	45,666
1908	C6H12O2	1313	990	122,792	46	249,093
134724	C3H4N1S1Br1	480	98	61,686	141	65,72
50930	C9H18	337	6	23,306	43	35,292
64555	C5H10N2	2668	1374	170,209	1534	209,514
113750	C9H20O1	405	96	40,403	164	57,396
114530	C8H16O1	1684	497	124,314	1326	198,587
61453	C12H24	5511	4	249,111	33	787,438
37233	C9H16	1901	100	64,897	975	188,779
60877	C12H24	5512	117	235,572	1055	594,571
63617	C3H4O1	13	9	12,919	10	12,957
72945	C4H5O1Cl1	175	82	35,002	104	36,479
113601	C12H24	5512	91	204,344	366	546,487
52322	C5H13N3	4054	1438	333,255	2368	498,184
215367	C6H8O1	1623	1268	82,459	1350	118,947
Durchschnitt		**1838,76**	**511 ±89**	**160,91**	**687 ±125,8**	**215,87**
Standardabweichung:		**2334,45**	**890,09**	**212,51**	**1258,40**	**309,67**
Median:		**738**	**109,5**	**62,1915**	**150**	**82,32**
75% Quantil:		**2360**	**515,5**	**219,3395**	**657,25**	**242,22**

Komplette MassStruct SQL Anfrage

```
1  SELECT substance.accession, Score
2  FROM substance, compound, library,
3
4      (SELECT inchi_key_1,
5       COUNT(DISTINCT PeakFragments.mz_cluster_id) AS Score
6       FROM substance, library, compound AS Candidates
7
8          (SELECT MIN(compound_id)
9           FROM compound
10          WHERE exact_mass BETWEEN 290.24 AND 290.28
11          GROUP BY inchi_key_1) AS FirstCandidates
12          LEFT OUTER JOIN
13          (SELECT fragments.structure, mz_cluster_id
14            FROM fragments, mz_cluster
15            WHERE fragments.mz_cluster_id = mz_cluster.id
16            AND    ((mz_cluster.mass between 123.035 AND 123.045)
17                  OR (mz_cluster.mass between 139.030 AND 139.040)
18                  OR (mz_cluster.mass between 165.040 AND 165.050)
19                  OR (mz_cluster.mass between 207.050 AND 207.060)
20                  OR (mz_cluster.mass between 249.060 AND 249.070)
21                  OR (mz_cluster.mass between 273.060 AND 273.070)))
22            AS PeakFragments
23           ON (PeakFragments.structure <= Candidates.mol_structure)
24
25      WHERE substance.compound_id = FirstCandidates.compound_id
26      AND substance.library_id  = library.library_id
27      AND library_name = 'pubchem'
28      AND Candidates.compound_id = FirstCandidates.compound_id
29      GROUP BY accession, inchi_key_1
30      ORDER BY Score DESC) AS Results
31
32 WHERE exact_mass BETWEEN 290.24 AND 290.28
33 AND substance.compound_id = compound.compound_id
34 AND substance.library_id  = library.library_id
35 AND library_name = 'pubchem'
36 AND compound.inchi_key_1 = results.inchi_key_1;
```

Beispielcode A.2. Ausführliche SQL Query von MassStruct. In dieser Abfrage wird zusätzlich berücksichtigt, dass Kandidaten aus unterschiedlichen Moleküldatenbanken vorhanden sind. Außerdem werden die Stereoisomere durch den ersten Teil des InChIKeys (Konnektivität) gefiltert und nur der zuerst auftretende verwendet.

MassStruct Ergebnisse der 1:1 Partitionierung

Tabelle A.8. Die Trainingsdaten sind zufällig in zwei gleichgroße Parititionen geteilt worden. Das MassStruct Training ist zuerst auf dem Einen durchgeführt und auf dem Anderen getestet worden. Im zweiten Schritt sind die Datensätze zum Trainieren und Testen getauscht worden. Die Angegebene Laufzeit beinhaltet nur die Zeit der SQL-Anfrage.

MassBank ID	CID	RRP	BC	WC	TC	Laufzeit
PR100121PR100122	13804	0,000	0	50902	50910	477
PR100113PR100114	834	0,000	0	34475	34491	283
PR100317	13804	0,001	0	50852	50910	448
PR100239	5319853	0,001	0	72031	72127	867
PR100390	165627	0,001	5	8735	8743	100
PR100329	717531	0,001	0	29768	29828	252
PR100198	439155	0,001	48	105480	105691	950
PR101031	5274585	0,001	16	46540	46649	526
PR100359	6441269	0,001	35	71951	72127	793
PR100296	92136	0,002	7	8717	8743	84
PR100076	34755	0,002	12	107657	108138	932
PR100363	442456	0,002	12	9711	9745	200
PR100277	160556	0,003	98	63536	63768	497
PR100249	92794	0,003	35	80451	80855	898
PR100447	5320863	0,003	106	70812	71109	854
PR100395	65065	0,003	17	11783	11839	113
PR100256	5280459	0,003	118	69911	70272	587
PR101033	5280459	0,003	118	69911	70272	893
PR101047	5280459	0,003	118	69911	70272	423
PR100386	1029	0,004	19	11782	11855	140
PR101022	5318759	0,004	153	54264	54545	687
PR100001PR100002	2901	0,004	3	4936	4974	26
PR100351	101781	0,004	209	71749	72127	1147
PR100248	5281673	0,004	81	59419	59828	879
PR101027	5281673	0,005	196	59464	59828	469
PR101007	5316673	0,005	285	78101	78553	639
PR101024	5316673	0,005	313	78082	78553	1020
PR100243	5282102	0,005	256	69786	70272	1327
PR100253	5281643	0,006	208	59338	59828	405
PR101012	5281643	0,006	232	59356	59828	2163
PR101021	5481882	0,007	178	89873	90914	1483
PR100240	5318645	0,007	232	55942	56492	738
PR101025	5282102	0,007	351	69641	70272	832
PR100335	65127	0,007	161	36266	36630	350
PR100254	5280804	0,008	327	59166	59828	318
PR100469	25674	0,009	39	8969	9090	97
PR100475	5281417	0,010	220	80697	82100	803
PR100175	6288	0,010	11	2027	2059	25
PR100315	99289	0,010	11	2027	2059	16
PR100366	5321576	0,010	120	26692	27134	425
PR100334	439574	0,012	5	2970	3036	28
PR100449	441031	0,012	5	2970	3036	25
PR100326	123938	0,012	427	41563	42144	400
PR100367	5321577	0,012	265	26740	27134	232

PR100436PR100437	439227	0,013	23	4863	4974	47
PR101030	5481224	0,018	351	68928	71109	436
PR100314	440018	0,018	1203	92875	95106	751
PR100221	137	0,019	10	4160	4316	6
PR100137	14982	0,019	0	1245	1296	17
PR100258	14982	0,019	0	1245	1296	22
PR100166	439579	0,020	512	47480	48887	443
PR100035PR100036	6057	0,020	0	13183	13733	55
PR100157PR100158	6322	0,021	165	14779	15256	184
PR100303	6322	0,021	165	14779	15256	160
PR100368	5320686	0,021	0	7828	8174	47
PR100163PR100164	5961	0,022	134	8505	8758	86
PR101023	5320686	0,022	14	7822	8174	119
PR100252	5484066	0,023	61	7750	8057	131
PR100322	5962	0,024	143	8158	8417	85
PR100153PR100154	439277	0,025	0	309	326	6
PR100420	70914	0,025	329	15535	15995	45
PR100093	637775	0,025	468	31226	32403	294
PR100349	5883291	0,025	84	8449	8815	183
PR100241	5481663	0,025	10	5710	6007	90
PR100244	5318767	0,026	21	8383	8815	173
PR100354	10621	0,028	12	7509	7950	125
PR100304PR100305	439232	0,029	280	14931	15546	86
PR100267PR100268	6950385	0,029	265	14002	14585	135
PR100162	33032	0,029	161	5843	6035	58
PR101009	5323562	0,030	304	10838	11203	89
PR100199	193653	0,030	35	10482	11125	109
PR101034	5323562	0,033	326	10792	11203	183
PR101046	5323562	0,033	326	10792	11203	57
PR100325	2724705	0,040	0	11343	12319	144
PR100448	9750	0,042	305	11068	11761	41
PR100456	5320835	0,044	187	4896	5160	86
PR100280	2761525	0,048	177	5843	6269	59
PR100306	439389	0,049	307	8199	8743	27
PR100320	10917	0,056	234	7721	8426	40
PR100299	439406	0,058	1486	27520	29428	225
PR100290	2761558	0,059	128	3785	4147	35
PR100260	107982	0,062	119	34592	39387	333
PR100259	1548943	0,064	71	41713	47808	444
PR100263	182232	0,083	1655	54277	63106	544
PR100338	16211048	0,085	350	5552	6269	57
PR100161	88513	0,085	2501	52074	59736	568
PR101055	11953815	0,091	4901	72002	82100	1212
PR100211	23724461	0,094	215	3515	4068	65
PR100013	5280567	0,101	555	13063	15656	220
PR100220	119	0,103	77	1210	1427	13
PR100067PR100068	439217	0,108	2371	32269	38088	377
PR101041	8655	0,109	453	14331	17731	245
PR100272	99478	0,110	474	5176	6035	58
PR100242	5281693	0,112	83	1419	1724	24
PR100291	5706676	0,113	313	3132	3643	30
PR100286	1502076	0,117	71	1648	2059	20

PR100212	73323	0,129	6532	55016	65319	776
PR100336	152306	0,129	3525	34738	42078	360
PR100279	5780	0,144	2239	15931	19254	197
PR100169PR100170	21236	0,159	531	3355	4147	43
PR100282	5706673	0,170	313	6314	9090	96
PR100006	24405	0,232	14812	52467	70324	761
PR100324	24405	0,232	14812	52467	70324	608
PR100380	1662	0,242	1124	7471	12319	96
PR100365	5320844	0,244	1585	32262	59828	642
PR100441	5281576	0,259	15049	48426	69191	604
PR100222	564	0,290	847	2592	4147	21
PR100215	439656	0,318	619	1367	2059	21
PR100121PR100122	138	0,323	447	1357	2573	21
PR100332PR100333	7971	0,500	0	0	682	6
PR100048	8871	0,500	0	0	682	8
PR100094PR100095	439335	0,500	0	0	2155	24
PR100143	439225	0,500	0	0	4187	44
PR100134PR100135	65359	0,500	0	0	6324	74
PR100357	5490298	0,500	0	0	7169	34
PR100356	5317025	0,500	0	0	9292	187
PR100371	5282151	0,500	0	0	11203	94
PR100069PR100070	5280951	0,500	0	0	28152	220
PR100229	5281666	0,500	0	0	62470	526
PR100139PR100140	6802	0,500	0	0	64733	528
PR100246	114776	0,500	0	0	70272	652
PR100370	5280441	0,500	0	0	78553	682
PR100251	5281807	0,500	0	0	101329	780
PR100378	34755	0,500	0	0	108138	379
PR100399	493570	0,500	3	0	102257	772
PR100127	7427	0,500	4	0	89036	735
PR100056PR100057	1052	0,500	1	0	16077	145
PR100100PR100101	6076	0,500	4	0	61985	555
PR100440	5282054	0,500	3	0	30008	268
PR100014PR100015	5202	0,500	2	0	15601	12
PR100362	442813	0,500	13	0	84869	616
PR100353	5281621	0,500	9	0	49954	847
PR100360	6450184	0,500	13	0	70272	331
PR100247	5280637	0,501	71	0	70272	171
PR100403	4644	0,501	12	0	11018	128
PR100227	5281654	0,501	93	0	60225	658
PR100023	227	0,501	7	0	4402	41
PR100312	5280378	0,501	86	0	52402	436
PR100072PR100073	439224	0,501	57	0	33551	339
PR100096PR100097	89	0,501	64	0	33591	148
PR100192	980	0,501	10	0	5213	58
PR100107PR100108	6115	0,501	1	0	520	6
PR100274	1051	0,502	84	0	27306	220
PR100323	449093	0,502	99	0	30440	106
PR100029	6106	0,502	16	0	4147	19
PR100275	1050	0,502	42	0	10541	114
PR100418PR100419	6723	0,502	52	0	12789	121
PR100043PR100044	24154	0,502	58	0	14257	81

PR100125PR100126	439213	0,502	58	0	14257	36
PR100091PR100092	6047	0,502	74	0	17006	151
PR101049	5748601	0,503	409	0	70272	251
PR100264	72276	0,503	431	0	63106	101
PR100262	9064	0,503	437	0	63106	602
PR100210	5610	0,504	32	0	4276	48
PR100219	14180	0,504	537	0	63501	675
PR100004PR100005	445858	0,504	197	0	23009	106
PR100089PR100090	91531	0,504	546	0	63336	653
PR100377	439498	0,505	167	0	16705	185
PR100385	64969	0,505	126	0	11476	35
PR100273	65059	0,506	808	0	73300	622
PR100165	165271	0,506	75	0	6621	63
PR100309	65110	0,506	848	0	72521	575
PR100054	10256	0,507	93	0	6578	85
PR100358	5319116	0,508	1151	0	70272	724
PR100423PR100424	6228	0,509	6	0	345	3
PR100177	4032	0,509	129	0	7376	86
PR100225	5281708	0,509	774	0	42047	413
PR100417	135	0,509	127	0	6810	55
PR100406	3845	0,509	275	0	14626	139
PR100010PR100011	1318	0,511	415	0	18859	198
PR100330	5324677	0,511	671	0	29828	27
PR100042	637760	0,513	792	0	30570	367
PR101045	69867	0,514	243	0	8783	98
PR100339	2761537	0,515	415	0	14102	143
PR100289	2761554	0,515	557	0	18585	173
PR100201PR100202	5430	0,516	515	0	16548	156
PR100319	6804	0,516	2388	0	72634	740
PR100186	967	0,517	353	0	10303	132
PR100115PR100116	6175	0,517	1244	0	35937	282
PR100266	11250133	0,522	472	0	10608	115
PR100193	378	0,522	607	0	13515	144
PR100297PR100298	72924	0,523	235	0	5213	44
PR100217	67701	0,523	200	0	4311	45
PR100292	1051	0,523	1271	0	27306	217
PR100039	5280567	0,524	740	0	15656	37
PR100003	5280536	0,524	870	0	17982	214
PR100261	637542	0,526	677	0	13098	169
PR100059	637511	0,526	298	0	5711	45
PR100110	689043	0,527	930	0	17204	76
PR100271	736715	0,528	424	0	7478	80
PR100209	6440982	0,529	4788	0	83130	885
PR100147	800	0,529	541	0	9312	111
PR100307	199	0,530	363	0	5960	50
PR100318	6131	0,531	3661	0	58125	565
PR100288	2761550	0,532	2084	0	32750	309
PR100383	3469	0,533	645	0	9895	118
PR100077	89594	0,534	767	0	11426	129
PR100364	5280781	0,536	5447	0	75207	535
PR100049PR100050	351795	0,537	987	0	13296	192
PR100473	5280569	0,539	1165	0	15122	129

PR101044	398554	0,539	941	0	12036	159
PR101042	637775	0,542	2723	0	32403	459
PR101043	10256	0,546	603	0	6578	30
PR100472	5280460	0,547	1936	0	20491	196
PR100384	1145	0,547	36	0	380	4
PR100474	5281416	0,548	1451	0	15122	141
PR100233	5280343	0,548	5051	0	52403	530
PR100392	112072	0,550	3867	0	38762	406
PR100213	6119	0,550	143	0	1427	15
PR100197	40539	0,551	1502	0	14607	149
PR100234	5281691	0,552	6319	0	60225	318
PR100337	7618	0,556	620	0	5586	71
PR100409PR100410	5570	0,557	502	0	4402	40
PR100016PR100017	65040	0,558	519	0	3805	33
PR100188	6604563	0,568	3620	0	26509	285
PR100295	6613	0,569	4228	0	30449	224
PR100400	6613	0,570	4248	0	30449	251
PR100152	8582	0,574	9918	0	66605	636
PR100257	107971	0,586	17442	0	101329	178
PR100470	444795	0,586	8302	0	48047	535
PR100228	5280863	0,595	10376	0	54468	500
PR100398	6441567	0,596	1862	0	9738	112
PR100027PR100028	6274	0,601	1910	0	9452	102
PR100226	440735	0,603	12217	0	59060	587
PR100391	439224	0,619	7995	0	33551	337
PR100379	4396761	0,628	2937	0	11476	149
PR100321	6274	0,632	2502	0	9452	114
PR100223PR100224	5280443	0,647	14943	0	50991	572
PR100373	3611	0,650	4016	0	13391	181
PR100308	916	0,651	4893	0	16208	128
PR100415	70639	0,656	4491	0	14432	162
PR100425	6433206	0,664	3199	0	9765	39
PR100230PR100231	5280445	0,687	20331	0	54468	615
PR100408	763	0,695	1027	0	2641	23
PR100119PR100120	6132	0,698	19228	0	48529	561
PR100414	4687	0,699	7494	0	18873	186
PR100185	637540	0,703	5319	0	13098	191
PR101040	6433206	0,714	4180	0	9765	57
PR100394	649	0,739	1992	0	4175	42
PR100413	70346	0,750	11168	0	22378	209

Literaturverzeichnis

[AHP+09] Alex, Alexander ; Harvey, Sophie ; Parsons, Teresa ; Pullen, Frank S. ; Wright, Patricia ; Riley, Jo-Anne: Can density functional theory (DFT) be used as an aid to a deeper understanding of tandem mass spectrometric fragmentation pathways? In: *Rapid Commun Mass Spectrom* 23 (2009), Sep, Nr. 17, 2619–2627. DOI: 10.1002/rcm.4163

[AWT03] Audi, G. ; Wapstra, A.H. ; Thibault, C.: The Ame2003 atomic mass evaluation: (II). Tables, graphs and references. In: *Nuclear Physics A* 729 (2003), Nr. 1, 337 - 676. DOI: 10.1016/j.nuclphysa.2003.11.003

[BLLP09] Böcker, Sebastian ; Letzel, Matthias ; Lipták, Zsuzsanna ; Pervukhin, Anton: SIRIUS: Decomposing isotope patterns for metabolite identification. In: *Bioinformatics* 25 (2009), Nr. 2, S. 218–224.

[BS05] Budzikiewicz, Herbert ; Schäfer, Mathias: *Massenspektrometrie*. 5. Wiley-VCH, 2005. – ISBN 3527308229

[But99] Butina, Darko: Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. In: *Journal of Chemical Information and Computer Sciences* 39 (1999), Juli, Nr. 4, 747–750. DOI: 10.1021/ci9803381

[BW11] Berglund, Michael ; Wieser, Michael E.: Isotopic compositions of the elements 2009 (IUPAC Technical Report). In: *Pure and Applied Chemistry* 83 (2011), 397–410. DOI: 10.1351/PAC-REP-10-06-02

[BWT+08] Bolton, Evan E. ; Wang, Yanli ; Thiessen, Paul A. ; Bryant, Stephen H. ; Wheeler, Ralph A. ; Spellmeyer, David C.: Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. Version: 2008. `http://dx.doi.org/10.1016/S1574-1400(08)00012-1`. Elsevier, 2008. – DOI 10.1016/S1574–1400(08)00012–1. – ISBN 1574–1400, 217–241

[CGP98] Cheng, Changfu ; Gross, Michael L. ; Pittenauer, Ernst: Complete Structural Elucidation of Triacylglycerols by Tandem Sector Mass Spectrometry. In: *Anal. Chem.* 70 (1998), Nr. 20, 4417–4426. DOI: 10.1021/ac9805192. – ISSN 0003–2700

[CLT01] Chernushevich, I. V. ; Loboda, A. V. ; Thomson, B. A.: An introduction to quadrupole-time-of-flight mass spectrometry. In: *J Mass Spectrom* 36 (2001), Aug, Nr. 8, 849–865. DOI: 10.1002/jms.207

[Con08] Consortium, The B.: Interoperability with Moby 1.0—It's better than sharing your toothbrush! In: *Briefings in Bioinformatics* (2008). DOI: 10.1093/bib/bbn003

[CTMD05] Churchwell, Mona I. ; Twaddle, Nathan C. ; Meeker, Larry R. ; Doerge, Daniel R.: Improving LC-MS sensitivity through increases in chromatographic performance: comparisons of UPLC-ES/MS/MS to HPLC-ES/MS/MS. In: *J Chromatogr B Analyt Technol Biomed Life Sci* 825 (2005), Oct, Nr. 2, 134–143. DOI: 10.1016/j.jchromb.2005.05.037

[DNH+92] Dalby, Arthur ; Nourse, James G. ; Hounshell, W. D. ; Gushurst, Ann K. I. ; Grier, David L. ; Leland, Burton A. ; Laufer, John: Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. In: *Journal of Chemical Information and Computer Sciences* 32 (1992), Nr. 3, 244-255. DOI: 10.1021/ci00007a012

[Dun08] Dunn, Warwick B.: Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. In: *Physical Biology* 5 (2008), Nr. 1, 011001 (24pp). `http://stacks.iop.org/1478-3975/5/011001`

[DZHS85] Dewar, Michael J. S. ; Zoebisch, Eve G. ; Healy, Eamonn F. ; Stewart, James J. P.: Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. In: *J. Am. Chem. Soc.* 107 (1985), Nr. 13, 3902–3909. DOI: 10.1021/ja00299a024

[FKD+00] Fiehn, O. ; Kopka, J. ; Dörmann, P. ; Altmann, T. ; Trethewey, R. N. ; Willmitzer, L.: Metabolite profiling for plant functional genomics. In: *Nat Biotechnol* 18 (2000), Nov, Nr. 11, 1157–1161. DOI: 10.1038/81137

[FMM+89] Fenn, J. B. ; Mann, M. ; Meng, C. K. ; Wong, S. F. ; Whitehouse, C. M.: Electrospray ionization for mass spectrometry of large biomolecules. In: *Science* 246 (1989), Oct, Nr. 4926, S. 64–71. DOI: 10.1021/ac00190a023

[GMRRW01] Gkoutos, G. V. ; Murray-Rust, P. ; Rzepa, H. S. ; Wright, M.: Chemical markup, XML and the World-Wide Web. 3. Toward a signed semantic chemical web of trust. In: *J Chem Inf Comput Sci* 41 (2001), Nr. 5, S. 1124–1130.

[GS00] Gilmore, I.S ; Seah, M.P: Static SIMS: towards unfragmented mass spectra — the G-SIMS procedure. In: *Applied Surface Science* 161 (2000), Nr. 3–4, 465 - 480. DOI: 10.1016/S0169-4332(00)00317-2. – ISSN 0169–4332

[HAK+10] Horai, Hisayuki ; Arita, Masanori ; Kanaya, Shigehiko ; Nihei, Yoshito ; Ikeda, Tasuku ; Suwa, Kazuhiro ; Ojima, Yuya ; Tanaka, Kenichi ; Tanaka, Satoshi ; Aoshima, Ken ; Oda, Yoshiya ; Kakazu, Yuji ; Kusano, Miyako ; Tohge, Takayuki ; Matsuda, Fumio ; Sawada, Yuji ; Hirai, Masami Y. ; Nakanishi, Hiroki ; Ikeda, Kazutaka ; Akimoto, Naoshige ; Maoka, Takashi ; Takahashi, Hiroki ; Ara, Takeshi ; Sakurai, Nozomu ; Suzuki, Hideyuki ; Shibata, Daisuke ; Neumann, Steffen ; Iida, Takashi ; Tanaka, Ken ; Funatsu, Kimito ; Matsuura, Fumito ; Soga, Tomoyoshi ; Taguchi, Ryo ; Saito, Kazuki ; Nishioka, Takaaki: MassBank: a public repository for sharing mass spectral data for life sciences. In: *J Mass Spectrom* 45 (2010), Jul, Nr. 7, 703–714. DOI: 10.1002/jms.1777

[Hal96] Halgren, Thomas A.: Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. In: *Journal of Computational Chemistry* 17 (1996), Nr. 5-6, S. 490–519. . – ISSN 1096–987X

[Hal99] Halgren, Thomas A.: MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. In: *Journal of Computational Chemistry* 20 (1999), Nr. 7, S. 730–748. . – ISSN 1096–987X

[HJK96] Hanser, Th. ; Jauffret, Ph. ; Kaufmann, G.: A New Algorithm for Exhaustive Ring Perception in a Molecular Graph. In: *Journal of Chemical Information and Computer Sciences* 36 (1996), Nr. 6, 1146-1152. DOI: 10.1021/ci960322f

[HKF+08] Hill, Dennis W. ; Kertesz, Tzipporah M. ; Fontaine, Dan ; Friedman, Robert ; Grant, David F.: Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. In: *Anal Chem* 80 (2008), Jul, Nr. 14, 5574–5582. DOI: 10.1021/ac800548g

[HM05] Hill, Alastair W. ; Mortishire-Smith, Russell J.: Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. In: *Rapid Communications in Mass Spectrometry* 19 (2005), Nr. 21, 3111–3118. DOI: 10.1002/rcm.2177

[HMRR06] Holliday, G. L. ; Murray-Rust, P. ; Rzepa, H. S.: Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. In: *J Chem Inf Model* 46 (2006), January, Nr. 1, 145–157. DOI: http://dx.doi.org/10.1021/ci0502698

[HNP95] Hellerstein, Joseph M. ; Naughton, Jeffrey F. ; Pfeffer, Avi: Generalized Search Trees for Database Systems. In: Dayal, Umeshwar (Hrsg.) ; Gray, Peter M. D. (Hrsg.) ; Nishio, Shojiro (Hrsg.): *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, Morgan Kaufmann, 1995. – ISBN 1–55860–379–4, S. 562–573

[HP01] Hassinen, Tommi ; Peräkylä, Mikael: New energy terms for reduced protein models implemented in an off-lattice force field. In: *Journal of Computational Chemistry* 22 (2001), Nr. 12, 1229–1242. DOI: 10.1002/jcc.1080. – ISSN 1096–987X

[HRM+06] Heinonen, Markus ; Rantanen, Ari ; Mielikäinen, Taneli ; Pitkänen, Esa ; Kokkonen, Juha ; Rousu, Juho: Ab Initio prediction of molecular fragments from tandem mass spectrometry data. In: *Proc. of German Conference on Bioinformatics (GCB 2006)*, 2006 (Lecture Notes in Informatics), S. 40–53

[HRM+08] Heinonen, Markus ; Rantanen, Ari ; Mielikäinen, Taneli ; Kokkonen, Juha ; Kiuru, Jari ; Ketola, Raimo A. ; Rousu, Juho: FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. In: *Rapid Commun Mass Spectrom* 22 (2008), September, Nr. 19, S. 3043–3052.

[HS07] Hoffmann, Edmond d. ; Stroobant, Vincent: *Mass Spectrometry: Principles and Applications*. 3. Auflage. John Wiley & Sons, 2007. – ISBN 0470033118

[HSRF08] Höltje, Hans-Dieter ; Sippl, Wolfgang ; Rognan, Didier ; Folkers, Gerd: *Molecular Modeling: Basic Principles and Applications*. 3. überarb. u. erg. Auflage. Wiley-VCH Verlag GmbH & Co. KGaA, 2008. – ISBN 3527315683

[HTF08] Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome H.: *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2008. – ISBN 9780387848570

[HWN11] Hildebrandt, Christian ; Wolf, Sebastian ; Neumann, Steffen: Database supported candidate search for Metabolite identification. In: *J Integr Bioinform* 8 (2011), Nr. 2, 157. DOI: 10.2390/biecoll-jib-2011-157

[JWD] James, C. A. ; Weininger, D. ; Delany, J. ; Daylight Chemical Information Systems, Inc. (Hrsg.): *Daylight Theory Manual*. Version 4.9. Daylight Chemical Information Systems, Inc., `http://www.daylight.com/dayhtml/doc/theory/index.html`. – Abgerufen im Oktober 2011

[KG00] Kanehisa, M. ; Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. In: *Nucleic Acids Res* 28 (2000), Jan, Nr. 1, S. 27–30. DOI: 10.1093/nar/27.1.29

[KGKN02] Kanehisa, M. ; Goto, S. ; Kawashima, S. ; Nakaya, A.: The KEGG databases at GenomeNet. In: *Nucleic Acids Res* 30 (2002), S. 42–46.

[KHL+07] Kuhn, Stefan ; Helmus, Tobias ; Lancashire, Robert J. ; Murray-Rust, Peter ; Rzepa, Henry S. ; Steinbeck, Christoph ; Willighagen, Egon L.: Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. In: *J Chem Inf Model* 47 (2007), Nr. 6, 2015-2034. DOI: 10.1021/ci600531a

[KLGM98] Kerber, A. ; Laue, R. ; Grüner, T. ; Meringer, M.: MOLGEN 4.0. In: *MATCH Commun. Math. Comput. Chem.* 37 (1998), S. 205–208.

[KLMV01] Kerber, A. ; Laue, R. ; Meringer, M. ; Varmuza, K.: MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation. In: *Advances in Mass Spectrometry* 15 (2001), S. 939–940.

[KMR06] Kerber, Adalbert ; Meringer, Markus ; Rücker, Christoph: CASE via MS: Ranking structure candidates by mass spectra. In: *Croatica chemica acta* 79 (2006), Nr. 3, 449–464. `http://cat.inist.fr/?aModele=afficheN&cpsidt=18330616`. – ISSN 0011–1643

[LEL06] Lottspeich, Friedrich ; Engels, Joachim W. ; Lay, Solodkoff Z.: *Bioanalytik*. 2. Aufl. 2006. Spektrum Akademischer Verlag, 2006

[LG03] Leach, Andrew R. ; Gillet, V.J.: *An Introduction to Chemoinformatics*. XV, 259p. Springer Netherlands, 2003. – ISBN 1402013477

[Luo03] Luo, Yu-Ran: *Handbook of bond dissociation energies in organic compounds*. CRC Press, 2003. – 394 S. – ISBN 0849315891, 9780849315893

[MRR99] Murray-Rust, Peter ; Rzepa, Henry S.: Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. In: *Journal of Chemical Information and Computer Sciences* 39 (1999), Nr. 6, 928-942. DOI: 10.1021/ci990052b

[MRR01] Murray-Rust, P. ; Rzepa, H. S.: Chemical markup, XML, and the World-Wide Web. 2. Information objects and the CMLDOM. In: *J Chem Inf Comput Sci* 41 (2001), Nr. 5, S. 1113–1123.

[MRR03] Murray-Rust, Peter ; Rzepa, Henry S.: Chemical markup, XML, and the World Wide Web. 4. CML schema. In: *J Chem Inf Comput Sci* 43 (2003), Nr. 3, 757–772. DOI: 10.1021/ci0256541

[MRRWW04] Murray-Rust, Peter ; Rzepa, Henry S. ; Williamson, Mark J. ; Willighagen, Egon L.: Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. In: *J Chem Inf Comput Sci* 44 (2004), Nr. 2, 462–469. DOI: 10.1021/ci034244p

[NB10] Neumann, Steffen ; Böcker, Sebastian: Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. In: *Anal Bioanal Chem* 398 (2010), Dec, Nr. 7-8, 2779–2788. DOI: 10.1007/s00216-010-4142-5

[NR03] Norvig, Peter ; Russell, Stuart: *Artificial Intelligence: A Modern Approach*. 2. A. International Edition. Prentice Hall, 2003. – ISBN 0130803022

[OBJ+11] O'Boyle, Noel M. ; Banck, Michael ; James, Craig A. ; Morley, Chris ; Vandermeersch, Tim ; Hutchison, Geoffrey R.: Open Babel: An open chemical toolbox. In: *J Cheminform* 3 (2011), Oct, Nr. 1, 33. DOI: 10.1186/1758-2946-3-33

[OGS+99] Ogata, H. ; Goto, S. ; Sato, K. ; Fujibuchi, W. ; Bono, H. ; Kanehisa, M.: KEGG: Kyoto Encyclopedia of Genes and Genomes. In: *Nucleic Acids Res* 27 (1999), Jan, Nr. 1, S. 29–34.

[PG00] Pichersky, Eran ; Gang, David R.: Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. In: *Trends in Plant Science* 5 (2000), Nr. 10, 439 - 445. DOI: 10.1016/S1360-1385(00)01741-6. – ISSN 1360–1385

[PW10] Pence, Harry E. ; Williams, Antony: ChemSpider: An Online Chemical Information Resource. In: *Journal of Chemical Education* 87 (2010), Nr. 11, 1123-1124. DOI: 10.1021/ed100697w

[RCC+92] Rappe, A. K. ; Casewit, C. J. ; Colwell, K. S. ; Goddard, W. A. ; Skiff, W. M.: UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. In: *Journal of the American Chemical Society* 114 (1992), Nr. 25, 10024-10035. DOI: 10.1021/ja00051a040

[Sch10] Schmid, Ernst-Georg: *Database-driven procurement of substances in the researching chemical industry - An algorithmic optimization approach*, Mercator School of Management - Fakultät für Betriebswirtschaftslehre - Technology and Operations Management - Wirtschaftsinformatik und Operations Research, Diss., June 2010. `http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hbz:464-20100809-090447-2`

[SGK+] Schymanski, Emma L. ; Gallampois, Christine M. J. ; Krauss, Martin ; Meringer, Markus ; Neumann, Steffen ; Schulze, Tobias ; Wolf, Sebastian ; Brack, Werner: Consensus Structure Elucidation Combining GC/EI-MS, Structure Generation and Calculated Properties. In: *Analytical Chemistry - Akzeptiert 02/2012*. DOI: 10.1021/ac203471y

[SHK+03] Steinbeck, Christoph ; Han, Yongquan ; Kuhn, Stefan ; Horlacher, Oliver ; Luttmann, Edgar ; Willighagen, Egon: The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. In: *Journal of Chemical Information and Computer Sciences* 43 (2003), Nr. 2, 493-500. DOI: 10.1021/ci025584y

[SHK+06] Steinbeck, Christoph ; Hoppe, Christian ; Kuhn, Stefan ; Floris, Matteo ; Guha, Rajarshi ; Willighagen, Egon L.: Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. In: *Curr Pharm Des* 12 (2006), Nr. 17, S. 2111–2120.

[SHT03] Stein, Stephen E. ; Heller, Stephen R. ; Tchekhovskoi, Dmitrii: An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. In: *Proceedings of the 2003 International Chemical Information Conference*, Infonortics, 2003, S. 131–143

[SMB09] Schymanski, Emma L. ; Meringer, Markus ; Brack, Werner: Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? In: *Anal Chem* 81 (2009), May, Nr. 9, 3608–3617. DOI: 10.1021/ac802715e

[SMW+05] Smith, Colin A. ; Maille, Grace O. ; Want, Elizabeth J. ; Qin, Chuan ; Trauger, Sunia A. ; Brandon, Theodore R. ; Custodio, Darlene E. ; Abagyan, Ruben ; Siuzdak, Gary: METLIN: A Metabolite Mass Spectral Database. In: *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology* Bd. 27. Louisville, Kentucky, 2005, S. 747–751

[SS94] Stein, Stephen E. ; Scott, Donald R.: Optimization and testing of mass spectral library search algorithms for compound identification. In: *Journal of the American Society for Mass Spectrometry* 5 (1994), September, Nr. 9, 859–866. DOI: 10.1016/1044-0305(94)87009-8

[Ste90] Stewart, J. J.: MOPAC: a semiempirical molecular orbital program. In: *J Comput Aided Mol Des* 4 (1990), Mar, Nr. 1, S. 1–105.

[Swe03] Sweeney, Daniel L.: Small molecules as mathematical partitions. In: *Anal Chem* 75 (2003), Nr. 20, S. 5362–5373.

[WB09] Wieser, Michael E. ; Berglund, Michael: Atomic weights of the elements 2007 (IUPAC Technical Report). In: *Pure and Applied Chemistry* 81 (2009), 2131–2156. DOI: 10.1351/PAC-REP-09-08-03

[WBD98] Willett, Peter ; Barnard, John M. ; Downs, Geoffrey M.: Chemical Similarity Searching. In: *Journal of Chemical Information and Computer Sciences* 38 (1998), Nr. 6, 983-996. DOI: 10.1021/ci9800211

[WD11] Weissberg, Avi ; Dagan, Shai: Interpretation of ESI(+)-MS-MS spectra—Towards the identification of "unknowns". In: *International Journal of Mass Spectrometry* 299 (2011), Nr. 2–3, 158 - 168. DOI: 10.1016/j.ijms.2010.10.024. – ISSN 1387–3806

[Wei88] Weininger, David: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In: *Journal of Chemical Information and Computer Sciences* 28 (1988), Nr. 1, 31-36. DOI: 10.1021/ci00057a005

[WHD+08] Werner, Erwan ; Heilier, Jean-François ; Ducruix, Céline ; Ezan, Eric ; Junot, Christophe ; Tabet, Jean-Claude: Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. In: *J Chromatogr B Analyt Technol Biomed Life Sci* 871 (2008), Aug, Nr. 2, 143–163. DOI: 10.1016/j.jchromb.2008.07.004

[WSMHN10] Wolf, Sebastian ; Schmidt, Stephan ; Müller-Hannemann, Matthias ; Neumann, Steffen: In silico fragmentation for computer assisted identification of metabolite mass spectra. In: *BMC Bioinformatics* 11 (2010), Nr. 1, 148. DOI: 10.1186/1471-2105-11-148. – ISSN 1471–2105

[WTK+07] Wishart, David S. ; Tzur, Dan ; Knox, Craig ; Eisner, Roman ; Guo, An C. ; Young, Nelson ; Cheng, Dean ; Jewell, Kevin ; Arndt, David ; Sawhney, Summit ; Fung, Chris ; Nikolai, Lisa ; Lewis, Mike ; Coutouly, Marie-Aude ; Forsythe, Ian ; Tang, Peter ; Shrivastava, Savita ; Jeroncic, Kevin ; Stothard, Paul ; Amegbey, Godwin ; Block, David ; Hau, David. D. ; Wagner, James ; Miniaci, Jessica ; Clements, Melisa ; Gebremedhin, Mulu ; Guo, Natalie ; Zhang, Ying ; Duggan, Gavin E. ; MacInnis, Glen D. ; Weljie, Alim M. ; Dowlatabadi, Reza ; Bamforth, Fiona ; Clive, Derrick ; Greiner, Russ ; Li, Liang ; Marrie, Tom ; Sykes, Brian D. ; Vogel, Hans J. ; Querengesser, Lori: HMDB:

the Human Metabolome Database. In: *Nucleic Acids Res* 35 (2007), Nr. suppl1, D521-526. DOI: 10.1093/nar/gkl923

[WWW89] Weininger, David ; Weininger, Arthur ; Weininger, Joseph L.: SMILES. 2. Algorithm for generation of unique SMILES notation. In: *Journal of Chemical Information and Computer Sciences* 29 (1989), Nr. 2, 97-101. DOI: 10.1021/ci00062a008

[WXS+09] Wang, Yanli ; Xiao, Jewen ; Suzek, Tugba O. ; Zhang, Jian ; Wang, Jiyao ; Bryant, Stephen H.: PubChem: a public information system for analyzing bioactivities of small molecules. In: *Nucleic Acids Res* 37 (2009), Jul, Nr. Web Server issue, W623–W633. DOI: 10.1093/nar/gkp456

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Halle, den 29.02.2012 Sebastian Wolf

Lebenslauf

Persönliche Daten

Name	Sebastian Wolf
geboren am	24. Februar 1984
in	Karl-Marx-Stadt (Chemnitz)
Staatsangehörigkeit	deutsch
Familienstand	verheiratet

Schulausbildung

1990 - 1992	Pestalozzischule Oberlungwitz
1990 - 1992	Humboldtschule Oberlungwitz
1994 - 2000	Lessing Gymnasium Hohenstein-Ernstthal
2000 - 2001	Will Sinclair High School Rocky Mountain House
2001 - 2003	Lessing Gymnasium Hohenstein-Ernstthal, Abschluss Abitur

Universitätsausbildung

10/2003 - 12/2008	Studium der Bioinformatik an der Martin-Luther-Universität Halle Wittenberg Abschluss als Diplom-Bioinformatiker
01/2009 - 12/2011	wissenschaftlicher Mitarbeiter am Leibniz-Institut für Pflanzenbiochemie, Halle (Saale) Arbeitsgruppe Massenspektrometrie & Bioinformatik

Halle (Saale), den 29. Februar 2012 Sebastian Wolf