Optimierung der Signalverarbeitung und Signalerkennung zur automatisierten NMR-Strukturbestimmung von Proteinen

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

Harald Donaubauer

aus

Riedl

im Jahr 2017

Das Promotionsgesuch wurde eingereicht am: 27.10.2017

Die Arbeit wurde angeleitet von:

Prof. Dr. rer. nat. Dr. med. Hans Robert Kalbitzer

Prüfungsausschuss

Vorsitzender:	Prof. Dr. rer. nat. Christoph Oberprieler
Erstgutachter:	Prof. Dr. rer. nat. Dr. med. Hans Robert Kalbitzer
Zweitgutachter:	Prof. Dr. rer. nat. Wolfram Gronwald
Drittprüfer:	Prof. Dr. rer. nat. Christine Ziegler

Unterschrift:

_____

Dipl.-Phys. Dipl.-Inf. (FH) Harald Donaubauer

Inhaltsverzeichnis

1 Einleitung

Da eine manuelle Bestimmung der Struktur eines Proteins sehr zeitaufwändig ist, war das Ziel dieser Arbeit, die automatische Strukturbestimmung von Proteinen mittels NMR-Spektroskopie mit dem Softwareprojekt AUREMOL (Wolfram Gronwald und Hans Robert Kalbitzer 2004) durch die Verbesserung der Signalidentifizierung weiter in Richtung Vollautomatisierung zu bringen. Auf eine Einführung in die Theorie der NMR soll an dieser Stelle verzichtet werden und es wird auf die einschlägige Fachliteratur (Hausser und Kalbitzer 1989; Claridge 2009; Cavanagh 2007; Levitt 2009) verwiesen.

Generell existieren zwei klassische Ansätze der automatischen Strukturbestimmung:

- Die „bottom-up“-Strategie nutzt eine umfangreiche Basis von experimentellen Daten. Mit diesen wird dann versucht, eine sequentielle Zuordnung zu erreichen, um damit die endgültige räumliche Struktur des Proteins zu bestimmen. Softwarelösungen, die diesen Ansatz verfolgen, sind z. B. GARANT (Bartels et al. 1997), PASTA (Leutner et al. 1998), CONTRAST (Olson und Markley 1994), oder AUTOASSIGN (Zimmerman et al. 1994). Der Nachteil dieser Strategie ist, dass kein Vorwissen aus anderen Strukturen genutzt wird.
- Die „top-down“-Strategie hingegen versucht den experimentellen Aufwand zu reduzieren, indem sie sich auf den Teil der Strukturauswertung fokussiert. Dazu werden die Informationen aus einer homologen Struktur verwendet, um zusätzliches Wissen basierend aus diesen Informationen durch eine Voraussage von chemischen Verschiebungen oder Rückgrat-Rotationswinkel zu erlangen. Um dann die NMR-Parameter aus der Teststruktur an die zu bestimmende Struktur anzupassen, werden zusätzlich externe statistische Datenbanken genutzt um über mehrere Iterationen die Struktur zu verfeinern, bis eine möglichst hohe Übereinstimmung mit der zu bestimmenden Struktur (also den experimentellen NMR-Daten) vorliegt.

Das molekülorientierte Konzept von AUREMOL basiert auf der „top-down“-Strategie. Dabei wird möglichst viel Vorwissen aus den internen AUREMOL-Datenbanken zur Strukturbestimmung genutzt. Zu Beginn der Auswertung wird zusätzlich zu den Informationen aus der internen AUREMOL-Datenbank noch so viel Vorwissen über das zu

bestimmende Protein wie möglich gesammelt. Dieses Vorwissen besteht aus der Primärsequenz des auszuwertenden Proteins oder die Parameter des Experiments (z. B. Druck, Temperatur oder Zusammensetzung der Pufferlösung).

Basierend auf diesem Vorwissen und einer Startstruktur (z. B. aus Homologie-Modelling oder ausgestreckter Strang) kann eine Start-Zuordnung der chemischen Verschiebungen des Proteins generiert werden. Danach werden zusätzliche Informationen aus dem Spektrum benötigt, welche durch die Module Peak-Picking, Volumenintegration durch Segmentierung und Signalidentifizierung durch die Bayessche Signalanalyse ermöglicht werden. Diese Module wurden im Rahmen dieser Arbeit erweitert und verbessert.

Durch den Abgleich der experimentellen Daten und der simulierten Daten wird die Strukturinformation beginnend mit der Startstruktur iterativ verfeinert. Diese Strukturinformationen dienen zur Berechnung von Diederwinkel, Wasserstoffbrücken und Atomabständen und gehen als Beschränkungen (sog. Restraints) in eine Moleküldynamik-Software ein. Die aus der Molekulardynamik erhaltene Struktur kann wiederum nach der Strukturvalidierung durch die AUREMOL-Module verbessert werden und man erhält wieder eine neue Struktur für die Molekulardynamik. Dieser Vorgang wird solange wiederholt, bis eine endgültige Struktur vorliegt. Der detaillierte Ablauf ist in (Wolfram Gronwald und Hans Robert Kalbitzer 2004) dargestellt.

Durch den technischen Fortschritt der Spektrometer erhöhte sich die digitale Auflösung der Spektren stets weiter. Dies hatte zur Folge, dass der Bedarf des Arbeitsspeichers und der Rechenaufwand der Routinen in AUREMOL stark anstieg. Daher war es ein Ziel, die Basisfunktionen hinsichtlich ihrer Laufzeit zu verbessern. Dazu wurden die wichtigsten Algorithmen parallelisiert und in objektorientierte und vor allem in speicheroptimierte Module überführt.

Da überwiegend alle Algorithmen dahingehend implementiert waren, dass für jede weitere verwendete Dimension ein eigener Quellcode existierte, wurden alle Module neu implementiert, um für alle Dimensionen lediglich einen zentralen Algorithmus zur Verfügung zu nutzen. Dies wurde durchwegs durch rekursive Ansätze realisiert.

Ein weiterer Vorteil der neuen Module ist zudem die Unabhängigkeit der AUREMOL zugrundeliegenden Basis von AMIX der Firma Bruker (Neidig et al. 1995). Somit wurden

alle neu- und weiterentwickelten Module dieser Arbeit von AMIX unabhängig gemacht und können in Zukunft leicht in ein neues Visualisierungstool eingebunden werden.

Im wesentlichen behandelt diese Arbeit die Weiterentwicklung und Optimierung der Verarbeitung der NMR-Rohdaten über die Klassifizierung der NMR-Signale bis hin zur Verbesserung deren Signalpositionen.

1.1 Die Verbesserung der Volumenberechnung von NMR-Signalen

Die Berechnung der Signalvolumen ist eine unabdingbare Information zur Erlangung einer dreidimensionalen Struktur eines Proteins in der NMR, da diese auf Informationen aus den Torsionswinkeln der J-Kopplung (Vuister und Bax 1993) und der Abstandsinformationen der NOE-Signale basiert. Daher ist es für die Strukturbestimmung wichtig, die Volumen möglichst exakt zu bestimmen (Neuhaus und Williamson 1989).

Damit das Volumen eines Signals berechnet werden kann, ist die Information der Position eines Signals nötig. Um diese zu bestimmen, sind im Softwarepaket AUREMOL diverse Module implementiert, die eine Signalklassifizierung vornehmen (automatisches Peak-Picken und die anschließende Signalklassifizierung durch das Bayes-Theorem oder durch manuelle Bestimmung des Benutzers).

Zur Bestimmung der Volumen gibt es verschiedene Ansätze, von denen hier einige kurz aufgeführt werden sollen:

- Manuelle Festlegung des Integrations-Bereichs. Hier werden alle Pixel des festgelegten Bereichs aufaddiert (z. B. in XWINNMR). Diese Methode ist aber enorm aufwändig, da jedes Signal interaktiv markiert werden muss.
- Multiplikation der Linienbreite auf halber Höhe eines Signals jeder Dimension mit der Intensität des Signals (Fejzo et al. 1990).
- Durch Anpassung von theoretischen Signalformen oder benutzerdefinierten Linienformen (Gauss oder Lorenz) (Brown und Huestis 1994; Sze et al. 1995; Eccles et al. 1991; Denk et al. 1986) an die experimentellen NMR-Signale. Die Angleichung benutzerdefinierter Signale als Referenz auf das Spektrum ist aber

stark davon abhängig, die Referenzsignale korrekt auszuwählen und benötigt zudem einen interaktiven Eingriff des Benutzers.

- Automatische Bestimmung der Integrationsbereiche durch eine stetig fallende Steigung bei wachsenden Abstand zum Signalmittelpunkt. Dabei darf die Steigung einen vorher interaktiven festgelegten Wert nicht unterschreiten (Shen und Poulsen 1990).
- Bestimmung der Volumen durch Fits von Linienformen und Amplituden, welche durch die Identifizierung von isolierten Signalen und einer Gruppierung von Linienformen von Signalen erhalten worden sind, auf die experimentellen Signale. Diese Informationen, welche zum Peak-Picken verwendet worden sind stellen somit auch die Grundlage bei AUTOPSY für die Integration dar (Koradi et al. 1998).

Das Softwarepaket AUREMOL stellt ein Modul zur Verfügung, welches auf der numerischen Methode der Segmentierung basiert (Neidig und Kalbitzer 1990). Dabei werden die Einzelintensitäten basierend auf der Signalform im digitalen Raster des Frequenzspektrums zum anteiligen Volumen zugeordnet. Dazu muss ein Bereich durch die Linienbreite eines Signals aus der Peakliste an dessen Position festgelegt werden. Danach wird der Ansatz der Integration basierend auf der Segmentierung angewandt. Hierzu werden alle Intensitäten und Positionen der enthaltenen Extrema als sog. Wachstumskeime festgelegt und absteigend bezüglich derer Intensität sortiert. Mithilfe des Region-Wachstums-Algorithmus (Geyer et al. 1995) werden alle Integrale dieser sortierten Extrema berechnet, indem die Region um diese Wachstumskeime um weitere sog. Seeds anwachsen, bis alle Intensitäten zu den Extrema zugeordnet wurden. Die Miteinbeziehung aller Extrema innerhalb des Integrationsbereichs soll verhindern, dass Seeds an den benachbarten Extrema vorbeiwachsen, falls sie der Wachstumsregel widersprechen und dem falschen Integral zugeordnet werden. Wichtig ist, dass nur Intensitäten in das Wachstum miteinbezogen werden, welche über dem Segmentierungslevel liegen. Somit sind im finalen Volumen V lediglich die Intensitäten I_i enthalten, welche in der Seedliste (Anzahl N Intensitäten) des zu integrierenden Signals anzutreffen sind:

$$V = \sum_{i=1}^{N} I_i \tag{1}$$

Wobei für die Intensitäten I_i aus der Formel 1, der Intensität des Extremums I_m an der Position des Signals und bei einem Segmentierungslevel S_l gilt:

$$I_i = \begin{cases} I_i, \text{ falls } I_i > S_l I_m \\ 0 \text{ sonst} \end{cases} \quad (2)$$

Da vorher die Integrationsbereiche interaktiv definiert werden mussten, wurde eine automatische Bestimmung der Integrationsgrenzen basierend auf den Signalnachbarschaften (also nur Signale, welche auch in der Peakliste festgelegt wurden) realisiert, welche jedoch den Nachteil hat, dass diese relativ unzuverlässig die Bereiche definiert und der Algorithmus dazu tendiert, zu große Integrationsbereiche festzulegen, was dann zu längeren Rechenzeiten führt. Durch die Abhängigkeit der Methode von den festgelegten Signalen aus der Peakliste, liefert diese Methode inkonsistente Volumen, falls ein Signal aus der Peakliste gelöscht oder hinzugefügt wird. Um diese Probleme zu lösen, wurde eine Verbesserung in Rahmen dieser Arbeit erarbeitet, welche die Größe des Bereiches adaptiv erweitert und die Abhängigkeiten vorher festgelegter Signale in der Peakliste auflöst.

Im Rahmen dieser Arbeit wurde eine Visualisierung der Volumen in AUREMOL eingeführt, welche es ermöglicht, ein oder mehrere Volumen grafisch darzustellen, um das Ergebnis der Integration visuell validieren zu können.

1.2 Das Modul Schwerpunktbestimmung zur Optimierung der genauen Position von Signalen

NMR-Signale werden in der Regel anhand der Position des Signal-Maximums aus den Intensitäten bezüglich des digitalen Rasters der Rohdaten des Spektrums bestimmt. In AUREMOL wird dazu jede maximale Intensität gewählt, welche durch Pixel niedrigerer Intensität umgeben ist, falls diese über dem vorher festgelegten Schwellwert liegt (Neidig et al. 1984; Cieslar et al. 1988). Diese Position des Extremums und deren ganzzahlige Koordinate wird in AUREMOL zur Berechnung der Koordinate in der ppm-Skala herangezogen. Da die Positionen der Intensitäten im Rohspektrum als ganzzahlige SI-Koordinaten vorliegen, jedoch die Positionen der NMR-Signale auf der ppm-Skala eine höhere Genauigkeit aufweisen, wird zur Berechnung der ppm-Werte der Mittelpunkt des

Pixels der Intensität gewählt. Diese Methode ist abhängig davon, welche digitale Auflösung für die Rohdaten des Spektrums gewählt wird. Daher wird der ppm-Wert um so ungenauer, je niedriger die Auflösung des Spektrums ist. Für eine optimale Positionierung der NMR-Signale ist jedoch eine möglichst genaue Bestimmung der ppm-Koordinaten nötig.

In STELLA (Kleywegt et al. 1990) wird die Signalform mit den zuvor (durch einen Lernvorgang) gespeicherten Signalen aus einer Datenbank geprüft, ob sich das gepickte Signal mit einem Signal aus der Datenbank durch das Cosinuskriterium deckt. Danach wird die Position des Signals am Extremum durch den Ansatz einer polynomialen Interpolation der umliegenden Intensitäten verbessert.

Der Ansatz CAPP (Garrett et al. 2011) definiert die Position eines Signals durch die Mittelung der Ellipsenzentren, welche zur Identifizierung eines Nutzsignals ermittelt werden.

Da in Rahmen der Erweiterung der Integration die Form des Volumens eines Signals aus einer Struktur jederzeit abrufbar war, konnte dies als zusätzliches Wissen verwendet werden, um die Position basierend auf dem Schwerpunkt der Volumenform zu bestimmen.

1.3 Klassifizierung von NMR-Signalen durch die Bestimmung der Bayesschen Wahrscheinlichkeit

Die Klassifizierung der NMR-Signale in Nutzsignale, welche eine oder mehrere Zuordnungen repräsentieren und in Störsignale stellt im Allgemeinen ein Problem dar, da jede Bewertungsmethode von der Güte des aufgenommenen Spektrums abhängt und sich eine automatische Bestimmung als schwierig darstellt.

Einige Methoden zur Identifizierung der Nutzsignale sind die

- interaktive Bestimmung der Positionen durch den Benutzer
- automatische Bestimmung durch einen durch den Benutzer festgelegten Schwellwert der Intensität (Neidig et al. 1984; Cieslar et al. 1988)
- automatisch durch Klassifikatoren wie z. B. durch neuronale Netze (Carrara et al. 1993; Corne et al. 1992) und mit RESCUE (Pons und Delsuc 1999)
- automatisch basierend auf der Berechnung der Ellipsen, welche am besten zu den Verläufen der Konturen aus dem Spektrum passen mit CAPP (Garrett et al. 2011)

- Filterung durch einen auf der Signalform basierenden Filter durch GIFA (Pons et al. 1996)
- automatisch basierend auf der Signal-Symmetrie, des lokalen Rausch-Levels und der Signalform durch AUTOPSY (Koradi et al. 1998) und PICKY (Alipanahi et al. 2009)

Um den Zeitaufwand für eine Strukturbestimmung zu optimieren, musste ein automatisierter Ansatz gefunden werden, welcher die Arbeit des Experimentators erleichtert bzw. beschleunigt.

Von einer Verwendung der linearen Diskriminanzanalyse nach Fischer (Fahrmeir und Brachinger 1996) wurde abgesehen, da diese Methode die Signale lediglich einer Klasse zuordnet. Damit geht die Information verloren, ob ein NMR-Signal nur knapp oder eindeutig einer Klasse zugeordnet worden ist.

Neuronale Netze haben den Nachteil, dass sich die Netzwerkstruktur im Falle eines Backpropagation-Netzes (Anzahl der Eingabeneuronen, Neuronen der verborgenen Schicht oder der Ausgabeneuronen unterscheiden) zwischen den verschiedenen Typen von Spektren erheblich unterscheiden kann, so dass für jeden Typ eines Spektrums ein Netz bestimmt werden muss.

Die alleinige Bewertung der NMR-Signale durch die Gaußsche Wahrscheinlichkeit ist zumeist nicht ausreichend, da viele Störsignale zu hohe Wahrscheinlichkeiten erhalten. Daher wird in dieser Arbeit der Bayes-Ansatz zur Signalidentifizierung aus (Antz et al. 1995) erweitert, um eine verbesserte Bewertung der Nutzsignale zu erreichen.

Ein weiteres Problem stellt sich dann, wenn der Benutzer zwar ein Spektrum prozessiert hat, jedoch keine optimalen Angaben der zu bestimmenden Bereiche der Signale und Störsignale hat. Diese werden bei Klassifizierungsverfahren benötigt, um den Lerndatensatz zu definieren. Daher kam die Motivation, eine Methode zu finden, welche die genaue Definition aufweicht und dem Nutzer erlaubt nur grobe Bereiche (im Folgenden Klassen) zu definieren. Diese Klassen werden lediglich dazu benutzt, theoretische Verteilungen an den aktuellen Datensatz anzupassen. Die Erstellung dieser Verteilungen geschieht mit einer Optimierung durch Simulated Annealing basierend auf Rohdaten aus einem simulierten Spektrum oder einem bereits zugeordneten experimentellen Spektrum.

Somit kann dieses Kapitel in zwei weitere Teile unterteilt werden:

- Erweiterung des bestehenden Algorithmus durch die Erfassung zusätzlicher Signal-Eigenschaften
- Erstellung von theoretischen Verteilungen basierend auf Referenz-Spektren, welche auf ähnliche Spektrumtypen angewendet werden können.

In dieser Arbeit bildet das Protein *PfTrx* in simulierter und in experimenteller Form hauptsächlich als zweidimensionales NOESY-Spektrum die Grundlage zur Generierung der Daten.

2 Materialien und Methoden

2.1 Software

2.1.1 Das Softwareprojekt AUREMOL

Am Institut für Biophysik und physikalische Biochemie der Universität Regensburg wurde 1999 eine Kooperation mit dem Spektrometerhersteller Bruker BioSpin GmbH begonnen. Dabei wurden C-Bibliotheken des AMIX-Viewer in AUREMOL eingebunden, so dass die durch das Institut erstellten Algorithmen in AUREMOL im AMIX-Viewer genutzt werden konnten. Da sich die Firma Bruker jedoch dazu entschlossen hat, die Weiterentwicklung des AMIX-Viewers einzustellen, war es ein zentrales Ziel, alle Funktionalitäten von dieser Basis zu lösen. Dies sollte es ermöglichen, Module zukünftig in einem alternativen Viewer einbinden zu können.

Die Ausgangssituation dieser Arbeit war das Softwarepaket AUREMOL in der Version 2.3.1 vom September 2009, welches komplett in ANSI-C implementiert war und direkt gegen die C-Bibliotheken des AMIX-Viewers gelinkt war. Diese Verbindung sollte aufgelöst werden und durch einen neuen objektorientierten Ansatz abgelöst werden. Zudem wurden in den neuen Modulen viele Schnittstellen bereitgestellt, welche den Datenzugriff auf die Rohdaten für multidimensionale Spektren erlaubt. Daher wurde die Neuimplementierung auf einen Datenzugriffsansatz mit nur einer zentralen Schnittstelle ausgerichtet. Die Algorithmen wurden stets durch eine rekursive Implementierung realisiert.

2.1.2 Die verwendete Entwicklungsumgebung, Framework und Bibliotheken zur Weiterentwicklung des Softwarepakets AUREMOL

Als Entwicklungsumgebung diente Microsoft Visual Studio 2008 bis 2013 mit dem C++-Framework Qt der Firma „The Qt Company“.

Zur Entwicklung wurde ein DELL Optiplex G655 unter Windows 7 bis Windows 8.1 verwendet. Alle Berechnungen dieser Arbeit wurden auf einem Dell PowerEdge-Server unter dem Betriebssystem Windows Server 2012 Datacenter 64-bit durchgeführt.

Zur Erstellung der Plots in dieser Arbeit wurde die Statistiksoftware R-Projekt in der Version 3.2.1 benutzt. Für die Erstellung der R-Protokolle wurde im Rahmen dieser Arbeit ein neues Modul in AUREMOL erstellt, welches aus den Daten ein R-Protokoll generiert.

Zur Visualisierung der Volumen und der Spektrumrohdaten wurde die Qt-Bibliothek QWTPlot3D verwendet, für die wiederum eine Schnittstelle geschaffen wurde. Für die Darstellung der zweidimensionalen Plots der Wahrscheinlichkeitsdichteverteilungen wurde die Bibliothek QWTPlot2D verwendet.

2.2 Die Teststruktur des Proteins *Plasmodium falciparum Thioredoxin* (*PfTrx*) als Grundlage der Datenbasis

Thioredoxine stellen eine Gruppe von redoxaktiven Proteinen dar, welche an zellulären Redox-Prozessen teilnehmen. *Thioredoxin*, *Glutaredoxin*, und *Tryparedoxin* haben gleiche Funktionalität und gehören alle der Überfamilie *Thioredoxin* an.

Ein Vorkommen von *Plasmodium falciparum* (funktionelles *Thioredoxin*) wurde im Malaria-Parasiten nachgewiesen. Daher wird dieses als ein wichtiges Protein für die Medikamentenforschung betrachtet und wurde häufig verwendet, den Nachweis eines Novel-22-kDa redox-aktiven Proteins im *P. falciparum* zu erbringen. Das Protein *Plasmoredoxin* (*Plrx*) ist der Superfamilie der *Thioredoxine* zuzuordnen. *Thioredoxin* ist hochkonserviert und einzigartig für die Malaria-Forschung und birgt ein sehr hohes Potential zur Verwendung in der Diagnostik.

Die Aufreinigung des Proteins und die Expressionsprotokolle wurden von Prof. Claudia Munte an der Universität in São Paulo in Brasilien erstellt und optimiert (Munte et al. 2009). Die Probe hat einen sehr hohen Anteil an Dithiothreitol (DTT) (10 *mM* in der finalen Konzentration), welcher die Ausbildung von Disulfidbrücken (zwischen Cys30 und Cys33 des aktiven Zentrums) verhindert. Dadurch tritt weder eine Oxidation noch eine unerwünschte Aggregation und Ausfall des Proteins ein. Zudem liegt *PfTrx* durch das DTT in reduzierter Form vor. Das zweidimensionale NOESY-Spektrum dieses Proteins diente weitgehend als Referenzspektrum in dieser Arbeit.

Zur Aufnahme des ^{1}H-^{1}H-NOESY-Spektrums diente eine Probe, welche 1 mM aus dem unmarkierten reduzierten PfTrx beinhaltete. Die Lösung bestand aus 92 % H_2O und 8 % D_2O bei einem pH-Wert von 7, welcher durch die Zugabe von 10 mM Puffer (Kaliumphosphat) erreicht wurde. Zur Referenzierung beinhaltete die Probe 0,1 mM DSS. Zudem enthielt die Probe 1 mM Natriumazid (NaN_3).

2.2.1 Die zur Aufnahme der Spektren verwendete Spektrometer

Von dieser Probe wurden mehrere NMR-Spektren bei einer Temperatur von 293 K aufgenommen. Dazu wurden die Spektrometer DRX-600, DRX-800 und DRX-900 von der Firma Bruker verwendet. Diese Spektrometer sind ausgestattet mit vier Radio-Frequenz-Kanälen und einem Triple-Resonanz-Kryoprobenkopf, welcher mit einer abgeschirmten z-Gradienten-Wicklung versehen ist.

2.2.2 Prozessierung der Spektren

Die Prozessierung der Spektren der Zeitdomäne erfolgte mit dem Softwarepaket TOPSPIN, welches von dem Spektrometer-Hersteller Bruker stammt. Die dadurch erhaltenen Daten in der Frequenzdomäne wurde im Softwarepaket AUREMOL weiter verarbeitet.

Alle chemischen Verschiebungen der Protonen wurden direkt auf die ^{1}H -Resonanzfrequenz der Methylgruppe des DSS referenziert. Die Resonanzen von ^{13}C und ^{15}N wurden unter Beachtung der IUPAC-Empfehlungen (Markley et al. 1998) indirekt referenziert. Die sequentiellen Rückgrat-Zuordnungen wurden auf der Basis der HNCA, HN(CO)CA, CBCA(CO)NH, CBCANH, ^{15}N-HSQC und ^{15}N-NOESY-HSQC-Experimente getroffen. Die Zuordnung der Seitenketten wurde von Prof. Clauda Munte erstellt. Im Folgenden wird für diese Arbeit größtenteils das homonukleare, zweidimensionale ^{1}H-^{1}H-NOESY-Spektrum des Proteins *PfTrx* (Abb. 1) zur Generierung der nötigen Datenbasis herangezogen.

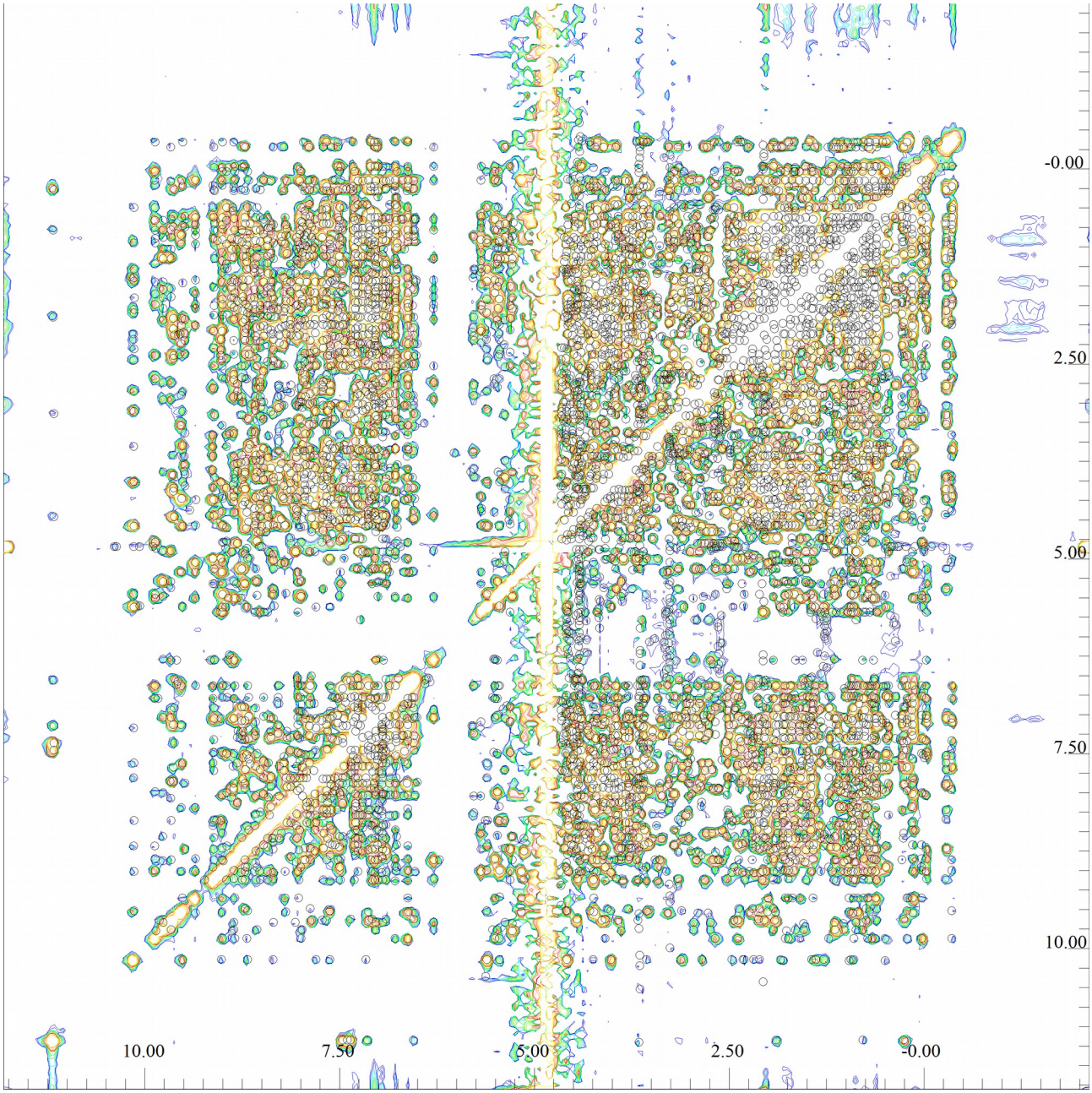

Abb. 1: Zugeordnetes (6738 zugeordnete Peaks) homonukleares, zweidimensionales ^{1}H-^{1}H-NOESY-Spektrum mit einer digitalen Auflösung von 1024x2048 Pixeln; Resonanzfrequenz 800,2 MHz; spektrale Breite 11160,714 Hz bzw. 13,9486 ppm; Offset bei 11,8187 ppm; Mischzeit 0,1 s; verwendeter Filter: Gauß-Lorentz-Transformation; Temperatur 293 K

2.2.3 Definitionen

Im Verlauf dieser Arbeit werden folgende wichtige Begriffe verwendet:

Intensität

Die Intensität bezogen auf ein Signal beschreibt den Wert in Richtung der Intensitätsachse an der festgelegten Signalposition am digitalen Raster (also die Position bezüglich der Frequenz-Achsen). Daher ist im Folgenden, falls die Intensität nicht auf das Volumen bezogen ist, stets die Intensität an der Signal-Position gemeint. Im Zusammenhang mit der Volumeninformation sind Intensitäten die Werte, welche in ein Signal-Volumen eingehen und damit der Signalform in der Frequenzdomäne entsprechen.

Digitale Auflösung

Der Begriff digitale Auflösung beschreibt im Folgenden die Anzahl der prozessierten Datenpunkte (SI).

$$SI = \frac{SW}{DR} \tag{3}$$

Wobei SW (spectral width) die spektrale Breite des Frequenzfensters in Hz und DR (digital resolution) die Frequenz zwischen benachbarten Datenpunkten in Hz/point beschreibt.

Masterliste

Die Masterliste (oder auch Peakliste) umfasst alle in einem Spektrum definierten Signale u. a. mit deren Eigenschaften Position (SI, ppm), Intensität, Volumen, Wahrscheinlichkeit, usw. Dies wird üblicherweise auch als NMR-Signal (oder Peak) bezeichnet und ist das Frequenzsignal bzw. die chemische Verschiebung resultierend aus der Fouriertransformation des FID‘s. Dieser kann durch Aufspaltung auch mehrere relative Extrema neben einem Hauptmaximum in seiner Form beinhalten.

2.2.4 Simulation eines zweidimensionalen ^{1}H-^{1}H-NOESY-Spektrums von *PfTrx* mit Rauschen

Zur Generierung der theoretischen Verteilungen, welche für die Diskriminierung benötigt werden bzw. zur Untersuchung der Verbesserungen der Integration wurde ein zweidimensionales ^{1}H-^{1}H-NOESY-Spektrum mit dem AUREMOL-Modul RELAX (Görler

und Kalbitzer 1997; Ried et al. 2004) in der Frequenzdomäne simuliert. Dabei wurden folgende Parameter für diese Simulation verwendet:

Tabelle 1: Übersicht Parameter für das simulierte zweidimensionale ^{1}H-^{1}H-NOESY-Spektrum durch das Modul RELAX für die Integration.

Linienform	Gauß
LB (Linienverbreiterung)	-6 Hz
GB (gaussbroadening-factor)	0,1 % des FID
Cutoff-Distanz (Maximaler Abstand der Kerne, welche noch in die Simulation mit einfließen)	0,5 nm
Mischzeit	0,3 s
Wiederholzeit	1,54 s
Larmorfrequenz	600,13 MHz
Linienbreite	25 Hz
Rausch-Anteil	1 %
digitale Auflösung (siehe 2.2.3)	1024x2048

Da in der Regel ein Spektrum immer Rauschen aufweist, wurde dem simulierten Frequenzspektrum ein Rauschanteil von 1 % hinzugefügt. Dies ergibt sich aus der Standardabweichung des Rauschens, welches durch das Signal (also dessen Intensität) vorgegeben wird, welches am nächsten bei 0,3 nm anzutreffen ist (Brunner 2006) .

2.3 Die wichtigsten Erweiterungen der Basisfunktionen von AUREMOL

2.3.1 Die Spektrumrohdaten der Frequenzdomäne

Um NMR-Signale in einem Spektrum zuordnen zu können, müssen die vom Spektrometer erstellten Rohdaten in AUREMOL eingelesen werden. Diese Rohdaten stellen die prozessierten Frequenzdomänendaten aus den entsprechenden Zeitdomänendaten dar. Diese Frequenzdaten werden als Binär-Dateien, welche die Intensitäten beinhalten, gespeichert. Der Aufbau der Datei wurde von Bruker rekursiv definiert. Dies hätte eine Umsetzung einer rekursiven Einleseroutine nahe gelegt. Jedoch wurde für eindimensionale, zweidimensionale, dreidimensionale und vierdimensionale Spektren jeweils eine eigene Einleseroutine der Rohdaten von AMIX zur Verfügung gestellt.

Die Motivation war nun, diese Funktionen zu einem zentralen objektorientierten, rekursiven Ansatz überzuführen und gleichzeitig von AMIX loszulösen, um zukünftige einfache Erweiterung durch eine geringere Fehleranfälligkeit zu gewährleisten. Dies wurde realisiert und für die in dieser Arbeit erweiterten Module verwendet.

2.3.2 Das Modul Maximum Peak-Picking

Steht ein bereits prozessiertes Spektrum für AUREMOL zur Verfügung, müssen die Signale für eine spätere Strukturbestimmung identifiziert werden.

Da in AUREMOL das sog. „Peak-Picking“ in mehreren Versionen als Kopie vorlag, wurden im Rahmen dieser Arbeit alle Peak-Peaking-Module lokalisiert und ausgetauscht. Diese lagen sowohl als komplette Kopien vor, als auch als Teilkopien für die jeweilig benötigte Dimension (also für eindimensionale, zweidimensionale, dreidimensionale und vierdimensionale Spektren). So existierte für eindimensionale bis vierdimensionale je ein eigenes Modul, was eine funktionelle Erweiterung sehr aufwändig machte, da der Quellcode mehrfach vorlag. Dies führte zur Inkonsistenz der verschiedenen Implementierungen. So wurden Fehlerbehebungen nur bei einer Kopie vorgenommen (z. B. eine Änderung am Peak-Picken eines zweidimensionalen Spektrums) und an anderer Stelle (bei einem dreidimensionalen Spektrum) vernachlässigt. Daher war es nötig, die Funktion des Maximum-Peak-Picking zu zentralisieren.

2.3.3 Zusammenfassung mehrerer Signale zu einem Signal

Wurden in einem Spektrum Signale gepickt, bei dem Multipletts vorkommen, war es nicht möglich, diese zu einem Signal zusammenzufassen, ohne dass die Information der Positionen der einzelnen Signale verloren ging. So musste eine Möglichkeit gefunden werden, Signale derart zusammenzufassen, ohne dass deren ursprüngliche Positionsinformation verloren geht.

Dazu wurde die Funktion „merge Peaks“ in AUREMOL eingeführt, welche die zusammenzuführenden Signale in der Masterliste zu einem Signal zusammenfasst und dabei eine neue Position generiert, welche auf dem geometrischen Zentrum der ursprünglichen Signale basiert. Die ursprünglichen Positionen (in ppm) der zusammengefassten Signale sind in der Masterliste unter dem Schlüssel SUBPEAKS wiederzufinden. Ein solch zusammengefasstes Signal stellt auch für die erweiterte Integration kein Problem mehr dar, da diese die Information aus den SUBPEAKS-Schlüssel der Masterliste nutzt, um die Volumen korrekt den Signalen zuordnen zu können.

2.3.4 Die Verschiebung von Signalpositionen zu deren nächstgelegenem Extremum

Wurde ein Signal z. B. manuell gepickt oder durch das Modul RELAX berechnet, konnte es sein, dass die Positionen von diesen Signalen nicht auf dem Extremum der zugrunde liegenden Signalform lagen. Für diese Signale war es nicht möglich, die Positionen nachträglich automatisch auf das am nächsten liegende Extremum zu transferieren, ohne dass dessen Zuordnung verloren geht. Auch die manuelle Verschiebung an ein Extremum, welches dann auch die Zuordungsinformation behalten würde, war umständlich, da die Verschiebung jedes einzelnen Signals manuell durch die Angabe der gewünschten neuen Position sehr aufwändig war. Dies war der Grund, die Routine „move to nearest maximum“ zu entwickeln.

Diese Routine wurde ebenfalls von der erweiterten Integration verwendet und daher werden die Ergebnisse dieser Methode durch die Resultate der Integration wiedergegeben.

2.4 Die Grundlagen zur Bestimmung der Rohdaten und die Erweiterungen zur Verbesserung der Bestimmung der Signalvolumen

2.4.1 Die Erfassung der Volumendaten durch Integration der NMR-Signale

Das Integrationsmodul der AUREMOL-Version 2.3.1 wurde als Grundlage (Geyer et al. 1995) für alle Erweiterungen und Verbesserungen im Rahmen dieser Arbeit verwendet (siehe Segmentierung Kapitel 1.1).

Da die ursprüngliche Integration aus der Version 2.3.1 nach Zuweisung aller Intensitäten durch die Segmentierung lediglich die Intensitäten der resultierenden Seedliste zu einem Volumen aufaddierte aber alle zusätzlichen Informationen wieder verworfen hatte, wurde eine Struktur erstellt, welche die Segmentierungs-Informationen behält. Diese Struktur diente als Basis für schnelle Zugriffe des Softwarepakets AUREMOL auf die Volumen selbst und den volumenbasierten Informationen der Signale aus der Peakliste.

Die Motivation war nun die verworfenen Volumenanteile (Einzelintensitäten) auch nachträglich ohne erneute Integration zugänglich zu machen. Um später auf diese Information wieder zugreifen zu können, wurden im Integrationsmodul während der Segmentierung die einzelnen Intensitäten abgegriffen und einzeln zugehörend zu jedem Signal separat in der Struktur *Integrations-Hash* gespeichert. Die Erstellung dieser Struktur war essentiell, um die Erweiterungen in dieser Arbeit überhaupt realisieren zu können.

Die Struktur *Integrations-Hash* setzt voraus, dass jedes Signal aus der Peakliste einen eindeutigen Schlüssel (die Peak-ID) bei der Erstellung der Peakliste erhält. Diese Peak-ID stellt den Hash-Schlüssel dar. Zu jedem Hash-Schlüssel existiert ein Hash-Wert, welcher in diesem Fall aus einer Liste mit den zu dem Signal zugeordneten Positionen der einzelnen Intensitäten (des Volumens) besteht.

Eine erheblicher Vorteil dieser zusätzlichen Struktur ist, dass alle am Volumen eines Signals beteiligten Positionen der Intensitäten abgespeichert sind und jederzeit ohne Neuintegration abgefragt werden können. Wird ein Volumen mit einer anderen Segmentierung ausgelesen, werden die Positionen mit den Intensitäten des Spektrums synchronisiert und Intensitäten unter dem Segmentierungslevel ignoriert. Diese gefilterten Intensitäten ergeben dann das neue Gesamtvolumen. Einen weiteren Vorteil bietet diese

Struktur dahingehend, dass es ermöglicht wird, das Rohspektrum einfach auszutauschen, jedoch die zugeordneten Positionen weiter zu benutzen. Diesen Effekt macht sich sowohl die Integration durch Glättung zunutze, um stark verrauschte Spektren integrieren zu können als auch die Bestimmung des Schwerpunkts der Signalposition, welche alle notwendigen Informationen aus diesem *Integrations-Hash* beziehen kann. Hierzu finden sich in der Klasse *CIntegrations-Hash-Analyzer* entsprechende Schnittstellen.

2.4.2 Visualisierung des Signalvolumens

Um die Volumenbeiträge der Signale zur Integration visuell beurteilen zu können, wurde eine Methode entwickelt, welche Signale von ein- und zweidimensionalen Spektren in einer Grafik darstellen kann. Damit hat der Benutzer die Möglichkeit, die Integrationsergebnisse noch einmal stichprobenartig zu verifizieren. Diese Methode wurde in dieser Arbeit intensiv genutzt, um die Ergebnisse der verschiedenen erweiterten Integrations-Modi darzustellen.

Hierzu wurde die C++-Bibliothek QWTPlot3D verwendet. Diese benötigt zur Visualisierung zwei Matrizen, wobei die erste Matrix aus den Intensitäten und die zweite Matrix aus den Farbcodes besteht. Um diese Matrizen zu erhalten, muss die Volumeninformation eines Signals aus dem *Integrations-Hash* ausgelesen werden. Da zu jedem Signal bei der Integration alle zugehörigen Positionen der Intensitäten gespeichert wurden, können die zugehörigen Intensitäten an deren jeweiligen Position mit der Klasse *CFileWindow* in eine Matrix aus Intensitäten übergeführt werden, welche die Projektion der Intensitätsachse auf die Frequenzachsen darstellt.

Diese Intensitäten der Matrix wurden unter Einbeziehung der Segmentierungsschwellen farblich dargestellt. Die positiven Intensitäten werden rot (respektive die negativen Intensitäten grün) dargestellt, falls diese über (bzw. unter) dem Segmentierungslevel liegen.

2.4.3 Automatische Größenermittlung des Integrationsbereiches für ein Signal aus der Peakliste und variable Erhöhung der Integrationsschritte

Um die Bestimmung der Größe des Integrationsbereiches zu optimieren, wurde im Rahmen dieser Arbeit eine dynamische Bereichseingrenzung (die sog. Integrationsbox) der relevanten Signalformen entwickelt. In der ursprünglichen Version wurde die Größe des Bereichs entweder manuell oder basierend auf der Nachbarschaft der gepickten

Signale festgelegt. Da aber die Größe der Integrationsbox viel zu groß war, führte dies zu hohen Laufzeiten der Integration, da jedes Extremum innerhalb der Integrationsbox in der Seedliste als Wachstumskeim aufgenommen werden muss. So enthielt die Integrationsbox viele Extrema, welche zur Bestimmung des Volumens des aktuellen Signals irrelevant waren.

Daher wurde der Ansatz der adaptiven Vergrößerung des Integrationsbereichs durch Mehrfachintegrationen verfolgt um nur für das Volumen relevante Extrema zu berücksichtigen. Zusätzlich soll die Abgrenzung überlappender Signale verbessert und auch das Umwachsen eines Nachbarschafts-Extremums innerhalb des Integrationsbereichs vollständig verhindert werden, falls die Integrationsschritte des Wachstumsalgorithmus nicht ausreichen würden (Standard waren 10 Iterationsschritte).

Dieser Ansatz ist in zwei Stufen aufgebaut:

1. Stufe: Es werden vom Algorithmus mehrere Integrationsläufe eines jeden Signals durchgeführt und nach jeder Integration überprüft, ob eine Intensität am Rande der Integrationsbox noch mit in das Volumen des aktuell integrierten Signals eingeht. Falls dies der Fall ist, wird die Integrationsbox um 30 % (Standardwert) in jede Richtung des Bereichs, in dem ein Volumenanteil am Rand anzutreffen ist, vergrößert. Dies wird solange wiederholt, bis nach einer Vergrößerung der Integrationsbox keine Intensitäten am Rand des Integrationsbereiches mehr in das Volumen mit einfließen. Als Startgröße hat dabei die Integrationsbox eine Ausdehnung von je drei Pixeln in jede Frequenzachse mit Positionen der Intensität des Signals als Mittelpunkt.
2. Stufe: Da der Wachstumsalgorithmus mit den üblich verwendeten 10 Integrationsschritten nicht bei jedem Signal verhindert, dass der Wachstumsalgorithmus einen anderen initialen Seed (anderes Extremum in der aktuellen Integrationsbox) ansatzweise umschließt, wurde eine automatische Erhöhung der Integrationsschritte eingeführt. Dazu wird direkt nach der Bestimmung der korrekten Größe der Integrationsbox (Stufe 1) eine Erhöhung der Iterationsschritte um den Wert 10 durchgeführt. Danach wird das vor der Erhöhung der Integrationsschritte ermittelte Volumen mit dem Volumen nach der Erhöhung verglichen. Falls eine Veränderung von mehr als 10 % vorliegt, werden

Iterationsschritte um weitere 10 Schritte erhöht. Dies wird solange wiederholt, bis die Änderung des Volumens 10 % unterschreitet.

2.4.4 Verbesserung der Integration stark verrauschter Signale

Eine weitere Schwäche der ursprünglichen Integration stellten stark verrauschte Spektren dar, da diese bei stark verrauschten Signalen stets zu einem Abbruch des Wachstumsalgorithmus führten. Der Grund war, dass der Algorithmus zerklüftete Konturen als Nebenextrema interpretierte und damit zur falschen Seedliste (nämlich die der Nebenextrema) zuordnete. Daher wurde in dieser Arbeit ein Glättungsverfahren eingeführt, welches das Spektrum temporär mit einstellbarer Größe eines Gauß-Glättungsfaktors (Standardmäßig 3 Pixel in jede Dimension) glättet. Ziel dabei war es, die Zerklüftung zu vermindern, so dass es dem Wachstumsalgorithmus möglich war, die durch Rauschen entstandene Nebenextrema zu überwinden, da die Kontur temporär glatt war. Die Integration wird dann mit der in diesem Kapitel aufgeführten Verbesserungen durchgeführt. Da durch die Glättung kaum eine Signalposition an seinem ursprünglichen Extremum der Intensitäten verbleibt, greift in diesem Fall die Methode 2.3.4, um die Integration von Signalen, deren Position nicht an einem Extrema liegt, zu ermöglichen.

Da die Integration alle Positionen der Intensitäten eines Signals separat in der Struktur *Integrations-Hash* speichert, kann nach der erfolgreichen Integration das geglättete Spektrum wieder verworfen werden. Danach wird das Original-Spektrum wieder eingelesen, damit das Signalvolumen anhand der Positionen (Koordinaten) der Intensitäten des Original-Spektrums erstellt werden kann. Da im *Integrations-Hash* lediglich die Positionen der Pixel in den Rohdaten gespeichert werden, können anhand dieser Positionsangaben die Intensitäten aus dem ungeglätteten Spektrum ausgelesen werden und dem entsprechenden Signal zugeordnet werden. Letztendlich wird das geglättete Spektrum dazu benutzt, die Volumenfläche anhand der geglätteten Signalkontur zu bestimmen.

2.4.5 Integration mehrerer Signale an gleicher Position

Falls sich zwei oder mehrere Signale eine Position am digitalen Raster teilen, wurde jedem Signal in der ursprünglichen Integration dasselbe Volumen zugewiesen. Um dies zu korrigieren, musste die Anzahl der Signale an der Position auf eins reduziert werden und

in den Signaleigenschaften der Peakliste nachgepflegt werden. Der Algorithmus wurde daher so gestaltet, dass es nicht nötig ist, manuelle Mehrfachzuordnungen festzulegen.

Dazu wurden zu Beginn der Integration alle Signale dahingehend untersucht, ob diese an einem Extremum mehrere Signale aus der Peakliste aufweisen. War dies der Fall, wurde des erste gefundene Signal belassen und alle weiteren Signale für den Algorithmus deaktiviert. Nach der Integration wurden diese deaktivierten Signale wieder aktiviert und mit einem prozentualen Teilvolumen in Abhängigkeit ihrer Intensität versehen. Die Information der beteiligten Intensitäten wurde jedoch redundant im *Integrations-Hash* unter Angabe seines prozentualen Anteils am Gesamtvolumen gespeichert, um später das richtige Volumen wieder extrahieren zu können. Somit gilt für die relativen Anteile R_i der Einzelvolumen V_i bei N Signalen, welche am Gesamtvolumen V an einer Signalform beteiligt sind:

$$R_i = \frac{I_i}{\sum_{s=1}^{N} I_s} \tag{4}$$

Damit gilt für das Volumen V_i eines korrigierten Signals:

$$V_i = V\,R_i \tag{5}$$

Damit gibt die Summe dieser Einzelvolumen V_i wieder das Gesamtvolumen V der Signalform, welche die Integration segmentiert hat. Hierzu finden sich in der Klasse *CIntegrations-Hash-Analyzer* entsprechende Funktionen, welche es erlauben diese Volumen lt. Formel 5 oder falls gewünscht auch ohne relative Volumenanteile auszulesen.

2.4.6 Integration von Signalen, deren Position nicht an einem Extremum der Signalform definiert ist

Dies ist eine wichtige Erweiterung der Integration, da diese nun die Möglichkeit bietet, auch Signale zu integrieren, welche nicht am Extremum der Signalform positioniert sind. Dies war vor allem dann nötig, wenn ein Spektrum simuliert wurde, bei dem sich die theoretischen Signalpositionen nicht zwingend am Extremum am digitalen Raster der Signalform befinden. Falls Positionen von Signalen manuell gesetzt bzw. manuell

verschoben wurden, können diese nun durch diese Erweiterung auch integriert werden. Die ursprüngliche Version der Integration lieferte in all diesen Fällen ein Volumen von 0, da der Algorithmus ein gepicktes Signal-Extremum am digitalen Raster der Signalform voraussetzte.

Um auch diese Signale integrieren zu können, wurde die Idee zur Verschiebung der Position an das nächste Extremum der Signalform mit in den Algorithmus aufgenommen. Dazu wurde die Erstellung der Seedliste für die Segmentierung derart erweitert, dass diese auch Signale ohne Position an einem Extremum aufnehmen darf.

Dazu wird vor Beginn der Integration ein jedes Signal daraufhin geprüft, ob dieses am Extremum positioniert ist. Falls nicht, wird die nächste Extremum-Position gesucht und sowohl das Signal als auch dessen nächst gelegene Extremum-Koordinate in einer Nicht-Extremum-Liste gespeichert.

Es kann vor der Integration vom Benutzer gewählt werden, wie die Signale mit Position ohne Extremum am digitalen Raster der Signalform gehandhabt werden sollen:

- Alle Signale werden an ihr nächstes Extremum gesetzt und analog zu 2.4.5 behandelt. Hier trägt die Intensität der zu verschiebenden Signale aus der Peakliste entscheidend zur Volumenaufteilung bei.
- Nur der nächste Nachbar aus der Peakliste zum Extremum erhält die Position des Extremums. Die restlichen Positionen, welche nicht auf einem Extremum liegen, bleiben unverändert. Hier muss der Algorithmus als erstes überprüfen, ob bereits ein Signal aus der Masterliste am Extremum liegt. Ist dies der Fall, werden keine der Nicht-Extremum-Signale verschoben. Ist die Position jedoch frei, wird den nächsten Nachbarn des Maximums erlaubt, dahin zu wandern, wobei der Rest der potentiell zu verschiebenden Signale an ihrer Stelle verbleiben. Diese blockierten nicht-Extremum-Signale werden dann in die Seedliste aufgenommen und vom Algorithmus als Pseudo-Extrema behandelt. Daher kommt es zu dem Effekt, dass das Volumen an den Positionen der nicht-Extremum-Signale abgeschnitten wird, da der Algorithmus nur absteigendes Wachstum erlaubt.

2.4.7 Integration von Multipletts und zusammengefassten Signalen

Da es möglich ist, mit dem Modul RELAX auch Multipletts zu simulieren, wurde die Integration erweitert, um sowohl Signale mit einer Multiplett-Aufspaltung aus der Simulation als auch zusammengefasste Signale (2.3.3) integrieren zu können.

Dazu wurde die Masterliste temporär so aufbereitet, dass die Signale ohne Multipletts vorlagen. Um dies zu erreichen, wurden die Multipletts aus der ursprünglichen Masterliste in einzelne Signale zerlegt. Diese Masterliste kann dann direkt integriert werden. Am Ende der Integration werden die Signale wieder zusammengefasst. Da während der Integration die Struktur *Integrations-Hash* erstellt wird, liegen dort die Signalpositionen der gesplitteten Signale vor. Daher muss dieser nach dem Zusammenfügen bereinigt werden, indem die Positionen der Intensitäten der gesplitteten Volumen ebenfalls zusammengeführt werden.

2.5 Die Berechnung des Schwerpunkts zur Verbesserung der Positionsbestimmung von NMR-Signalen

Bislang war es im Softwarepaket AUREMOL auf drei Arten möglich, die Position eines Signals zu bestimmen:

- automatisches Peak-Picking legt die Position des Signals an das Extremum im digitalen Raster der Signalform fest
- manuelles Picken eines Signals im Spektrums (Position am Extremum oder nicht)
- manuelle nachträgliche Änderung der Position des Signals

Eine Möglichkeit zur Verbesserung der Signal-Positionen für die automatische Strukturbestimmung (entweder Bereiche eines Spektrums oder alle Signale des Spektrums aus der Peakliste) ist die Berechnung der Signal-Position basierend auf dem Massenschwerpunkt des Volumens eines Signals. Hierzu werden alle Intensitäten, die einen Beitrag zum Signalvolumen liefern zur Berechnung herangezogen.
Es gilt die allgemeine Formel für die Bestimmung des physikalischen Schwerpunkts, falls alle Massenteile diskret vorliegen:

$$\vec{S_P} = \frac{\sum_{n=1}^{N} I_n * \vec{S_n}}{\sum_{n=1}^{N} I_n} \tag{6}$$

Wobei N die Gesamtzahl aller homogenen Einzelmassen beschreibt. Der Vektor $\vec{S_P} \in \mathbb{R}$ legt die Koordinaten des physikalischen Schwerpunkts fest. I_n beschreibt die einzelnen Massenteile (also die Einzelintensitäten) mit $\vec{S_n} \in \mathbb{N}$ an deren jeweiligen Positionen am digitalen Raster.

Die Segmentierungstiefe dient hier sowohl zur Festlegung der Ausdehnung des Bereichs der die in das Volumen eingehenden Intensitäten umfasst als auch für die Höhe der Abschneidung des Volumens am Segmentierungslevel. Der Wert unter der Segmentierung wird daraufhin von allen beteiligten Intensitäten, welche in die Berechnung des Schwerpunkts eingehen, subtrahiert oder weggelassen.

Damit ergibt sich die Definition für den Schwerpunkt der reduzierten Einzelmassen (Intensitäten bzw. Voxel) bei einer Segmentierungstiefe S_l:

$$\vec{S_s} = \frac{\sum_{n=1}^{N} \left(|I_n| - |I_{max}| * S_l \right) * \vec{S_n}}{\sum_{n=1}^{N} \left(|I_n| - |I_{max}| * S_l \right)} \tag{7}$$

Wobei N die Gesamtzahl aller Voxel beschreibt. Der Vektor $\vec{S_s} \in \mathbb{R}$ legt die Koordinaten des physikalischen Schwerpunkts auf der ppm-Skala fest. Der Beitrag $|I_n| - |I_{max}| * S_l$ beschreibt die reduzierten Einzelintensitäten an deren jeweiligen Positionen. I_n beschreibt die einzelnen Massenteile (also die Einzelintensitäten) mit $\vec{S_n} \in \mathbb{N}$ an deren jeweiligen Positionen am digitalen Raster und I_{max} die Intensität am Extremum des Signals.

2.6 Signalidentifizierung durch die Bestimmung der Bayesschen Wahrscheinlichkeit von NMR-Signalen

2.6.1 Die Ausgangssituation des Moduls der Bayesschen Wahrscheinlichkeits-Berechnung

In der ursprünglichen Fassung des Bayesschen Peak-Picking-Algorithmus (Antz et al. 1995) wurden folgende Eigenschaften aller Peaks zur Diskriminierung von NMR-Signalen (im Folgenden nur als Signale bezeichnet) durch das Protein und Störsignalen durch Rauschen, Artefakte oder Wasserstreifen erfasst:

- Intensität des NMR-Signals
- Verhältnis des Volumens bei einer Segmentierungstiefe von 0,5 zur Intensität des NMR-Signals
- Verhältnis des Volumens bei einer Segmentierungstiefe von 0,2 zur Intensität des NMR-Signals
- Verhältnis des Volumens bei einer Segmentierungstiefe von 0,01 zur Intensität des NMR-Signals
- Peak-Symmetrie bezüglich der diagonalen NMR-Signals **(äußere Symmetrie)** (Schulte et al. 1997)

Dabei wurden die drei Volumeninformationen zu einem multivariaten diskriminierten Volumen zusammengefasst, so dass sich insgesamt drei Eigenschaften für die Berechnung der Bayesschen Wahrscheinlichkeit ergeben.

Die Motivation war nun die, dass durch eine Erweiterung der Anzahl von Signaleigenschaften und die Erweiterung der verwendeten Klassen *Signal* und *Rauschen* durch die Klasse *Wasser* eine bessere Klassifizierung erreicht werden kann.

Die Erhöhung der Anzahl der verwendeten Eigenschaften sollten schlechte Beiträge anderer Eigenschaften (bedingt durch die Rohdaten) durch zusätzliche Eigenschaften ausgeglichen werden, was zu einer Verbesserung und einer Erhöhung der Stabilität bezüglich der Diskriminierung führt.

2.6.2 Geglättete Wahrscheinlichkeitsdichteverteilungen zur Berechnung der Bayesschen Wahrscheinlichkeiten

Zur Berechnung der Bayesschen Wahrscheinlichkeit wurden geglättete Wahrscheinlichkeitsdichteverteilungen verwendet (Schulte et al. 1997).

Diese Methode wurde aufgrund der häufigen Verwendung in dieser Arbeit bezüglich deren Laufzeit verbessert, fehlerbereinigt und durch einen adaptiven Glättungsfilter erweitert. Zudem kann die resultierende Verteilung mittels der C++-Bibliothek QWTPlot direkt in AUREMOL visualisiert werden. Zudem wurde die Klasse *CRPlot* implementiert, welche aus den Rohdaten der geglätteten Verteilung ein R-Skript generiert, welches dann im Softwarepaket R zur besseren Visualisierung der Verteilung ausgeführt werden kann.

2.6.3 Bestimmung der Parameter aller Eigenschafts-Verteilungen durch Optimierung der Maximum-Likelihood-Funktion mittels Simulated Annealing

Da die ursprünglichen geglätteten Verteilungen den Nachteil hatten, dass sie nicht auf andere Spektren übertragbar sind, weil die Funktionen der verschiedenen Wahrscheinlichkeitsdichteverteilungen nicht bekannt waren, war es nötig, die Parameter der Verteilungen eines richtig zugeordneten Spektrums zu ermitteln, um diese auf ein anderes noch nicht zugeordnetes Spektrum übertragen zu können.

Die Annahme war, dass die Wahrscheinlichkeitsdichteverteilungen der Signal-Eigenschaften einer Verteilungen der Kombination aus Normalverteilung (N) und logarithmischer Normalverteilung (LOGN) entsprechen.
Dazu wurden die freien Parameter der entsprechenden Maximum-Likelihood-Funktion mittels Simulated Annealing optimiert. Die entsprechende Energielandschaft wurde durch die Rohdaten der Signaleigenschaften des Spektrums gebildet.

Die Motivation zur Anwendung theoretischer Verteilungen war, dass bei Vorliegen einer theoretischen Verteilung zu einem bekannten Spektrumtyp, diese auch auch auf ein neu automatisch gepicktes Spektrum desselben Typs angewendet werden soll. Zudem ist es möglich, dass nur sehr wenig NMR-Signale vorhanden sind, was es unmöglich macht eine aussagekräftige Verteilung zu generieren. Ein weiteres Problem lag darin, dass die markierten Signale auch Störsignale enthalten. Diese Klasse *Signal* musste daher grob bereinigt werden. Dies geschieht durch das Entfernen von Störsignalen mit einer geringern

gaußschen Wahrscheinlichkeit aus dieser Klasse. Der restliche Signaldatensatz wurde dazu verwendet, adäquate Anfangswerte für die Anpassung der theoretischen Verteilungen zu bestimmen. Dabei geht die Breite der theoretischen Verteilungen nicht verloren und eventuell verbleibende Störsignale in der Klasse *Signal* werden mit einer geringen Wahrscheinlichkeit bewertet.

2.6.4 Erweiterung des Moduls Bayessches Peak-Picking um weitere Eigenschaften und Einführung der Klasse *Wasser*

Durch das im Rahmen dieser Arbeit erweiterte Peak-Picken war es möglich, alle NMR-Signale eines Spektrums zu picken. Daher kann nun auch der Beitrag der Störsignale stärker in die Diskriminierung der Nutzsignale von den Störsignalen eingehen. Generell könnten in diesem neu entwickelten Modul beliebig viele Klassen definiert werden. Im Rahmen dieser Arbeit wurden die Klassen *Signal*, *Rauschen* und *Wasser* genauer untersucht und ausgewertet.

Jede dieser Klassen kann mit beliebig vielen Eigenschaften befüllt werden, da im Rahmen dieser Arbeit viel Wert auf Wiederverwendung auch für andere Methoden und Dynamik der C++ Klassen gelegt wurde.

Zu der Dynamik des entwickelten Moduls, die es nun erlaubt viele verschiedene Eigenschaften und Klassen zu verwenden, ist es auch möglich, die Berechnung der Bayesschen Wahrscheinlichkeit für NMR-Signale von Spektren beliebiger Dimension durchzuführen. Zusätzlich erlaubt diese Methode auch die Verwendung anderer Spektrumtypen (neben NOESY-Spektren). Dies ist möglich, da nun ausreichend viel Eigenschaften zur Klassifizierung vorhanden sind, denn die äußere Symmetrie ist bei heteronuklearen Spektren nicht mehr gegeben.

Bei der Einteilung der Klassen werden die entsprechenden Bereiche des Spektrums markiert. Diese Bereiche werden gespeichert und können in einer zentralen Datenbank verwaltet werden. Damit ist es möglich die Bereiche eines Spektrums für andere des gleichen Typs zu verwenden. Dies soll verhindern, dass der Benutzer falsche Bereiche festlegt und soll einen interaktiven Eingriff unnötig machen.

Die Markierungen der Bereiche (also der Klassen) können in diesem Modul im Falle von einer Dimension und zwei Dimensionen per Maus definiert werden. Bei

höherdimensionalen Spektren müssen jedoch ppm-Bereiche manuell durch die Eingabe in einer Maske deklariert werden.

Die Verwaltung und der Aufbau der Klassen und der Eigenschaften übernehmen zum Großteil die C++-Klassen *CLearndatasetBuilder* und *CSpectrumPropertyCalculator*.

Der Vorteil der Modularisierung der Klassen und die Definition einer entsprechenden Schnittstelle erlauben es, diese auch künftig für andere Klassifizierungsverfahren einzusetzen.

Hauptaufgabe dieser C++-Klassen ist, die Eigenschaften von NMR-Signalen aus einem Spektrum zu erfassen. Dazu werden alle Eigenschaften der NMR-Signale berechnet und in eine Hauptmatrix übertragen. Die Zeilen der Matrix repräsentieren dabei die verschiedenen NMR-Signale und die Spalten definieren deren Eigenschaftstypen (Intensität, Volumen, usw.).

Die markierten Bereiche aus NMR-Signalen, werden aus der Hauptmatrix in die jeweiligen Klassenmatrizen *Signal*, *Rauschen* und *Wasser* übertragen.

2.6.5 Evaluation der optimalen Parameter zur Generierung der Eigenschaften

2.6.5.1 Methoden zur Berechnung der Eigenschaften von Signalen für die Bestimmung der Bayesschen Wahrscheinlichkeit

Im Folgenden wurde das NOESY-Spektrum des Proteins *PfTrx* (siehe 2.2) verwendet, welches sowohl als Simulation aus dem Modul RELAX, sowie als bereits zugeordnetes NOESY-Spektrum vorliegt. Alle bereits richtig zugeordneten Signale wurden durch die Klasse *Signal* definiert. Um für die Klassifizierung weitere Klassen zu erhalten, wurden Störsignale durch das automatische Peak-Picking hinzugefügt, wobei die bereits zugeordneten Peaks nicht mehr gepickt wurden. Somit konnten die Klassen *Rauschen* und *Wasser* festgelegt werden.

Für die Bestimmung der optimalen Parameter wurden dann die folgenden Kombinationen der Klassen untersucht:

- *Signal* und *Rauschen*
- *Signal*, *Rauschen* und *Wasser*

Hierbei gibt es einen wesentlichen Unterschied der experimentellen Zuordnung zur simulierten Zuordnung. Denn RELAX liefert sowohl mehrere Zuordnungen an einer Position, als auch Peaks, welche nicht an einem Extremum liegen. Die Signale aus der Simulation wurden belassen und lieferten später die Grundlage für die Generierung der theoretischen Verteilungen.

Da für die Peak-Eigenschaften auch verschiedene Berechnungsmethoden verwendet wurden (z. B. Segmentierungslevel bei den Volumen), mussten die optimalen Parameter für jede Eigenschaft evaluiert werden. Dazu wurden sogenannte Hitlisten eingeführt, welche die optimalen Parameter für eine jede einzelne Eigenschaft wieder gibt.

Eine Erläuterung der verwendeten Eigenschaften soll einen Überblick über den Umfang der Berechnungen schaffen.

2.6.5.2 Die Berechnung der Eigenschaften aus den Rohdaten und die Definition der Parametersätze

1. Eigenschaft: Berechnung der äußeren Symmetrie eines Signals bezüglich der Symmetriediagonalen zu einem symmetrischen Signal bei homonuklearen Spektren

Handelt es sich um ein homonukleares zweidimensionales Spektrum, ist es möglich, eine Bewertung der Peaks bezüglich dessen symmetrischen Partner auf der gegenüberliegenden Seite der Symmetrieachse zu treffen. Dies wird durch die Verwendung des sog. Cosinuskriteriums ermöglicht (Schulte et al. 1997). Dieser Score bewertet die Ähnlichkeit der beiden Signalformen und liegt im Bereich von -1 (sehr schlecht) bis 1 (sehr gut). Voraussetzung zur Berechnung sind zwei gleich große Flächen, welche im Folgenden Muster genannt werden. In Abb. 3 ist der Ablauf der Ermittlung des Cosinusscores dargestellt.

Um die ursprüngliche Qualität der Symmetrieeigenschaft weiter zu verbessern, wurde die Symmetriebestimmung im Rahmen dieser Arbeit erweitert.

Um den Symmetrie-Wert (Cosinus-Score) zu bestimmen, werden die Volumen der Signale ausgelesen und alle im Volumen eines Peaks enthaltenen Intensitäten in dieses Muster projiziert. Da jedoch die Volumenform nicht generell exakt um das Zentrum des Musters verteilt ist, wird die Größe der Integrationsbox soweit vergrößert, bis sich die Koordinate des Signales aus der Masterliste im Zentrum des Musters befindet. Alternativ kann die

Originalposition der Integrationsbox belassen werden, jedoch akzeptiert man dadurch, dass der symmetrische Partner zum Referenzmuster verschoben ist. Das eingelesene Muster wird in die Muster-Matrix M_1 gespeichert.

Die Ermittlung des symmetrischen Peaks erfolgt in mehreren Stufen. Zu Beginn wird die Koordinate aus dem Ausgangspeak in die symmetrische Koordinate gespiegelt. An dieser errechneten Position am digitalen Raster wird geprüft, ob an dieser Position ein entsprechender Peak in der Masterliste existiert.

Falls nicht, wird das nächstgelegene Extremum mit gleichem Vorzeichen der gefundenen Koordinate entlang des steilsten Anstiegs bzw. Abstiegs gesucht. Dabei wird der digitalen Auflösung des Spektrums Rechnung getragen. So ist beispielsweise ein zwei Pixel entferntes Extremum entlang der Achse niedrigerer digitaler Auflösung näher, als ein 2 Pixel entferntes Extremum entlang der Achse höherer digitaler Auflösung. Diese Suche wird durch die C++-Klasse *CExtremaHandler* realisiert, welche im Rahmen dieser Arbeit entwickelt wurde und sich wiederum durch den rekursiven Ansatz auf Spektren aller Dimensionen auszeichnet.

Wurde das Extremum innerhalb der festgelegten Schrittweite (Entfernung in Pixel) gefunden, wird dieses als Symmetrie-Partner akzeptiert. Falls nicht, wird die Schrittweite um einen Pixel erhöht (also der Suchbereich erweitert) und erneut gesucht, bis der Partner gefunden wurde. Jedoch nur bis zur maximal festgelegten Schrittweite.

Wurde auf diese Weise ein existierender Peak aus der Masterliste gefunden, werden dessen Abmessungen der Integrationsbox aus der Masterliste verwendet (siehe Abb. 2).

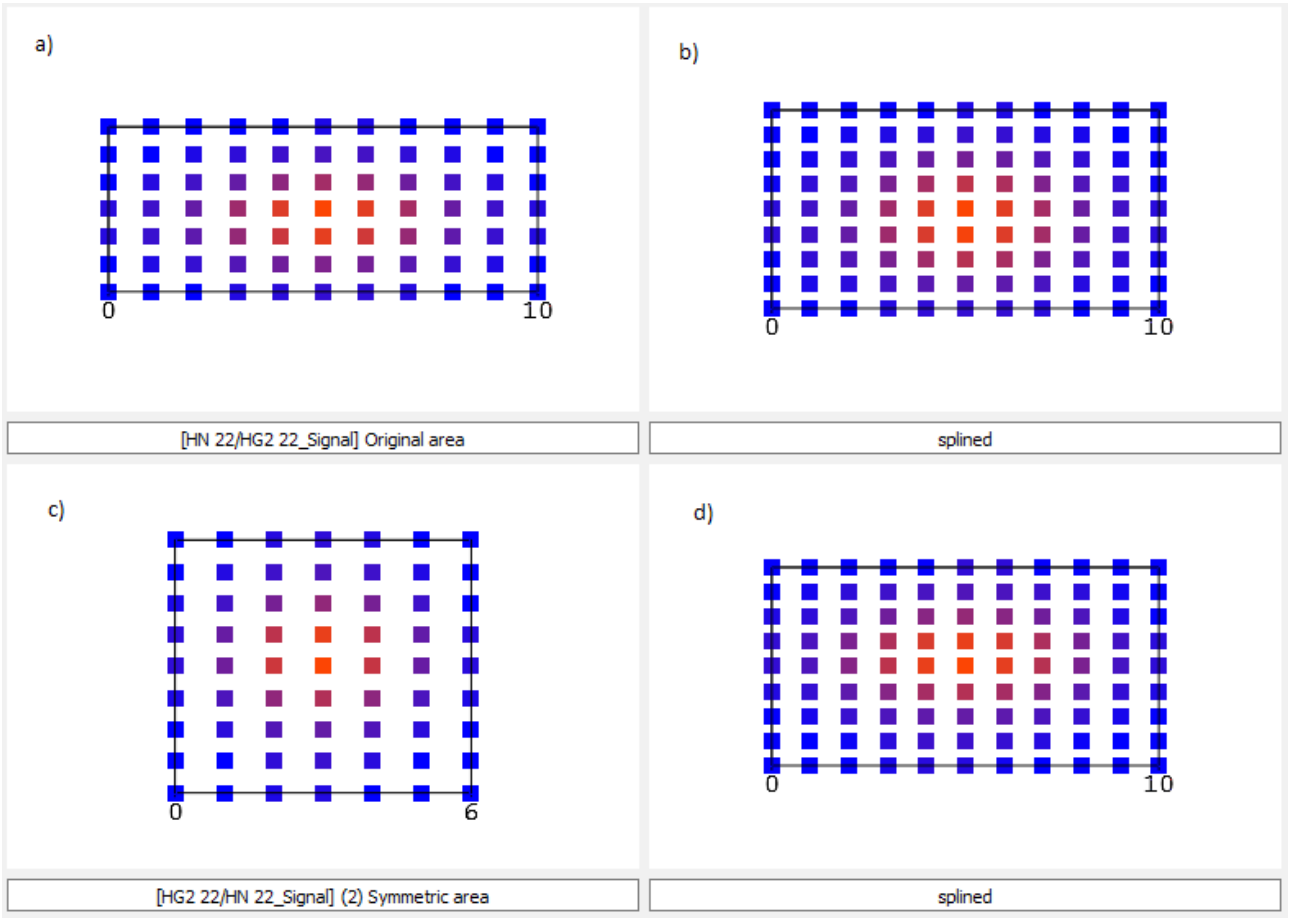

Abb. 2: Screenshot der Visualisierung der äußeren Symmetrie des Signals HN 22/HG2 22; a) zeigt das Referenz-Signal HN 22/HG2 22; c) zeigt das gefundene symmetrische Signal HG2 22/HN 22; b) und d) zeigen die jeweiligen auf gleiche Größe gesplinten Muster, aus welchen der Cosinuswert ermittelt wurde.

Beispiel:

Im Softwarepaket AUREMOL kann ein Peak ausgewählt werden und mit der Taste ‚s' eine Visualisierung der Symmetrie dieses Signals ausgegeben werden. Im folgenden Beispiel wurde das Signal „HN 22/HG2 22" ausgewählt (Abb. 2a). Das Spektrum war ein H^1-H^1-NOESY Spektrum des Proteins *PfTrx* mit einer digitalen Auflösung von 1024x2048 Pixeln. Die Resonanzfrequenz lag bei 800 MHz in beide Frequenzachsen. Die spektrale Breite war in beiden Frequenzachsen 11160,71 Hz (13,95 ppm).

*Tabelle 2: Reduzierter Auszug aus der Masterliste des **ausgewählten Signals**. Die Größe der Volumengrundfläche beträgt hier 1+861-855 = 7 Pixel in der F1-Frequenzachse und 1+368-358=11 Pixel in der F2-Frequenzdomäne. Die Position des Signals liegt im digitalen Raster an 858/363.*

PEAKLABEL: **HN 22/HG2 22**_Signal												
PEAKDESCRIPTION:												
s1	ppm1	s2	ppm2	intensity	volume	prob	width1	width2				
858	0.13347	**363**	9.34639	24713	462940	0.960548	36.083	29.227	**855**	**861**	**358**	**368**

Der gefundene Symmetrie-Partner war das Signal „HG2 22/HN 22“. Jedoch lag dieser nicht an der erwarteten Position des digitalen Rasters bei 1716 (s2) und 182 (s1). Es wurde nach 2 weiteren Such-Schritten in Richtung des Extremums die korrekte Position von s2 bei 1714 gefunden (Abb. 2c)).

*Tabelle 3: Reduzierter Auszug aus der Masterliste des **gefundenen symmetrischen Signals**. Die Größe der Volumengrundfläche beträgt hier 1+185-179 = 7 Pixel in der F1-Frequenzachse und 1+1718-1710=9 Pixel in der F2-Frequenzdomäne. Die Position des Signals befindet sich im digitalen Raster an 182/1714.*

PEAKLABEL: **HG2 22/HN 22**_Signal												
PEAKDESCRIPTION:												
s1	ppm1	s2	ppm2	intensity	volume	prob	width1	width2				
182	9.34174	**1714**	0.14493	25623	390436	0.962318	41.479	22.395	**179**	**185**	**1710**	**1718**

Die jeweiligen Muster b) und d) aus Abb. 2 stellen die durch Splines auf eine gemeinsame Größe gebrachten Muster dar, aus welchen dann der Cosinus-Score berechnet wurde. Dabei wurde die jeweilige höchste Integrationsbox-Ausbreitung verwendet. Das heißt die Größe der Volumengrundfläche aus c) entlang der F1-Achse (symmetrisches Signal) mit 9 Pixeln und aus a) die Ausbreitung des Volumens des Ausgangs-Signals entlang der F2-Frequenzachse mit 11 Pixeln. Damit war die Größe der Muster M_1 und M_2 jeweils 9x11 Pixel.

Kann kein Peak aus der Masterliste gefunden werden, wird die Größe der Integrationsbox des Referenzpeaks benutzt und (basierend auf ppm) in die entsprechende Größe auf dessen Position gegenüber der Symmetrieachse umgerechnet.

Das so ermittelte Muster wird in die Matrix M_2 eingelesen und an der Symmetrie-Diagonalen gespiegelt, damit der symmetrische Peak die gleiche Orientierung der Signalform wie der Ausgangspeak hat.

Um den Informationsverlust durch das Splinen zu minimieren, werden diese beiden Muster auf die größte gemeinsame Größe gebracht, falls das Auflösungsverhältnis nicht größer als 2 oder kleiner als 0.5 ist. Ist dies nicht der Fall, wird die kleinste gemeinsame Größe berechnet. Optional kann jedes Muster vor dem Splinen noch um ein gefundenes Extrema bzw. gefundenen Peak (analog zum Ausgangsmuster M_1) zentriert werden. Dabei wird das Muster derart vergrößert, dass der Peak (in engeren Sinne dessen Extemum) im Mittelpunkt des Musters liegt.

Die Muster M_1 und M_2 können danach noch durch folgende Methoden variiert werden.

- **Beschneidung der Muster durch eine Segmentierungstiefe bezüglich der Intensitäten**

  Falls zu den Mustern auch Peaks aus der Masterliste existieren, werden die Größen der Muster auf die Integrationsbereiche mit der entsprechenden Segmentierung angepasst. Zudem werden alle Intensitäten innerhalb der Muster, welche unter dem vorgegebenen Segmentierungslevel liegen auf 0 gesetzt. Dies betrifft vor allem die Intensitäten in den Eckbereichen der Muster.

- **Modifikation der Muster-Rohdaten (projizierte Intensitäten)**

  Für jeden Symmetrievergleich werden die Intensitäten der Muster wahlweise durch folgende drei Variationen modifiziert:

  Jede Intensität im Muster wird

  - durch den lokalen Rausch an dessen Stelle reduziert
  - durch dessen Signal zu Rausch Verhältnis ersetzt
  - durch den lokalen Rausch an dessen Stelle ersetzt

Diese beiden Muster M_1 und M_2 werden je zu einem Vektor aufgereiht und danach mit dem Cosinuskriterium C bewertet.

Es gilt:

$$C = \frac{\sum_{i=1}^{N} M_{1i} M_{2i}}{\sqrt{\sum_{i=1}^{N} M_{1i} M_{1i} \sum_{i=1}^{N} M_{2i} M_{2i}}} \tag{8}$$

Wobei N die Anzahl der Pixel in Muster M_1 und M_2 darstellen. Bei der Berechnung des Cosinusscores werden nur Werte aus den Mustern mit einbezogen, bei denen sowohl im Referenzmuster M_1 als auch im Symmetriemuster M_2 ungleich 0 (siehe Beschneidung der Muster durch eine Segmentierungstiefe bezüglich der Intensitäten) sind.

Alternativ wird C auch als Mittel der Spalten- und Zeilenkoeffizienten der Pixel repräsentiert durch eine entsprechende Matrix berechnet. Die Annahme war, dass zu jedem Signal der symmetrische Partner die direkte und die indirekte Frequenzachse zum Referenz-Signal vertauscht hat. Da die Auflösungen in die jeweilige Frequenzdomäne stark unterschiedlich sein können, soll die Mittelung der aufgespannten Vektoren je in die beiden Achsen das Cosinuskriterium bei stark unterschiedlichen Auflösungen verbessern.

Seien S die Anzahl der Spalten und Z die Anzahl der Zeilen, dann gilt für den Zeilen- und Spalten-Score C_{ZS}:

$$C_{ZS} = \frac{1}{S+Z} \left(\sum_{z=1}^{Z} \frac{\sum_{s=1}^{S} M_{1sz} M_{2sz}}{\sqrt{\sum_{s=1}^{S} M_{1sz} M_{1sz} \sum_{s=1}^{S} M_{2sz} M_{2sz}}} + \sum_{s=1}^{S} \frac{\sum_{z=1}^{Z} M_{1sz} M_{2sz}}{\sqrt{\sum_{z=1}^{Z} M_{1sz} * M_{1sz} \sum_{z=1}^{Z} M_{2sz} M_{2sz}}} \right) \tag{9}$$

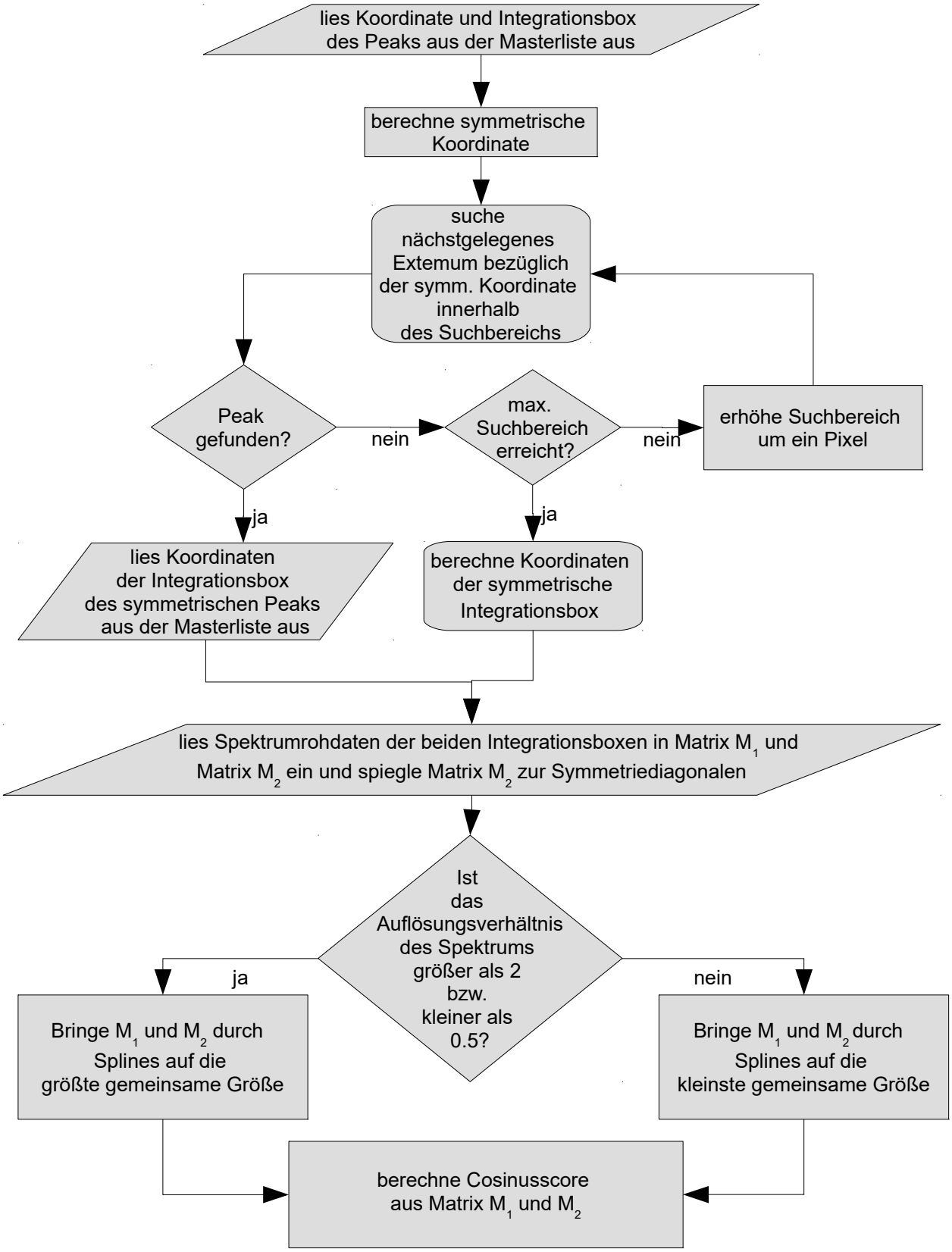

Abb. 3: Ablaufdiagramm der Symmetriesuche zu einem Kreuzsignal bezüglich der Diagonalen im Falle eines homonuklearen Spektrums.

2. Eigenschaft: Die innere Symmetrie eines Signals

Die Bestimmung der Muster der inneren Symmetrie erfolgt analog zur äußeren Symmetrie, nur dass in diesem Fall kein Symmetrie-Partner gesucht wird. Hier wird lediglich das Muster des Referenz-Signals verwendet und um 270° rotiert. Damit wird erreicht, dass der Cosinus-Wert die Güte der symmetrischen Ausdehnung der Signalform in beide Frequenzdomänen wieder gibt.

3. Eigenschaft: Die gaußsche Signalwahrscheinlichkeit basierend auf dem lokalen Rauschen

Da ein Extremum auch ein Störsignal repräsentieren kann, müssen die gepickten Peaks bewertet werden. Das primäre Ziel ist es, Signalpeaks von Stör-Signalpeaks (z. B. *Rauschen* oder *Wasser*) zu unterscheiden um eine korrekte Zuordnung zur entsprechenden Klasse zu erhalten.

Um dies zu bewerkstelligen, wurden im Rahmen dieser Arbeit die Ansätze aus (Trenner 2006) durch die Einführung einer von der Abhängigkeit zur digitalen Auflösung bezüglich der Pixelnachbarschaft um das Extremum erweitert.

Die Verwendung der Normalverteilung zur Berechnung der Wahrscheinlichkeiten

Um die gaußsche Wahrscheinlichkeit zu berechnen, dass ein gepicktes NMR-Signal tatsächlich ein Nutzsignal darstellt, wurde in AUREMOL die Normalverteilung mit Erwartungswert μ und Standardabweichung σ verwendet:

$$f(x)=\frac{1}{\sigma\sqrt{(2\pi)}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{10}$$

Mit μ=0 in (10) folgt:

$$f(x)=\frac{1}{\sigma\sqrt{(2\pi)}}\exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{11}$$

Dabei stellt x die Mittelwerte der Nachbarschaftsintensitäten in einem Raster um die Signalposition herum dar und die Standardabweichung σ entspricht der minimalen Standardabweichung an der Position des Signals.

Daher geht Formel (11) über in:

$$f(x)=\frac{1}{\sigma\sqrt{(2\pi)}}\exp\left(\frac{-\left(\frac{1}{N}\sum_{i=1}^{N} I_i\right)^2}{2\sigma_m^2}\right) \tag{12}$$

Wobei N die Gesamtzahl der Nachbarschaftspixel I_i eines Signals innerhalb eines Bereichs am digitalen Raster darstellt.

Durch die Standardisierungsformel (mit μ=0)

$$z=\frac{x}{\sigma}=\frac{\frac{1}{N}\sum_{i=1}^{N} I_i}{\sigma_m} \tag{13}$$

kann aus Formel (12) die Dichtefunktion der Standardnormalverteilung erhalten werden, da durch diese Transformation die Kurve so verändert wird, dass sie der Standardnormalverteilung mit μ=0 und σ=1 entspricht:

$$\phi(z)=\frac{1}{\sqrt{(2\pi)}}\exp\left(\frac{-z^2}{2}\right) \tag{14}$$

Durch diese Transformation kann die Verteilungsfunktion der Standardnormalverteilung zur Berechnung der Wahrscheinlichkeiten P_r benutzt werden, um zu bewerten, ob es sich bei einem gepickten NMR-Signal um ein Rauschsignal handelt.

$$P_r(Z\le z)=\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{z}\exp\left(\frac{-t^2}{2}\right)dt \tag{15}$$

Da die Bewertung der Signalpeaks von Interesse ist, gilt für die Wahrscheinlichkeit P_s, dass ein gepicktes NMR-Signal ein Nutzsignal ist:

$$P_s(X\ge x)=1-P_r(X\le x) \tag{16}$$

Da bei vielen Spektren der Effekt auftritt, dass übermäßig viele Wahrscheinlichkeiten über 99 % liegen, wurde anstatt der kumulativen Verteilungsfunktion (Abb. 4) die logarithmisch kumulative Verteilungsfunktion (Abb. 6) verwendet.

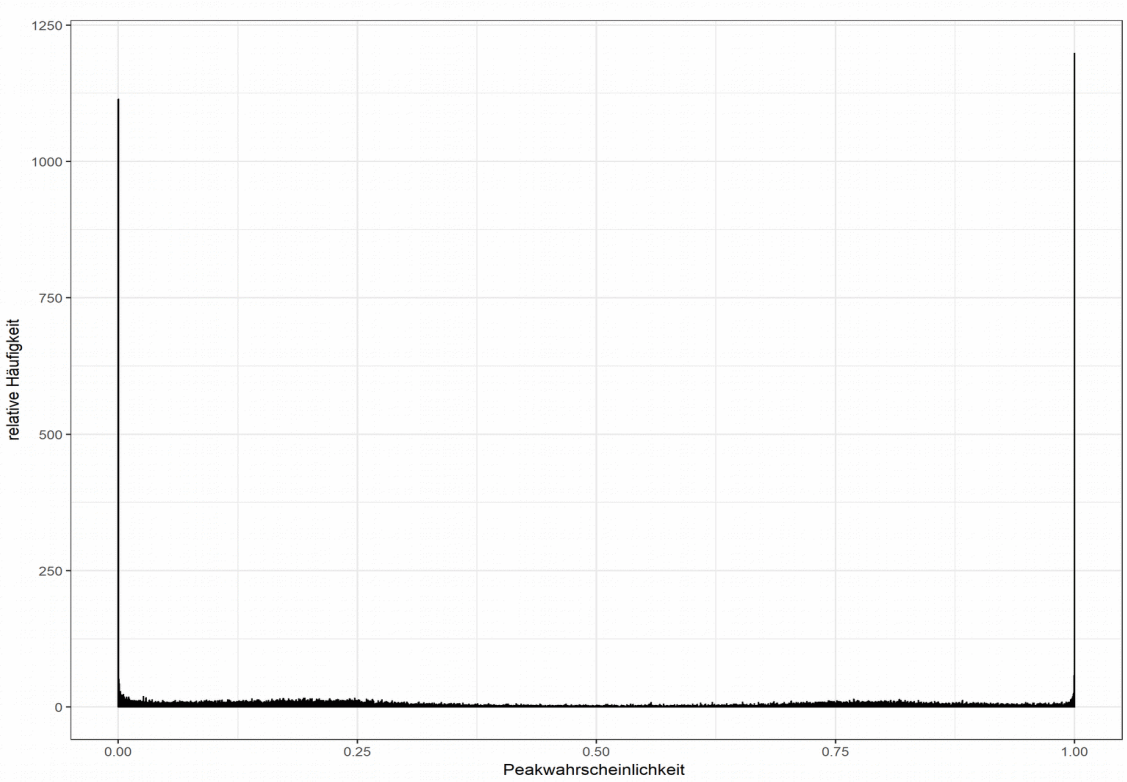

Abb. 4: Wahrscheinlichkeitsverteilung der normalen kumulativen Dichtefunktion der gaußschen Wahrscheinlichkeit aller Peaks.

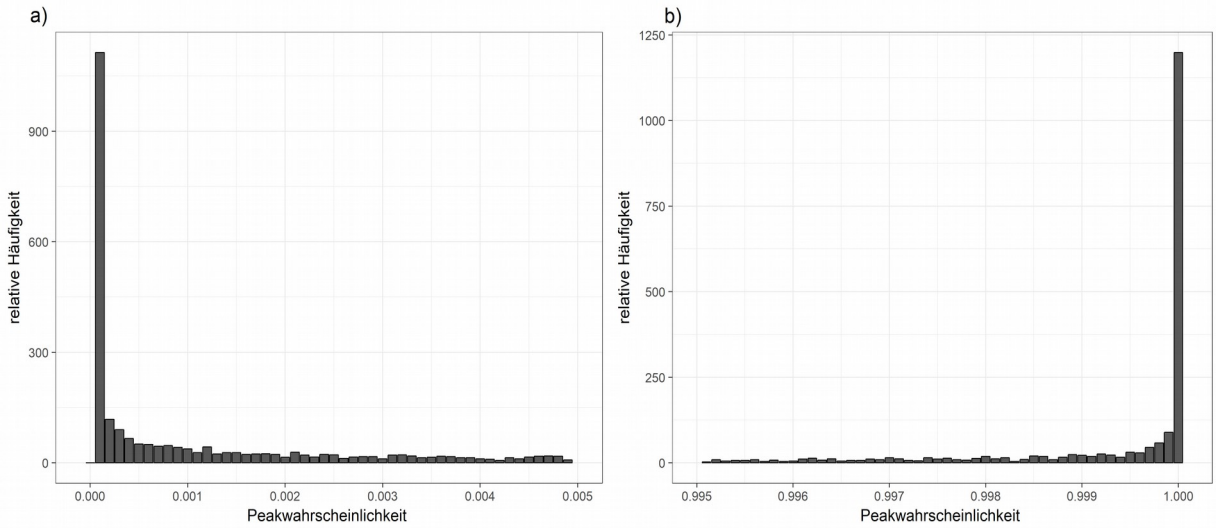

Abb. 5: Vergrößerung der Bereiche sehr kleiner Wahrscheinlichkeiten (a) nahe bei 0 und sehr großer Wahrscheinlichkeiten (b) nahe bei 1 aus Abb. 4.

Dazu muss eine veränderte Transformationsformel z_l verwendet werden und Formel (13) geht über in:

$$z_l = \frac{\ln\left(\left|\frac{1}{N}\sum_{i=1}^{N} I_i\right|\right)}{\sigma_m} \tag{17}$$

Zur Berechnung der Wahrscheinlichkeiten wurde in AUREMOL die Funktion gsl_cdf_ugaussian_P aus der freien GNU-Scientific-Library verwendet. Dabei stellt der Übergabeparameter den Wert aus der Transformationsformel (13) oder (17) dar.

Die Verteilung der Wahrscheinlichkeiten in Abb. 6 ist stark entzerrt und weist einen aussagekräftigeren Verlauf als in Abb. 4 auf.

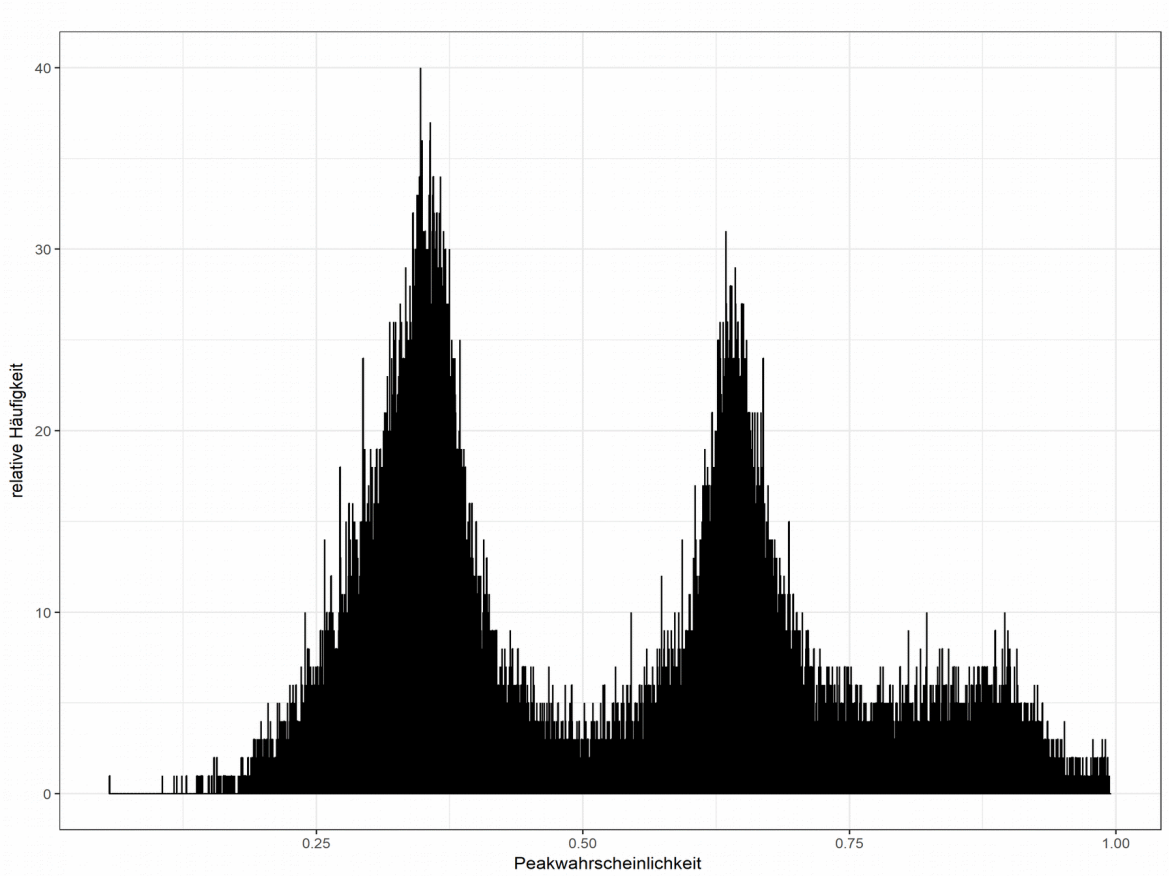

Abb. 6: Entzerrte Wahrscheinlichkeitsverteilung der logarithmisch kumulativen Dichtefunktion der gaußschen Wahrscheinlichkeit.

Die Variationen der Tranformationsfomeln

Die Größe des ursprünglichen festen Gaußfilters der Formeln (13) und (17) (Intensitäten I_i im Bereich von 3x3 Pixel bei zwei Dimensionen) wurde durch die digitale Auflösung in jeder Frequenzdomäne angepasst. Dadurch ergab sich z. B. für die Transformationsformel (13) bei Verwendung eines festen 3x3 großen Gaussfilter:

$$z = \frac{\frac{1}{9}\sum_{i=1}^{9} I_i}{\sigma_m} \tag{18}$$

Bei einer digitalen Auflösung eines zweidimensionalen Spektrums von 1024x2048 Pixeln wird die Größe des verwendeten Bereiches für die Intensitäten I_i wie folgt dynamisch erweitert:

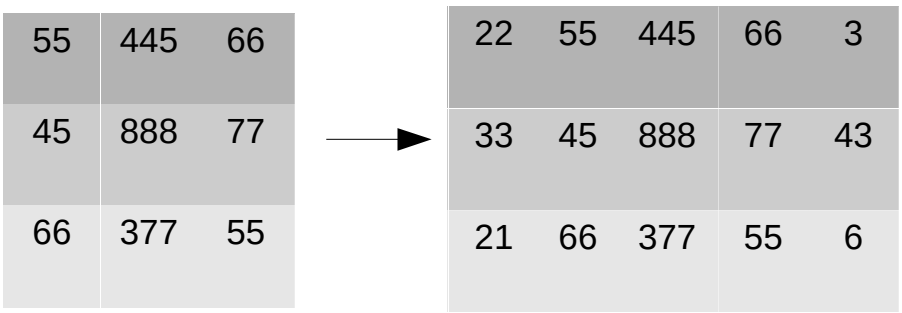

Abb. 7: Dynamische Erweiterung des Gaußfilters bezüglich der digitalen Auflösung; Gaußfilter mit fester Größe von 3x3 (links); dynamische Größe basierend auf der digitalen Auflösung von 1024x2048 Pixeln des Spektrums (rechts).

Der Effekt ist, dass das Intensitätsmittel (also der Mittelwert der einbezogenen Intensitäten aus dem Gausskernel) zum lokalen Rausch-Verhältnis vergrößert wird, wenn es sich um einen Signalpeak handelt, da die Vergrößerung des Filters nun zusätzlich niedrigere Werte ins Mittel einbringt. Im Falle von Rausch würde tendenziell das Mittel erhöht werden, da ein Bereich des erweiterten Filters bereits den nächsten Rauschpeak erfasst. Somit wird der Bereich für die Intensitäten I_i aus den Transformationsformeln (13) und (17) zusätzlich abhängig von der digitalen Auflösung des Spektrums.

Eine weitere Möglichkeit ist, dass der Bereich durch das Intensitätsmittel aus den Einzelintensitäten I_i der Transformationsformeln (13) und (17) der realen Fläche bestimmt

wird, welches das Peak-Volumen aufspannt. Dadurch werden auch Peaks mit einer sehr kleinen Volumenausbreitung (im Extremfall, falls das Volumen lediglich durch die Intensität eines einzelnen Pixels bestimmt wird) korrekt in die Berechnung eingebracht. Die hierfür benötigte Information wird bei der Integration der Volumen der Signale ermittelt und in die Struktur *Integrations-Hash* abgespeichert.

Zur Berechnung des lokalen und globalen Rauschens wurde der iterative Ansatz abermals durch einen rekursiven mit all seinen bereits beschriebenen Vorteilen abgelöst. Dies ist in der Klasse *CNoiseLevelCalculator* realisiert, welche auch die Berechnung auf mehreren Prozessorkernen beherrscht.

Im Rahmen der Neuimplementierung wurde auch die Berechnung der minimalen Standardabweichung σ_m aus den Transformationsformeln (13) und (17) so verändert, dass statt der festen Größe von 10 Pixel des „sliding windows“ wieder analog zu (Trenner 2006) auf eine Größenabhängigkeit von der digitalen Auflösung übergegangen wurde.

Es gilt:

$$S = \frac{N}{100} \tag{19}$$

Wobei S die Größe des „sliding windows“ und N die Gesamtanzahl der Pixel entspricht. Jedoch darf die Pixelweite von 10 Pixel nicht unterschritten werden.

Im Detail erfolgt die Berechnung nun derart, dass zu jedem Pixel in jede Dimension je eine Säule in eine andere Dimension die Datenbasis für das „sliding windows“ gewählt wird. Bei einem eindimensionalen Spektrum bedeutet dies, dass es kein „sliding windows“ geben kann, da nur ein Pixel in jede (hier genau eine) Dimension existiert. Daher wird dieser Wert zur Berechnung der Varianz (und damit der Standardabweichung) direkt benutzt. Ist das Spektrum zweidimensional, werden Vektoren zu jeder Frequenzdomäne gebildet. Die kleinsten Standardabweichungen aller „sliding windows“ werden nacheinander entlang jeder Frequenzachse gebildet und in diese Vektoren gespeichert.

Dies geschieht analog für alle Dimensionen, wobei die Speicherung der minimalen Standardabweichungen jedes Vektors in Strukturen einer Dimension tiefer erfolgen. Sprich ein zweidimensionales Spektrum hat zwei Vektoren gefüllt mit den minimalen

Standardabweichungen, ein dreidimensionales Spektrum hat analog drei Matrizen gefüllt mit den minimalen Standardabweichungen usw.

Durch die Sicherstellung von einem von 0 abweichenden minimalen Mittelwert wird verhindert, dass eine minimale Standardabweichung den Wert 0 erhält und damit zu einer undefinierten Transformationsformel (13) oder (17) führt. Dazu wird bei der Bildung der Standardabweichungen jeder Mittelwert von 0 durch ein EPSILON ersetzt, welches die Eingenschaft hat, dass die Addition von 1 zu diesem Wert die nächste numerisch darstellbare Zahl ist.

Das lokale Rauschen kann nun für jede Intensität des digitalen Rasters lt. (Trenner 2006) ermittelt werden.

4. Eigenschaft: Volumenfehler

Die Berechnung des Volumenfehlers wurde unverändert aus (Trenner 2006) übernommen und setzt sich aus dem Fehler der Signalüberlappung und der Schätzung mittels Konfidenzintervall zusammen.

5. Eigenschaft: Kreuz-Rauschen eines Peaks

Bei dieser Eigenschaft wurde als Ausgangsposition die Signal-Position am digitalen Raster gewählt und danach für jede Dimension eine Strecke gebildet, welche durch die Größe des resultierenden Integrationsbereichs durch die Integration bestimmt wird. Dies entspricht annähernd der Linienbreite der Volumenform an einer bestimmten Segmentierungstiefe. So würde im Falle eines zweidimensionalen Spektrums ein Kreuz am Peak aufgespannt, welche die Basis für die Berechnung des Keuz-Rauschens bildet. Für alle auf diesen Linien liegenden Positionen des Rasters wurde das lokale Rauschen berechnet und aufsummiert.

6. Eigenschaft: Die Schwerpunkt-Abweichung von der in der Peakliste festgelegten Position des Signals am digitalen Raster

Zur Bestimmung des Schwerpunktabstandes wird die Information aus dem Volumen dazu verwendet, den Schwerpunkt eines Peaks zu berechnen (siehe Kapitel Schwerpunkt eines Signals). Hier wurde der Abstand der Position des Massenschwerpunkts des Peaks zu seiner Position in der Masterliste in ppm berechnet und als Eigenschaft erfasst. Die Schwerpunktbestimmung wurde mit verschiedenen Segmentierungstiefen bestimmt.

7. Eigenschaft: Die Intensität des Signals

Die Signalintensität entspricht der Intensität der Position des Signals auf dem digitalen Raster aus der Masterliste. Die Festlegung dieser Position (also die Aufnahme des Signals in die Masterliste) kann sowohl manuell als auch durch automatisches Peak-Picken oder durch Bestimmung des Massenschwerpunkts erfolgen.

8. Eigenschaft: Das Verhältnis der Peakintensität zum lokalen Rauschen

Analog zur Peakintensität wird die Intensität eines Signals aus der Masterliste ausgelesen und durch den lokalen Rausch an der Signal-Position dividiert.

9. Eigenschaft: Das Volumen eines Signals

Die Volumen der Signale werden direkt aus dem *Integrations-Hash* ausgelesen, wobei die Volumen auch hier für verschiedene Segmentierungstiefen abgefragt werden können. Die entsprechenden Details sind im Kapitel verbesserte Integration aufgeführt.

10. Eigenschaft: Die Volumengrundfläche eines Signals

Für diese Eigenschaft wurde lediglich der Anteil der digitalen Pixel aufsummiert, welche in das Volumen eines Signals eingeflossen sind. Auch diese Eigenschaft wurde durch die Angabe verschiedener Segmentierungstiefen variiert, da sich durch eine veränderte Segmentierung die Ausbreitung der Volumenform vergrößern (niedrige Segmentierung oder verkleinern (höhere Segmentierung) kann.

11. Eigenschaft: Das Verhältnis der Peakintensität zur Volumengrundfläche

Diese Eigenschaft stellt das Verhältnis der Intensität des Signals zu dessen Volumengrundfläche dar. Auch hier wurden analog zu obigen Eigenschaften verschiedene Segmentierungsschwellen getestet.

12. Eigenschaft(en): Die Linienbreite(n) eines Signals

Die Berechnung der Linienbreite wurde unverändert verwendet. Die Linienbreiten lagen in der Materliste bezüglich jeder Frequenzachse vor. Die Linienbreite wurde mittels der Levenberg-Marquardt-Optimierung von Lorentz- oder Gaußfunktionen berechnet und danach in der Masterliste gespeichert. Jede Linienbreite pro Frequenzachse stellt eine eigenständige Eigenschaft dar. Ein eindimensionales Spektrum hat somit eine einzige

Eigenschaft der Linienbreite und ein zweidimensionales Spektrum liefert zwei Eigenschaften der Linienbreite.

13. Eigenschaft: Die Summe der Linienbreiten eines Signals

Für diese Eigenschaft wurde die Summe aus allen Linienbreiten eines Signals gebildet.

14. Eigenschaft(en): Das Verhältnis der Peakintensität zur Linienbreite

Hier ist das Verhältnis von der Signalintensität zur Linienbreite auf halber Höhe dargestellt. Auch hier wurde die Linienbreite separat in jede Frequenzachse wie zuvor bei der Linienbreite als Eigenschaft definiert.

15. Eigenschaft: Die Summe der Verhältnisse der Peakintensität zu Linienbreiten

Analog zur 13.ten Eigenschaft wurde hier wieder die Summe aus den Eigenschaften 14 gebildet.

16. Eigenschaft: Die volumenbasierte Linienbreite

In diesem Fall wurde die Breite der durch das Volumen aufgespannten Signalform in Pixel von jeder Frequenzachse aus dem *Integrations-Hash* bestimmt. Auch diese Eigenschaft wurde durch die Angabe verschiedener Segmentierungstiefen variiert, da sich durch eine veränderte Segmentierung die Ausbreitung der Volumenshapes vergrößern (niedrige Segmentierung) oder verkleinern (höhere Segmentierung) kann und somit die Linienbreite verändert. Damit würde sich bei einer Segmentierungstiefe von 0,5 die Linienbreite auf halber Höhe ergeben. Die Genauigkeit dieser Berechnung der Linienbreite ist in hohem Maße von der Auflösung des digitalen Rasters abhängig.

17. Eigenschaft: Das Verhältnis der Peakintensität zur volumenbasierten Linienbreite

Analog zur Eigenschaft 14 wurde hier das Verhältnis der Peakintensität zur volumenbasierten Linienbreite gebildet. Die Anzahl dieser Eigenschaften ist wieder abhängig von der Dimension des Spektrums.

18. Eigenschaft: Die Summe aus dem Verhältnis der Peakintensität zur volumenbasierten Linienbreite

Analog zur 13.ten Eigenschaft wurde hier wieder die Summe aus den Eigenschaften 18 gebildet.

2.6.5.3 Die Vorbereitung der Datenbasis aus dem ^{1}H-^{1}H-NOESY-Spektrum von *PfTrx*

Da die ursprüngliche Version der Bayesschen Wahrscheinlichkeitsberechnung nur positive Intensitäten erlaubte und auch die Anzahl der Peaks bereits vorab durch das adaptive Pickverfahren (Trenner 2006) zu stark reduziert wurde, waren die für eine Diskriminierung erforderliche Anzahl von Signalen und Störsignalen nicht mehr ausreichend gegeben.

Um die Eigenschaften 1-18 für die Diskriminierung bestimmen zu können, wurden zu den bereits zugeordneten 6738 Signalen alle noch nicht gepickten Extrema hinzugefügt, so dass letztendlich ein vollständig gepicktes Spektrum mit 50436 Peaks vorlag. Dabei wurden alle Extrema unter der Intensitätsschwelle von 434 ignoriert. Diese Schwelle wurde durch das globale Rauschlevel bestimmt (Trenner 2006).

Da es zu testen galt, welche Eigenschaften am besten klassifizieren, wurden die Klassen *Signal*, *Rauschen* und *Wasser* interaktiv markiert. Jedoch gehen in den markierten Bereich *Signal* nur die Peaks ein, die zuvor bereits richtig zugeordnet waren (also keine Störsignale), so dass die Abschätzung durch Reduktion der Signalklasse außer Acht gelassen werden konnte.

Danach wurden sowohl die Linienbreiten als auch die Volumen bei einer Segmentierungtiefe von 0,001 berechnet. Alle volumenbasierten Eigenschaften konnten durch die Struktur *Integrations-Hash* ohne wiederholte Integration bestimmt werden.

2.6.5.4 Berechnung der Rohdatensätze

Die Berechnung der Daten (also die Peakeigenschaften) wurde in mehreren Schritten durchgeführt, um den Speicherbedarf des Arbeitsspeichers zu minimieren und die Berechnung auf mehrere CPU-Prozesse auslagern zu können.

Bei der Berechnung wurden vier Eigenschaftstypen unterschieden, die bezüglich ihrer Berechnungsmethode zusammengefasst werden konnten.

Für alle Eigenschaften wurden zusätzlich verschiedene Skalierungen getestet, um eine effektivere Wahrscheinlichkeitsdichteverteilung zu erhalten. Dazu wurden zusätzlich folgende Skalierungen verwendet:

- logarithmische Skalierung, falls die Maximalwerte größer 3 oder kleiner als -3 sind

- Verschiebung der Daten ins Positive, bis der kleinste Wert am 0-Punkt der x-Achse liegt
- logarithmische Skalierung und danach Verschiebung ins Positive

Wichtig dabei war, dass die gesamten Daten skaliert wurden und nicht jede Klasse einzeln, da im Falle der Verschiebung die Information der Daten zerstört worden wäre.

Zum Abschluss wurden die Daten, falls negative Werte vorhanden waren, zusätzlich als Absolutwerte genommen. So war je ein Datensatz sowohl mit positiven und negativen Werten und ein Datensatz mit ausschließlich positiven Werten vorhanden.

Folgende Berechnungs-Variationen wurden bei den verschiedenen Eigenschaftstypen vorgenommen:

Die volumenbasierten Eigenschaften

1. Volumen mit den Segmentierungstiefen bei 0,0001, 0,1, 0,2 und 0,5
2. keine veränderte Skalierung, logarithmische Skalierung, Verschiebung ins Positive und Verschiebung ins Positive mit anschließender Logarithmierung
3. Werte aus Datensatz als Absolutwerte genommen oder unangetastet

Die nicht volumenbasierten Eigenschaften

1. keine veränderte Skalierung, logarithmische Skalierung, Verschiebung ins Positive und Verschiebung ins Positive mit anschließender Logarithmierung
2. Werte aus Datensatz als Absolutwerte genommen oder unangetastet

Eigenschaften abhängig von der gaußschen Signalwahrscheinlichkeit (im Falle des Volumenflächen-Filters)

1. Volumen mit den Segmentierungstiefen bei 0,0001, 0,1, 0,2 und 0,5
2. fester Gaußfilter von 3x3 Pixel, dynamischer Gaußfilter in Abhängigkeit der digitalen Auflösung des Spektrums und Volumengrundfläche als Gaußfilter zur Berechnung der gaußschen Signalwahrscheinlichkeit (siehe Eigenschaft 3 aus 2.6.5.2)

Symmetrie-Eigenschaften

1. Volumen mit den Segmentierungsschwellen bei 0,0001, 0,1, 0,2 und 0,5

2. keine veränderte Skalierung, Verschiebung ins Positive
3. Werte aus Datensatz als Absolutwerte genommen oder unangetastet
4. Manipulation der Muster-Daten:
    a) Rohdaten unverändert
    b) Subtraktion der durchschnittlichen Intensität des ganzen Musters von den einzelnen Muster-Intensitäten
    c) Subtraktion des lokales Rauschens eines jeden Musterelements von jeder einzelnen Intensität
    d) Austausch der Muster-Intensitäten durch die Verhältnisse der Intensität zum lokalen Rauschen eines jeden Pixels
5. erlaubte Suchweite des diagonalen Peaks im Falle der äußeren Symmetrie von 0 bis 3
6. Splines: splinen auf größte gemeinsame Fläche oder auf kleinste gemeinsame Fläche
7. Cosinusberechnungstyp: Spalten-Score C_{ZS} (siehe Formel 9) oder Score C des aufgereihten Vektors (ursprüngliche Berechnung aus Formel 8)
8. Musterbereich: nachträglich zentriert um den Peak oder aus der Größe der Integrationsbox
9. minimal erlaubte Muster-Größe von 3 bis 5 Pixeln in jeder Frequenzdomäne
10. Abtrennung entlang der Volumengrundfläche bei den Segmentierungsschwellen 0,0001, 0,1, 0,2 und 0,5 (die Muster sind in diesem Fall also nicht mehr rechteckig).

Nach der Festlegung der optimalen Parameter war noch von Interesse, welche Eigenschaften überhaupt einen positiven Beitrag zum Endergebnis liefern. Dazu wurden die zuvor evaluierten Eigenschaftsparameter verwendet um alle möglichen Kombinationen der Eigenschaften zu testen. Im Gegensatz zu den Tests zur Erlangung der optimalen Parameter wurde jedoch hier der gesamte Datensatz zur Wiedererkennung der Klassenzugehörigkeit der Peaks verwendet. D. h. es wurden anhand der markierten Klassen die Verteilungen zu jeder Klasse generiert und anschließend auf alle Peaks

angewendet. Danach wurde untersucht, wie oft ein Signal oder ein Störsignal korrekt wiedererkannt wurde.

Die beste Kombination kann somit als Ergebnis des erweiterten Algorithmus angesehen werden und kann mit der ursprünglichen Version verglichen werden.

Im Folgenden wurden drei Tests bezüglich der Klassenanzahl (also die markierten Bereiche) auf das experimentelle Spektrum durchgeführt.

2.6.5.5 Die Bereiche zur Festlegung der Klassen *Signal* und *Rauschen*

Tabelle 4: Bereich der Klasse Signal

Frequenzachse w1	Frequenzachse w2
4,0 bis 0,0 ppm	10,0 bis 6.7 ppm

Tabelle 5: Bereiche der Klasse Rauschen vereint mit der Klasse Wasser

Frequenzachse w1	Frequenzachse w2
-0,7 bis -2,1 ppm	11,5 bis 5,5 ppm
-0,7 bis -2,1 ppm	5,49 bis 4,01 ppm
2,2 bis 0,6 ppm	-0,9 bis -1,6 ppm

Hier wurden die eigentlichen Klassen *Wasser* und *Rauschen* zu einer Rauschen-Klasse vereint, so dass für die Diskriminierung nur zwei Klassen vorlagen.

2.6.5.6 Die Bereiche zur Festlegung der Klassen *Signal*, *Rauschen* und *Wasser*

Tabelle 6: Bereich der Klasse Signal

Frequenzachse w1	Frequenzachse w2
4,0 bis 0,0 ppm	10,0 bis 6.7 ppm

Tabelle 7: Bereiche der Klasse Rauschen

Frequenzachse w1	Frequenzachse w2
-0,7 bis -2,1 ppm	11,5 bis 5,5 ppm
2,2 bis 0,6 ppm	-0,9 bis -1,6 ppm

Tabelle 8: Bereiche der Klasse Wasser

Frequenzachse w1	Frequenzachse w2
-0,7 bis -2,1 ppm	5,49 bis 4,01 ppm

Hier galt es somit mittels drei Klassen zu diskriminieren.

2.6.5.7 Die Generierung der Verteilungen der verschiedenen Klassen

Da die Rohdaten nun alle vorlagen und die Klassen definiert wurden, konnten die Wahrscheinlichkeitsdichteverteilungen einer jeden Klasse und Eigenschaft generiert werden.

Hierzu wurden verschiedene Glättungsgrößen zur Erstellung der Wahrscheinlichkeitsdichtefunktionen getestet:

- adaptive Variation des Glättungsfaktors, bis die gewünschte Anzahl an Extrema (4, 3, 2 oder 1) in der Wahrscheinlichkeitsdichteverteilung vorzufinden waren.
- fester Glättungsfaktor von 1, 2, 3, 4, 5 und 10 Pixeln

Somit wurden z. B. für eine Eigenschaft und einer Klasse 10 Wahrscheinlichkeitsdichteverteilungen verschiedener Glättungsfaktoren generiert, die jeweils für die Berechnung der Bayesschen Wahrscheinlichkeit als Basis dienten. So konnte evaluiert werden, welche Glättungsmethode für die Diskriminierung die beste Verteilung lieferte. Diese Verteilungen wurden dann als Binärdateien gespeichert (serialisiert), so dass diese zur Berechnung der endgültigen Hitliste herangezogen werden konnte ohne erneut erstellt werden zu müssen.

2.6.5.8 Generierung der Hitlisten durch Berechnung verschiedener Parametersätze zur Ermittlung der Bayesschen Wahrscheinlichkeiten der Signale

Die vorliegenden Verteilungen konnten verwendet werden, um die Bayessche Wahrscheinlichkeit eines jeden NMR-Signals einer jeden Klasse (hier nur die NMR-Signale der markierten Klassen) aufgrund einer einzigen Eigenschaft zu berechnen.

Sind die Ereignisse voneinander unabhängig, gilt für die Bayessche Wahrscheinlichkeit, dass ein NMR-Signal zur Klasse *Signal* K_s gehört:

$$p(K_s|E_1,\ldots,E_z)=\frac{\prod_{i=1}^{N} P(E_i|K_s)}{\sum_{j=1}^{M}\prod_{i=1}^{N} P(E_i|K_j)} \quad (20)$$

Wobei M die Anzahl der Klassen und N die Anzahl der verwendeten Eigenschaften E_i darstellt.

Da jede Eigenschaft eine andere Wahrscheinlichkeitsverteilung liefert, muss für jede Eigenschaft ein Wahrscheinlichkeits-Threshold bestimmt werden, der festlegt, welche Wahrscheinlichkeit die *Signal*-Klasse von den anderen Klassen am besten trennt.

Dazu wurde die Hitliste so erstellt, dass diese für jeden Parametersatz den Wahrscheinlichkeits-Threshold der Peak-Wahrscheinlichkeit insofern bestimmt, dass die maximale Anzahl an richtig positiven Peaks aller Klassen erreicht wird bei gleichzeitiger maximaler Anzahl an richtig negativen Peaks.

Die Hitliste wird dann absteigend nach der Anzahl der richtig positiven Peaks sortiert. Diese Hitliste wurde für jede Eigenschaft erstellt und die zugehörigen Datensätze archiviert, um den Datensatz mit den optimalen Parametern für die spätere Auswertung wieder auszulesen.

2.6.5.9 Beste Kombination der Eigenschafts-Verteilungen zur Berechnung der Bayesschen Wahrscheinlichkeit

Um ein optimales Ergebnis der Bayesschen Wahrscheinlichkeit zu erreichen, wurde zum Abschluss noch bei allen vorhandenen NMR-Signalen aus der Masterliste getestet, welche Kombination der Wahrscheinlichkeitsdichteverteilungen der Eigenschaften (also die Verteilung, welche mit den optimalen Parametern bestimmt wurde) das beste Resultat lieferte.

2.7 Theoretische Verteilungen zur Berechnung der Bayesschen Wahrscheinlichkeit

Da alle Eigenschaften als Verteilungen (siehe vorangegangenes Kapitel) vorlagen und deren optimale Berechnungs-Parameter bekannt sind, wurden mit diesen Parametern für jede Eigenschaft die Daten generiert, welche auf der Basis eines korrekt zugeordneten, experimentellen oder simulierten Spektrums erstellt wurden. Ziel war es, eine Funktion zu finden, welche die Verteilungen der Wahrscheinlichkeiten einer jeden Eigenschaft am besten beschreibt. Zudem sollten diese Verteilungen ermöglichen, auch auf sehr wenig Signale anwendbar zu sein (beschränkte Statistik), welche keine Generierung von Verteilungen durch ihren geringen Umfang zulassen.

2.7.1 Die Varianten der möglichen Verteilungen

2.7.1.1 Die Normalverteilung

Es wird angenommen, dass eine zufällig ausgewählte Eigenschaft eines Peaks X der Normalverteilung mit dem Mittelwert μ und der Standardabweichung σ unterliegt (LIMPERT et al. 2001), falls diese die Wahrscheinlichkeitsdichte f_N besitzt.

$$f_N(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{21}$$

2.7.1.2 Die logarithmische Normalverteilung

Weiter wird angenommen, dass eine zufällig ausgewählte Eigenschaft eines Peaks X der logarithmischen Normalverteilung mit den Parametern μ und σ unterliegt (LIMPERT et al. 2001), falls diese die Wahrscheinlichkeitsdichte f_{LOGN} besitzt.

$$f_{LOGN}(x)=\begin{cases}\frac{1}{\sigma\sqrt{2\pi}\,x}e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}, \text{falls } x>0\\ 0, \text{falls } x\le 0\end{cases} \tag{22}$$

2.7.1.3 Die kombinierte Wahrscheinlichkeitsdichte

Durch eine Kombination zweier Wahrscheinlichkeitsdichten (Mischverteilung) ergibt sich dann:

$$f_k(x)=\lambda f_1(x)+(1-\lambda)f_2(x) \tag{23}$$

Wobei λ das Gewicht repräsentiert, wie stark der Beitrag der jeweiligen Funktion f1 und f2 zur Gesamtfunktion f_k ist.

Die Funktionen f1 und f2 können sowohl (21) als auch (22) entsprechen. Somit ergeben sich folgende mögliche Wahrscheinlichkeitsdichten:

N-Verteilung

$$f_N(x)=\lambda\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{24}$$

LOGN-Verteilung

$$f_{LOGN}(x)=\begin{cases}\lambda\frac{1}{\sigma\sqrt{2\pi}x}e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}, \text{falls } x>0\\ 0, \text{falls } x\leq 0\end{cases} \tag{25}$$

NN-Verteilung

$$f_{NN}(x)=\lambda\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}+(1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \tag{26}$$

NLOGN-Verteilung

$$f_{NLOGN}(x)=\begin{cases}\lambda\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}+(1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi}x}e^{-\frac{(\log(x)-\mu_2)^2}{2\sigma_2^2}}, \text{falls } x>0\\ \lambda\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \text{falls } x\leq 0\end{cases} \tag{27}$$

LOGNN-Verteilung

$$f_{LOGNN}(x)=\begin{cases}\lambda\frac{1}{\sigma_1\sqrt{2\pi x}}e^{-\frac{(\log(x)-\mu_1)^2}{2\sigma_1^2}}+(1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}, \text{falls } x>0\\ (1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}, \text{falls } x\le 0\end{cases} \quad (28)$$

LOGNLOGN-Verteilung

$$f_{LOGNLOGN}(x)=\begin{cases}\lambda\frac{1}{\sigma_1\sqrt{2\pi x}}e^{-\frac{(\log(x)-\mu_1)^2}{2\sigma_1^2}}+(1-\lambda)\frac{1}{\sigma_2\sqrt{2\pi x}}e^{-\frac{(\log(x)-\mu_2)^2}{2\sigma_2^2}}, \text{falls } x>0\\ 0, \text{falls } x\le 0\end{cases} \quad (29)$$

2.7.1.4 Die Maximum-Likelihood-Funktionen

Um die Parameter der Verteilungsfunktionen der Formeln (24)-(29) zu erhalten, wurde die Maximum-Likelihood-Schätzung (R.A. Fisher and the making of maximum likelihood 1912-1922 1997) für die freien Parameter Mittelwert μ_1, Mittelwert μ_2, Standardabweichung σ_1, Standardabweichung σ_2 und das Gewicht λ verwendet, indem man die Maximum-Likelihood-Funktion maximiert. Dieses Verfahren wurde der Momentenmethode (Hazewinkel 2002) vorgezogen, da der Stichprobenumfang sehr hoch werden kann und dessen einfache Berechnung durch die numerische Iteration sehr rechenintensiv wird. Zudem ist der erlangte Schätzer nicht erwartungstreu (Koutrouvelis und Meintanis 2002).

Somit gilt mit (23) für die zu maximierende Likelihood-Funktion L_k:

$$L_k(\mu_1,\mu_2,\sigma_1,\sigma_2,\lambda)=\prod_{i=1}^{N} f_k(x_i) \quad (30)$$

Wobei k den Verteilungstyp aus den Formeln (24)-(29) darstellt. Die Maximierung wurde durch Simulated Annealing (Kirkpatrick et al. 1983) realisiert.

2.7.2 Die Rohdaten der Eigenschaften als Grundlage für die Optimierung

Mittels der bereits evaluierten optimalen Parameter aus Abschnitt 2.6.5.8 konnten die Rohdatensätze der Eigenschaften zu jeder Klasse generiert werden. Diese Rohdaten jeder Peak-Eigenschaft aus Abschnitt 2.6.5.2 bilden dann die Energielandschaft für die Optimierung. Die Optimierung wird mittels Simulated Annealing durchgeführt.

2.7.2.1 Generierung der Referenz-Verteilungen durch Simulated Annealing

Um den richtigen Verteilungstyp zu finden, wurden alle möglichen Kombinationen der Formeln (24) - (29) für jede Klasse und jede Eigenschaft evaluiert. Das heißt, es kann z. B. für die Klasse *Signal* einer Eigenschaft eine N-Verteilung zum Zuge kommen, während gleichzeitig bei der Klasse Rauschen eine NN-Verteilung angewendet wird. Diese beiden Typen wurden dann zur Berechnung der Bayesschen Wahrscheinlichkeit verwendet. Als Zielfunktion der Optimierung wurde die Maximum Likelihood-Funktion (30) gewählt.

Um eine schnelle und korrekte Optimierung durchführen zu können, galt es, den Optimierungsalgorithmus auf das vorliegende Problem abzustimmen:

2.7.2.2 Festlegung der erlaubten Konfigurationen durch Einschränkung der Parametergrenzen

Um den Konfigurationsraum effektiv durchlaufen zu können, war es nötig, nur erlaubte Konfigurationen zuzulassen. Dazu war eine Abschätzung nötig, welche die Grenzen aller zu bestimmenden Parameter für die Auswahl der erlaubten Konfigurationen festlegt.

Die Grenzen des Parameters σ der Formeln (24)-(29) wurden aus der geglätteten Verteilung mit der kleinsten möglichen Glättungsgröße von einem Pixel erhalten. Dazu wurde das minimale und das maximale σ verwendet, welches bei der Glättung der Daten des 3 Pixel breiten „sliding Window“ erhalten wurde.

Die Grenzen des Mittelwertes μ wurden durch den kleinsten bzw. größten Rohdatenwert festgelegt. Der Korrelationsparameter λ wird nur im Intervall [0,1] erlaubt.

2.7.2.3 Die Modifikation des Metropolis-Kriteriums

Da es bei einer großen Datenmenge dazu kommen kann, dass zu oft Energieverschlechterungen zugelassen werden, wurde die Gewichtsfunktion $g(\zeta,a_{BE})$ eingeführt, wobei ζ den Beitrag der Gewichtsfunktion festlegt und a_{BE} die Anzahl der bereits akzeptierten Energieverschlechterungen darstellt.

$$g(\zeta, a_{BE}) = e^{-\frac{\zeta}{\sqrt{a_{BE}+1}}} \tag{31}$$

Diese erschwert nach jeder Akzeptanz einer schlechteren Energie eine nachfolgende Energieverschlechterung innerhalb eines Temperaturschrittes, macht diese aber nie unmöglich.

Somit gilt für die Übergangswahrscheinlichkeit von Zustand Z_{alt} zu Z_{neu} folgende Metropolis-Funktion:

$$W(Z_{alt} \rightarrow Z_{neu}) = \begin{cases} e^{-\frac{E_{neu}-E_{alt}}{T}} - e^{-\frac{\zeta}{\sqrt{a_{BE}+1}}}, \text{ falls } E_{neu} > E_{alt} \\ 1, \text{ sonst} \end{cases} \quad (32)$$

Im Folgenden wurde ζ=1,5 gesetzt, da dies eine adäquate Balance zwischen Laufzeit und gutem Ergebnis darstellt.

2.7.2.4 Die Ermittlung der Start-Temperatur und die Wahl des Abkühlverfahrens

Um zu verhindern, dass der Algorithmus mit einer zu niedrigen Energie startet, wurde ein sogenannter „random walk“ durch die Energielandschaft durchgeführt. Dieser Durchgang wurde solange durchgeführt, bis 20 Energieverschlechterungen statt fanden. Von diesen Energieverschlechterungen diente die höchste als Starttemperatur und dessen verwendete Parameter als Startparameter.

Als Abkühlfunktion für die Temperatur wurde die logarithmische Abkühlung mit einem Abkühlfaktor von 0,95 verwendet. Dieser gewährleistet, dass das System nicht zu schnell einfriert.

2.7.2.5 Auswahl der erlaubten Nachbarschaftskonfigurationen

Bei einem jedem Übergang von Zustand Z_{alt} zu Z_{neu} wird zufällig ein Parameter ausgewählt, welcher dann durch einen zufälligen Wert innerhalb der zulässigen Grenzen ersetzt wird. Mit diesem einen neu bestimmten Parameter wurde dann E_{neu} berechnet und evaluiert, ob dieser neue Parametersatz eine Energieverbesserung induziert.

2.7.2.6 Definition des Abbruchkriteriums

Bedingt durch den Algorithmus existieren zwei Abbruchkriterien a_R (Anzahl der nicht akzeptierten Energieverschlechterungen innerhalb eines Temeraturschrittes) und a_{OR} (Anzahl der Energieverschlechterungen nach Absenken der Temperatur). Diese stellen sicher, dass die Optimierung endet, falls sich sowohl bei den Energieverbesserungen innerhalb eines Temperaturschrittes als auch nach Absenken der Temperatur keine

signifikante Verbesserung mehr einstellt. Sowohl a_R als auch a_{OR} werden zurückgesetzt, falls eine Energieverbesserung eintritt. Für das aktuelle Optimierungsproblem darf sowohl a_R als auch a_{OR} den Wert 10 nicht übersteigen.

2.7.2.7 Mehrfache Läufe der Optimierung zur Stabilisierung der Ergebnisse

Um zu verhindern, dass die Optimierung trotz des Metropolis-Kriteriums nur ein lokales Minimum findet, werden für jede zu optimierende Eigenschaft 5 unabhängige Durchläufe durchgeführt. Der gefundene Parametersatz mit der geringsten Energie wurde als Endergebnis angesehen und gespeichert.

2.7.2.8 Die Festlegung der besten Verteilungs-Kombination zur Berechnung der Bayesschen Wahrscheinlichkeit

Da jeder Eigenschaftstyp pro Klasse eine andere Wahrscheinlichkeitsdichteverteilung aufweist, galt es zu bestimmen, welcher Verteilungstyp (24) - (29) die Rohdaten am besten beschreibt. Dazu wurde jeder Verteilungstyp optimiert und abgespeichert. Um die beste Verteilungskombination zu erhalten, wurde analog zu Abschnitt 2.6.5.8 die Bayessche Wahrscheinlichkeit einer jeden Eigenschaft für jede Kombination getestet. Die beste Kombination, mit der die meisten Peaks korrekt wiedererkannt wurden stellte das Ergebnis dar und wurde ebenfalls abgespeichert. Die abgespeicherten Verteilungsfunktionen ermöglichten damit die Anwendung auf andere Spektren.

2.7.2.9 Anpassung der Parameter der Verteilungen auf ein anderes Spektrum

Hier wird der Rohdatensatz eines auszuwertenden Spektrums entsprechend der optimalen Eigenschafts-Parameter berechnet. Für jede Eigenschaft wurden die optimalen Berechnungs-Parametern mit dem dann die Verteilungen generiert wurden erstellt.

Das Problem war nun, dass die Verteilungen durch abweichende Daten zum Ursprungsdatensatz andere Positionierungen der Kurven aufweisen können, da die Werte anders skalieren. Somit musste die theoretische Verteilung auf den neuen Datensatz angepasst werden.

Dazu wurden die Parameter der theoretischen Verteilung auf das neue Spektrum angepasst, indem lediglich die Parameter σ_1 und σ_2 festgehalten werden, d. h. aus der nachfolgenden Nach-Optimierung ausgeschlossen werden. Es wurden nur die neuen Parameter μ_1, μ_2 und λ ermittelt.

Somit geht (30) über in folgende zu maximierende Funktion L_k.

$$L_k(\mu_1, \mu_2, \lambda) = \prod_{i=1}^{N} f_k(x_i) \tag{33}$$

Damit jedoch die Störsignale im Signalbereich nicht dominieren und dadurch die Verteilung an die falsche Position rücken kann, musste dieser Datensatz noch bereinigt werden. Hierzu wurde der Signaldatensatz anhand seiner Gaußschen Wahrscheinlichkeit (siehe Eigenschaft 3 aus 2.6.5.2) absteigend sortiert und 20 % der Peaks mit der höchsten Wahrscheinlichkeit behalten.

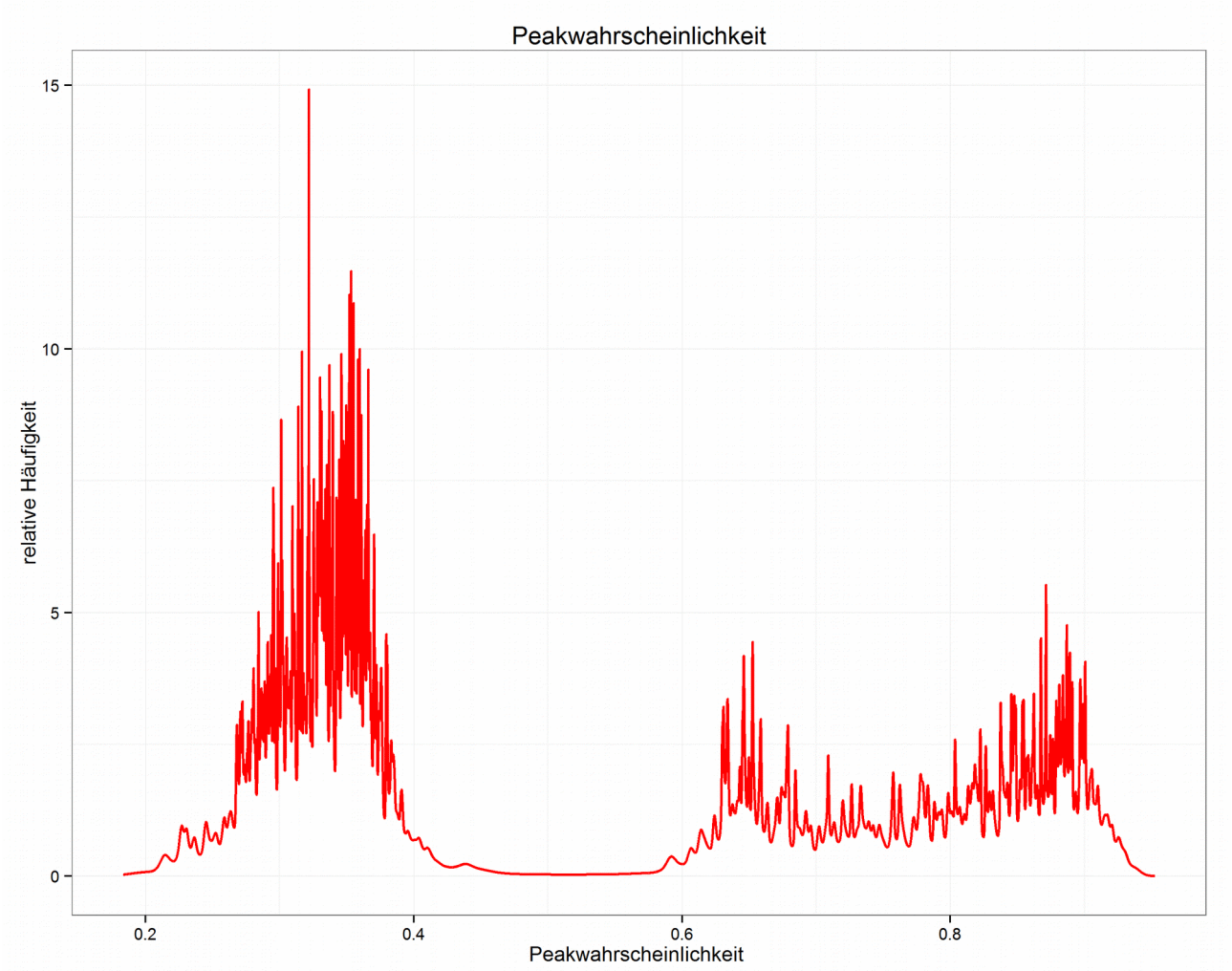

Abb. 8: Wahrscheinlichkeitsdichteverteilung der Gaußschen Wahrscheinlichkeit der Klasse Signal. Der zugrundeliegende Datensatz besteht aus allen Peaks der Peakliste innerhalb der Klasse Signal. Sie beinhaltet somit auch die Störsignale.

Abb. 8 zeigt exemplarisch am Beispiel der Wahrscheinlichkeitsdichteverteilung der gaußschen Wahrscheinlichkeit aller Peaks in der Signalklasse die Dominanz der Rauschpeaks. Es ist leicht zu erkennen, dass die Störsignale einen beträchtlichen Teil einnehmen (im Bereich der Signalwahrscheinlichkeit von 0-0,4).

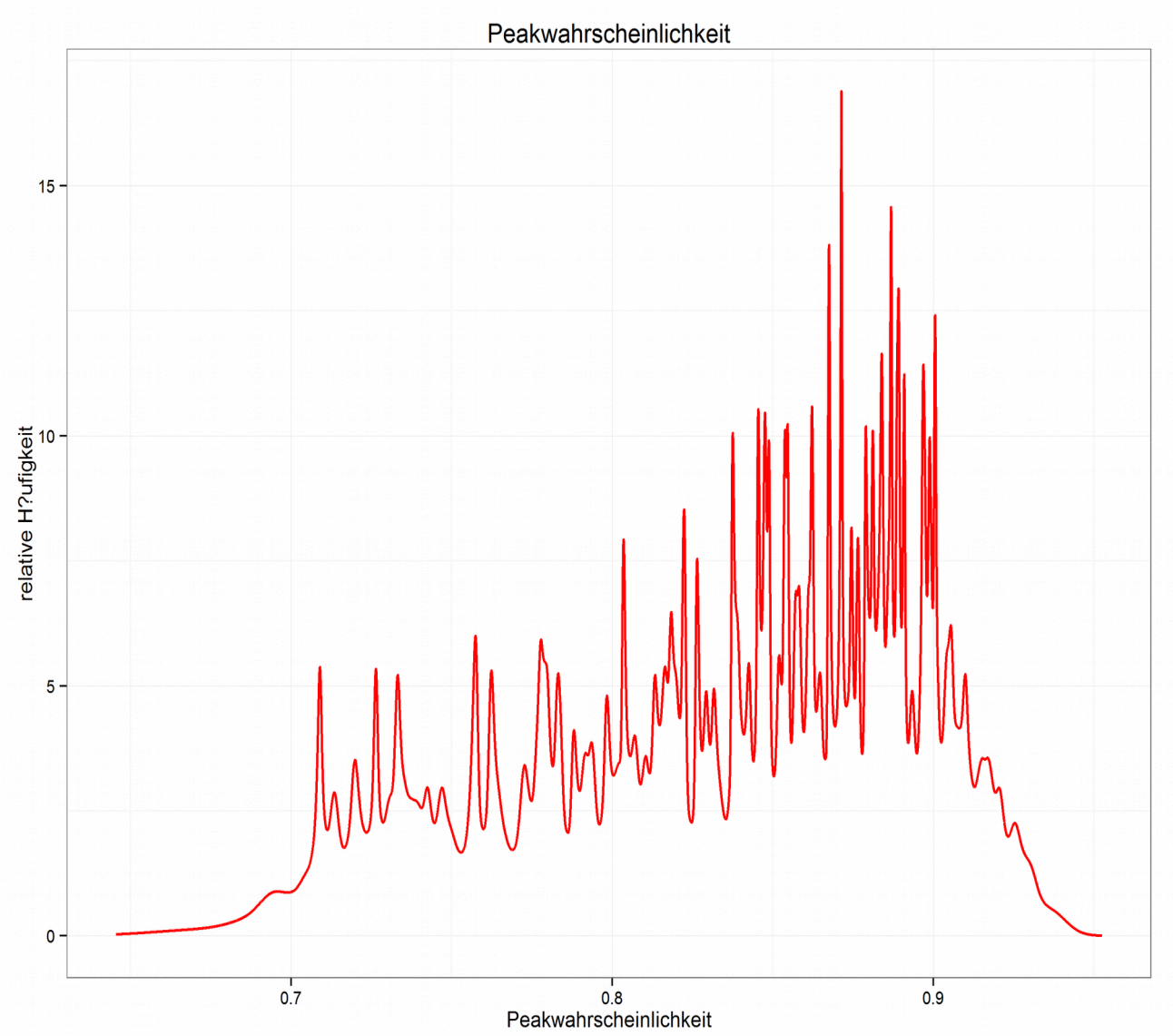

*Abb. 9: Wahrscheinlichkeitsdichteverteilung der Gaußschen Wahrscheinlichkeit von bereits bekannten Peaks der Klasse Signal. Hier wurden **nur die Peaks** des Signalbereichs miteinbezogen, welche als Signalpeaks bereits korrekt zugeordnet waren (siehe 2.2).*

Nach der Reduzierung der Verteilung aus Abb. 8 sollte die Verteilung der Wahrscheinlichkeiten in Abb. 10 nur noch einen dominierenden Bereich aufweisen, so dass die theoretische Verteilung an die korrekte Position gesetzt werden kann und der

tatsächlichen Verteilung der bekannten Signal-Peaks aus Abb. 9 näher kommen als der Verteilung mit allen Peaks aus Abb. 8.

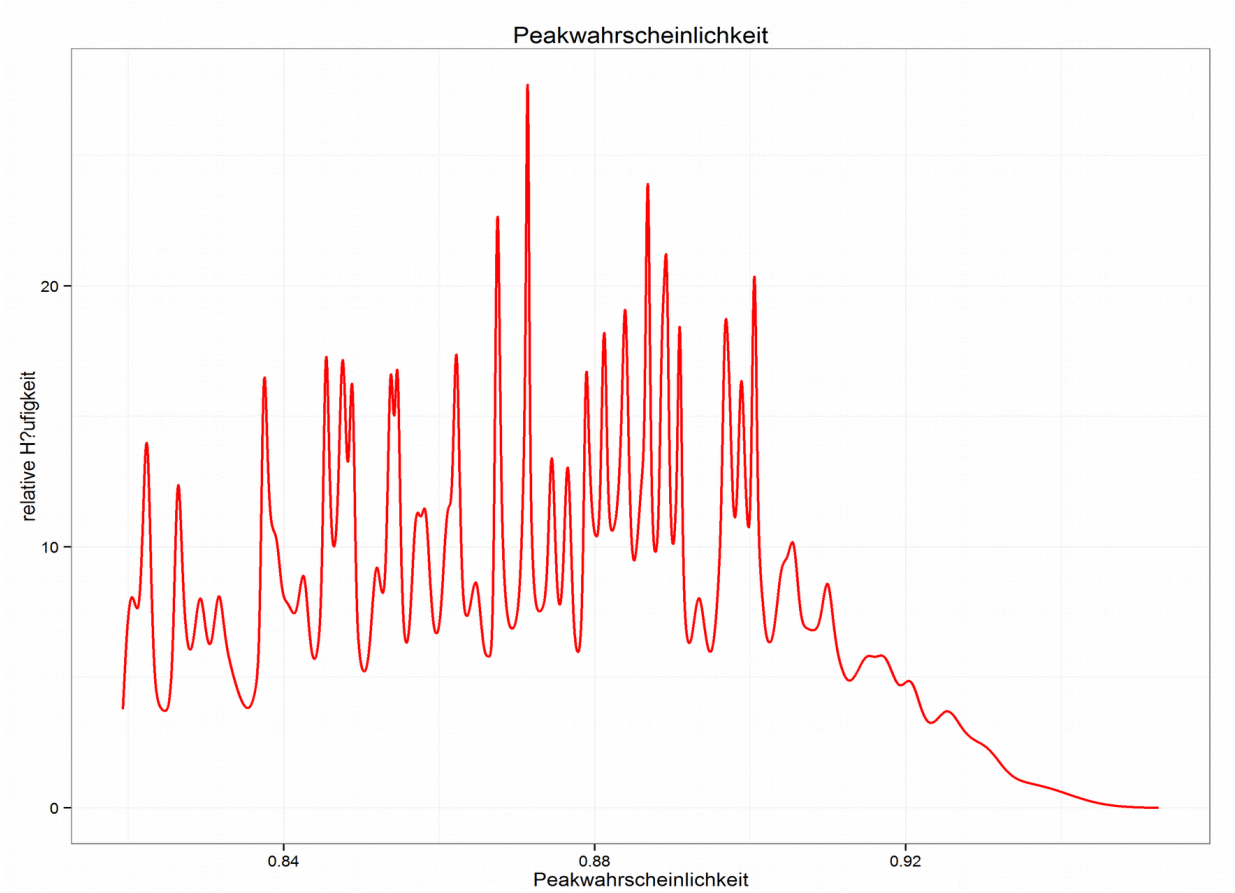

Abb. 10: Reduzierte Wahrscheinlichkeitsdichteverteilung der Gaußschen Wahrscheinlichkeit der Klasse Signal. Hier wurden alle Peaks des Signalbereichs miteinbezogen und dann auf die besten 20 % reduziert.

3 Ergebnisse

3.1 Erweiterung der Basisfunktionen von AUREMOL

3.1.1 Die Definition der Spektrumrohdaten der Frequenzdomäne des Spektrometerherstellers Bruker

Um ein Spektrum auswerten zu können, müssen die von der Equipmentkopplung zum Spektrometer erstellten Rohdaten in AUREMOL eingelesen werden. Diese Rohdaten stellen die prozessierten Frequenzdomänendaten aus den entsprechenden Zeitdomänendaten dar. Die Rohdaten der Frequenzdomäne liegen in diskretisierter Form vor. Dabei werden die Intensitäten jedes Pixels aneinandergereiht und als Vektor gespeichert. Der Aufbau des Vektors wurde vom Spektrometerhersteller Bruker rekursiv definiert. Dies legt eine Umsetzung einer rekursiven Einleseroutine nahe.

Die Anzahl der diskreten Pixel werden im Folgendem stets als digitale Auflösung bezeichnet.

3.1.2 Die Verbesserung der Einleseroutine von AUREMOL durch einen zentralen rekursiven Algorithmus

Der ursprüngliche iterative Einleseansatz wurde verworfen und durch eine rekursive Methode ersetzt. Dadurch konnte auf multiple Implementierungen verschiedener Dimensionen der Spektrumdaten verzichtet werden. Weiterhin lässt sich der Programmcode auf eine einzige zentrale Routine reduzieren und beugt Fehler durch das sogenannte „Copy&Paste“ vor.

Dies wurde in der Klasse *CFileWindow* realisiert, auf welche alle in dieser Arbeit entwickelten Module basieren und dadurch eine n-dimensionale Einsatzmöglichkeit erlauben. Die Instanz des *CFileWindows* erlaubt sowohl das Auslesen als auch das Schreiben der Spektrumrohdaten. Der Entwickler kann über Schnittstellen sowohl direkt auf den Vektor zugreifen, falls dieser weiß welche Position valid ist oder er kann durch die Angabe der Koordinaten Daten schreiben oder lesen.

Im Detail wird beim Auslesen der Rohdatendatei eine zweifache Rekursion Post-Order durchlaufen. Das heißt, es wird beginnend mit der untersten Rekursionsstufe die Lese/Schreibe-Operation ausgeführt.

Da die rekursive Form der prozessierten Rohdaten variiert, werden zusätzliche Informationen in den PROCS-Dateien beim Prozessieren angelegt. Um ein Spektrum in AUREMOL einlesen zu können, müssen folgende Informationen in den PROCS-Dateien vorhanden sein:

- Dimension (je eine PROCS-Datei pro Frequenzdomäne)
- digitale Auflösung als die Anzahl der Pixel in Richtung jeder Dimension (SI)
- Größe des Clusters (XDIM)

Diese Informationen werden durch die Instanz der Klasse *CProcFileHandler* entsprechend ausgelesen. Dazu wird die Anzahl der PROCS-Dateien ermittelt und für das *CFileWindow* bereitgestellt, damit diese die binären Rohdaten extrahieren und das digitale Raster der Intensitäten bereitstellen kann.

Das Speichern eines Spektrums erfolgt analog, jedoch immer auf Grundlage der bereits vorhandenen PROCS-Dateien.

3.1.3 Die Erweiterung des Moduls Maximum Peak-Picking zur Bestimmung der Positionen der Extrema von Signalen in einem NMR-Spektrum

Steht ein fertig prozessiertes Spektrum für AUREMOL zur Verfügung, müssen die Signale für eine spätere Strukturbestimmung identifiziert werden.

Da in AUREMOL das Peak-Picking in mehreren, verschiedenen Kopien vorlag, wurden sie Rahmen dieser Arbeit entfernt und durch eine zentrale, rekursive Routine ersetzt. Diese lagen sowohl als komplette Kopien vor, als auch als Teilkopien für die jeweilig benötigte Dimension. So existierten für eindimensionale bis vierdimensionale Spektren je eigene Module, was die Fehleranfälligkeit erhöhte.

Daher war es das Ziel, die Funktion der Peak-Picking-Routine zu zentralisieren und den Algorithmus neu zu konzipieren, damit Spektren beliebiger Dimension mit einer zentralen, rekursiven Funktion gehandhabt werden können.

Jedes Extremum kann mit Hilfe der neu eingeführten Klasse *CExtremaHandler* rekursiv über die Klasse *CFileWindow* bestimmt werden. Diese Hilfsklasse nutzt die Rekursionsfähigkeiten des *CFileWindows* aus und ermöglicht die Reduktion auf eine zentrale Funktion für alle Dimensionen.

Dazu wird ein Bereich in ein sog. „SubFileWindow“ gelesen, welcher aus dem Mittelpunkt und dessen direkten Nachbarpunkten besteht. Durch die Klasse *CExtremaHandler* wird die minimale bzw. maximale Intensität des Bereichs bestimmt. Die Größe des Bereichs hängt von der Anzahl der Dimensionen des Spektrums ab. Die Ausdehnung des Bereichs beträgt entlang jeder Dimension drei Pixel mit dessen Intensitäten.

Das SubFileWindow (also der Bereich) wird entlang der direkten Dimension iterativ über alle Pixel bewegt. Dabei werden um jeden Pixel rekursiv die Nachbarpixel der weiteren Frequenzachsen und die Nachbarpixel der direkten Dimension ergänzt.

Durch das komplette Neudesign dieses Moduls und der dynamischen Speicherverwaltung der Peakliste war es nun möglich, alle Extrema, deren Intensität >0 (bzw. <0 bei negativen Signalen) zu picken und in der Peakliste abzuspeichern (siehe Beispiel aus Abb. 11b). Ursprünglich war die Anzahl der erlaubten NMR-Signale auf ca. 65534 beschränkt. So konnte ein niedriger Schwellwert der Intensität eines Maximums dazu führen, dass die Pick-Routine das Picken abgebrochen hat. Das Ergebnis war eine Peakliste, welche lediglich die ersten 65534 NMR-Signale (Abb. 11a) enthielt und die anderen Extrema ignorierte.

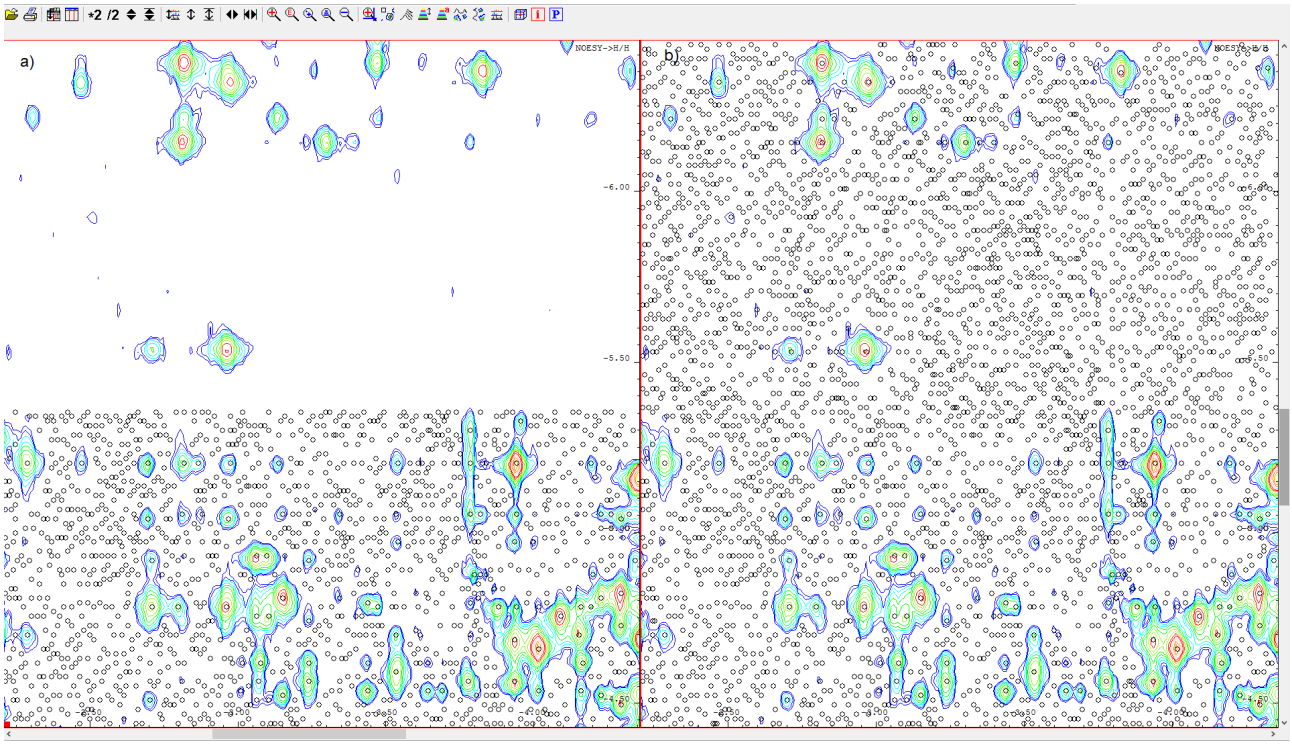

Abb. 11: Ausschnitte des zweidimensionalen Spektrums des Proteins PfTrx. In a) erkennt man die Grenze von 65534 erlaubten Peaks, bei der die ursprüngliche Pick-Routine abgebrochen hat. Der Bereich b) zeigt den Ausschnitt des gleichen Spektrums, jedoch mit allen 163727 Peaks.

Optional kann gewählt werden, ob der Schwellwert für die Intensität automatisch anhand von lokalen Rauschniveaus (Koradi et al. 1998) bestimmt werden soll.

3.1.4 Die Erweiterung der geglätteten Wahrscheinlichkeitsdichteverteilungen zur Bestimmung von theoretischen Verteilungen und Verbesserung der ursprünglichen Methode zur Ermittlung der Bayesschen Wahrscheinlichkeiten

In dieser Arbeit werden sehr oft Wahrscheinlichkeitsdichteverteilungen untersucht. Daher wurde eine Methode entwickelt, welche an die Glättungsfunktion von (Schulte et al. 1997) angelehnt ist. Der Hauptunterschied ist der, dass die Größe des Glättungsfilters nun frei wählbar ist. Zudem kann die Anzahl der gewünschten Extrema vorgegeben werden. Dazu wird der Glättungsfilter solange adaptiert, bis nur noch die gewünschte Anzahl an Extrema vorzufinden ist.

Im Folgenden werden Verteilungen mit dieser Methode generiert und untersucht. Diese erweiterte Funktionalität wurde später bei der Erstellung der theoretischen Verteilungen benötigt, um die Grenzen der Parameter während der Optimierung festzulegen. Zusätzlich

wurde eine Verwendung dieser Methode auch ohne die Verwendung theoretischer Verteilungen im Kapitel zur Berechnung der Bayesschen Wahrscheinlichkeit als Alternative analysiert.

3.1.5 Zusammenfassung mehrerer Signales zu einem Signal

Wurden in einem Spektrum NMR-Signale gepickt, bei dem Multipletts vorkommen, war es nicht möglich, diese zu einem Signal zusammenzufassen, ohne dass die Information von den Positionen der zusammengefassten Signale verloren ging. Die Methode „merge Peaks“ ermöglicht die Zusammenfassung von mehreren Signalen zu einem, wobei hier jede ppm-Position der Nebensignale und des Hauptsignals als Multiplett-Information in der Masterliste abgelegt wurde.

Die Position des zusammengefassten Signale wird als geometrisches Mittel der Positionen der zusammenzufassenden Signale in der Masterliste hinterlegt und kann nachträglich manuell oder durch die Verschiebung auf den Schwerpunkt verändert werden. Die Information der ursprünglichen Positionen bleiben in der Multiplett-Information aber stets erhalten.

3.1.6 Die Verschiebung der Signalposition auf das zugehörige Extremum

Liegt die Position eines Signals aus der Peakliste nicht auf einem Extremum, ist es möglich, die Position des Signals nachträglich durch das Modul „move peaks to nearest maximum“ auf das am nächsten liegende Extremum zu transferieren, ohne dass Informationen des Signals in der Peakliste verloren gehen.

Problemstellungen:

- Signale können z. B. nach der Durchführung von Titrationsreihen oder Druckreihen nach jedem Schritt aus dem vorhergehenden Spektrum (auch mit anderer digitaler Auflösung) importiert werden. Anschließend kann der Benutzer alle Positionen der Signale auf das nächstliegende Maximum versetzen lassen. Dabei werden nicht die Positionsangaben in SI (also die digitale Position im Raster des Spektrums) verwendet, da ansonsten die Positionen der Signale in Spektren verschiedener Auflösungen falsch wären. Daher werden die Verschiebungen ausschließlich in ppm berechnet und am Ende in einem Protokoll festgehalten, welches es dem

Benutzer ermöglicht, die durch die Positionsänderung definierten Signale zu identifizieren.

- Zur Strukturbestimmung muss gewöhnlich mehr als ein Spektrum ausgewertet werden. Das NMR-Signal eines Kernspins i ist stets durch seine Resonanzfrequenz δ_i bestimmt. Die Kreuzsignale S_k eines n-dimensionalen Spektrums sind daher durch die Frequenzen δ_j , δ_j … bestimmt. Die Signal-Position in einem Spektrum wird durch diese Frequenzen definiert. In verschiedenen experimentellen Spektren weichen diese Frequenzen oft leicht voneinander ab. Da gewöhnlich zur Strukturbestimmung mehrere Spektren parallel ausgewertet werden müssen, ist es nötig die Information der Maximum-Positionen der Kreuzsignale von Spektrum A auf Spektrum B zu übertragen. Daher muss diese Routine ein Signalmaximum S_i^A aus Spektrum A auf das Signalmaximum S_i^B in Spektrum B übertragen und dort seine Koordinate festlegen.

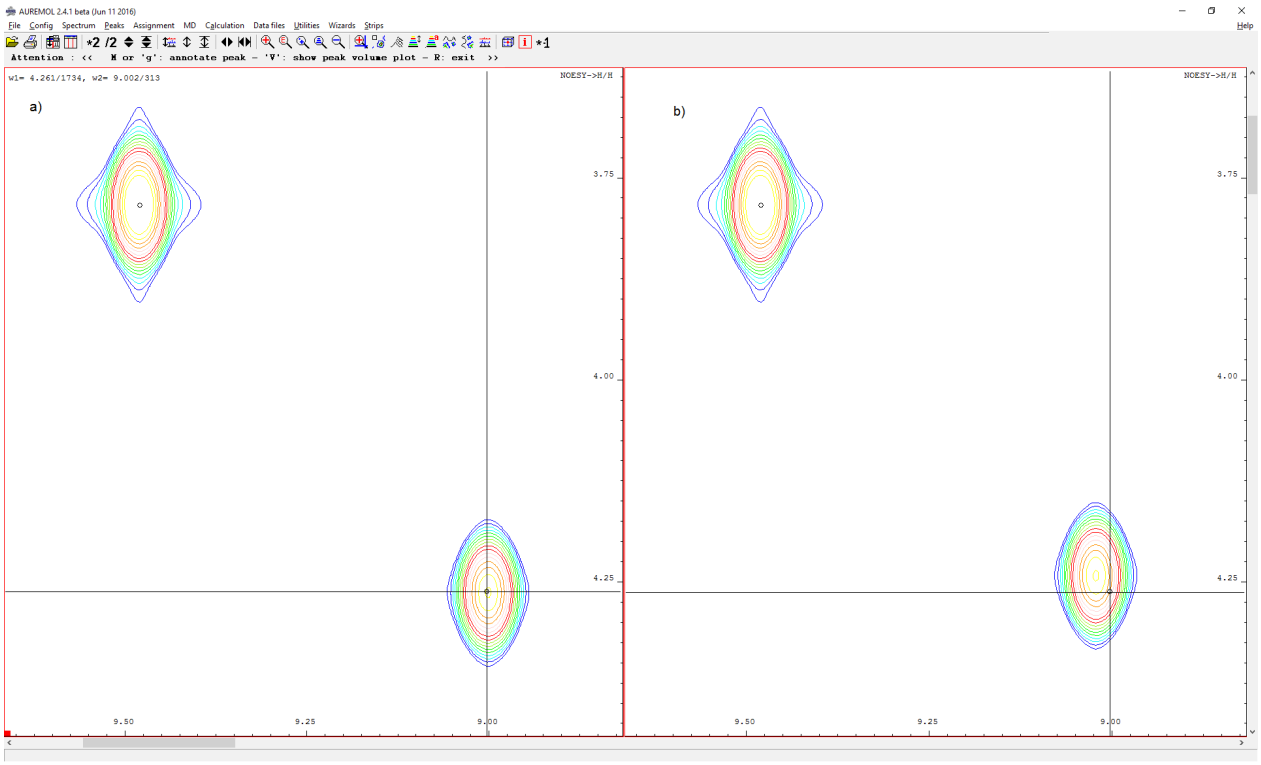

Abb. 12: Ausschnitt zweier Peaks aus je einem Spektrum, nachdem sich die Position des in den Bildern a) und b) markierten Signals geändert hat. In Abb. a) wird das Ausgangsspektrum dargestellt und die Position des unteren Signals markiert. In Abb. b) hat sich der Shift des Signals verändert und die ursprüngliche Position aus Spektrum b) passt nicht mehr zur Maximum-Position des Signals in Abb. b).

Um das jeweils nächstgelegene Extremum im Spektrum zu bestimmen, wird die Intensität der Signalposition zentral mit dessen benachbarten Intensitäten ausgelesen und in einem Bereich (analog zur Vorgehensweise beim Maximum-Picken) eingelesen. Wird eine höhere oder gleiche Intensität unter den Nachbarn gefunden, markiert der Algorithmus diese höhere (bzw. im negativen Bereich niedrigere) Intensität temporär als neues Extremum. Dieses temporäre Extremum des Bereichs aus dem Spektrum bildet wiederum das neue Zentrum des um einen Pixel versetzten nächsten Suchbereichs, um den abermals die Nachbarpixel mit den jeweiligen Intensitäten rekursiv eingelesen werden.

Dies wird solange wiederholt, bis eine temporäre Position in seiner direkten Nachbarschaft keine Positionen mit höherer Intensität mehr aufweist. Dies entspricht einer Suche entlang des steilsten Anstiegs (bzw. Abfalls bei negativen Intensitäten) der Intensitäten innerhalb des ausgewählten Bereichs der Positionen mit deren Intensitäten. Trifft die Routine auf Nachbarn gleicher Intensität, fährt dieser solange fort, bis das „Plateau“ aus Punkten gleicher Intensität abfallen (bzw. aufsteigen) würde.

Diese gefundene Extremum-Position gilt als finale Position des Signals und wird in der Peakliste abgespeichert.

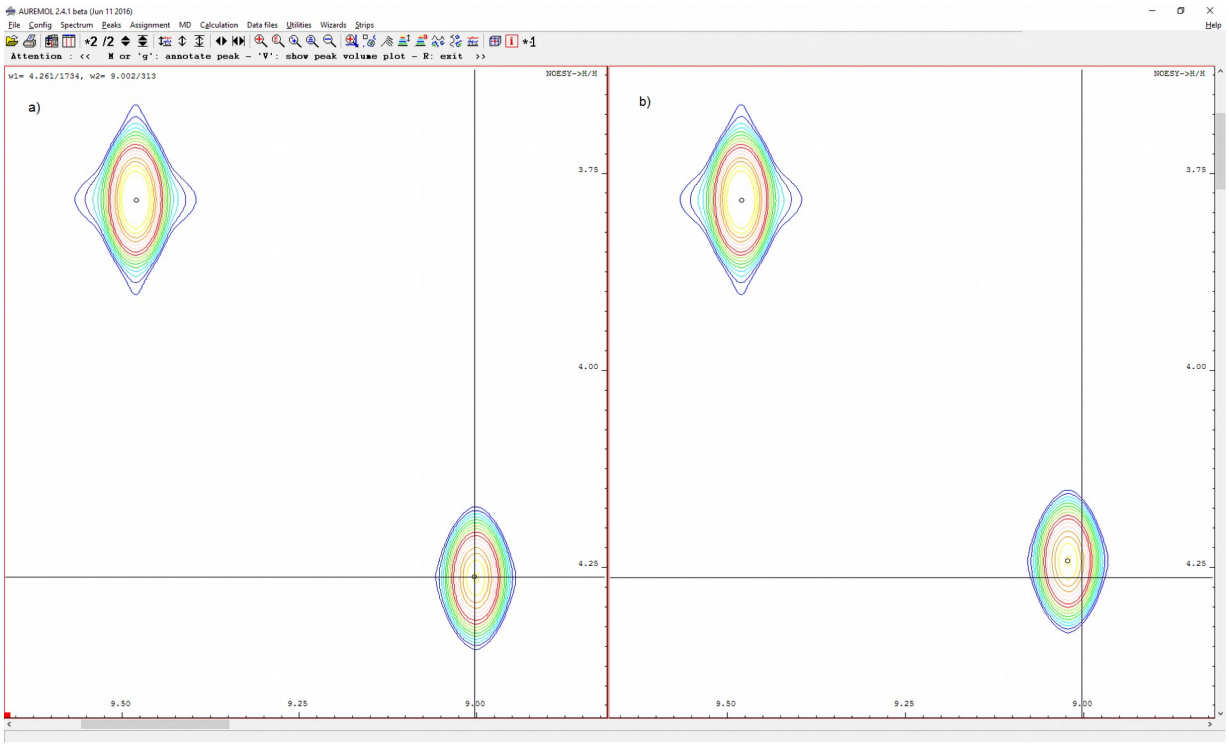

Abb. 13: Ausschnitt zweier Peaks aus je einem Spektrum, nachdem sich die Position des in den Bildern a) und b) markierten Signals geändert hat. In Abb. a) wird das Ausgangsspektrum dargestellt und die Position des unteren Signals markiert. In Abb. b) hat sich der Shift des Signals verändert und die ursprüngliche Position aus Spektrum b) passt nach Anwendung der Routine wieder zur Maximum-Position des Signals in Abb. b).

Würde die Routine für mehr als ein Signal dieselbe Extremum-Position bestimmen, wird das Signal verschoben, dessen Ausgangsposition die geringste euklidische Distanz zum gefundenen Extremum aufweist. Somit ist nur für das Signal ein Positionstransfer erlaubt, dessen Position dem gefundenen Extremum an nächsten liegt. Die weiteren Signale werden dann auf deren Position belassen. Diese können nun durch die Erweiterung der Integrationsroutine integriert werden, da ursprünglich die Signale, welche keine Extremum-Position in der Peakliste definierten, nicht integriert werden konnten.

Diese Entscheidung kann der Benutzer zusätzlich beeinflussen, indem er im obigen eintretenden Fall (ein Extremum für mehrere Signale gefunden) folgendes festlegt:

- es wird kein Signal aus seiner Position auf das nächste Extremum verschoben
- alle Positionen der Signale werden auf das Extremum verschoben (auch von diesen Signalen kann mit der erweiterten Integration das Volumen bestimmt werden)

Um zu vermeiden, dass einem Signal ein Extremum zugewiesen werden würde, welches durch eine niedrige Auflösung digital (also in Pixeln) näher liegt als ein anderes, das in

Richtung der Dimension mit höherer Auflösung liegt, wird das Auflösungsverhältnis mit in das Distanz-Kriterium eingebracht.

Somit gilt für die euklidische Distanz d von der Ausgangsposition P_i eines Kreuz-Signals zur Position des gefundenen Extremums M_i unter Berücksichtigung der digitalen Auflösung a_i, wobei i die jeweilige digitale Auflösung in Pixel in der jeweiligen Dimension von 1 bis N darstellt:

$$d=\sqrt{\sum_{i=1}^{N}\left(\frac{P_i-M_i}{a_i}\right)^2} \tag{34}$$

3.2 Verbesserte Integration der Signalvolumen

Dieser Abschnitt zeigt die Verbesserung der Integration zu der ursprünglichen Version von AUREMOL vom September 2009 auf. Da die Verbesserungen mehrere Variationen aufweisen, werden diese im Folgenden separat aufgeführt und abschließend anhand von simulierten Spektren des Proteins *PfTrx* in verschiedenen Auflösungen und mit der Zugabe von Rauschen evaluiert.

3.2.1 Automatische Größenermittlung der Integrationsbox und dynamische Erhöhung der Integrationsschritte

Diese Methode bestimmt adaptiv die Größe des zu verwendenden Integrationsfensters in mehreren Iterationen. Dazu wird das Integrationsfenster eines Signals solange vergrößert und die Integrationsschritte variiert, bis sich das Volumen nicht mehr signifikant ändert.

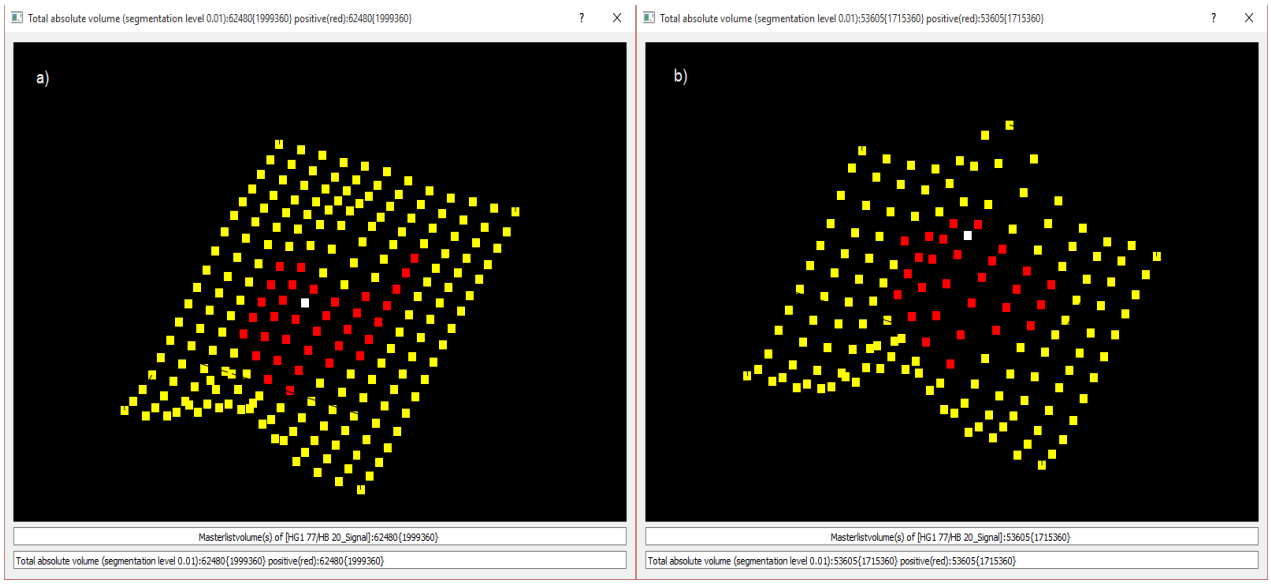

Abb. 14: Integration mit und ohne adaptive Ermittlung der Integrationsbereiche. Die roten Pixel stellen jeweils das Volumen des NMR-Signals dar. Screenshot a) zeigt die Integration ohne automatische Größenermittlung der Integrationsbox und adaptiver Erhöhung der Integrationsschritte. Screenshot b) zeigt die Integration mit automatischer Größenermittlung der Integrationsbox und adaptiver Erhöhung der Integrationsschritte.

Man kann erkennen, dass das Signal in Abb. 14a längere Ausläufer in Richtung des Nachbarsignals aufweist, als bei aktiviertem erweitertem Modus in Abb. 14b. Es ist eine schärfere Abgrenzung zum Volumen des Nachbarsignals erkennbar.

3.2.2 Integration stark zerklüfteter Signale

Um ein stark zerklüftetes oder ein stark verrauschtes Spektrum zu evaluieren, wurde für die Abb. 15a und 15b das simulierte Spektrum mit Rauschen aus Kapitel 2.2.4 verwendet. Dazu wurde der Rauschanteil auf 2 % erhöht, um Rauscheffekte in der Signalform zu erreichen. Die Auflösung des zweidimensionalen Spektrums liegt hier bei 2kx4k.

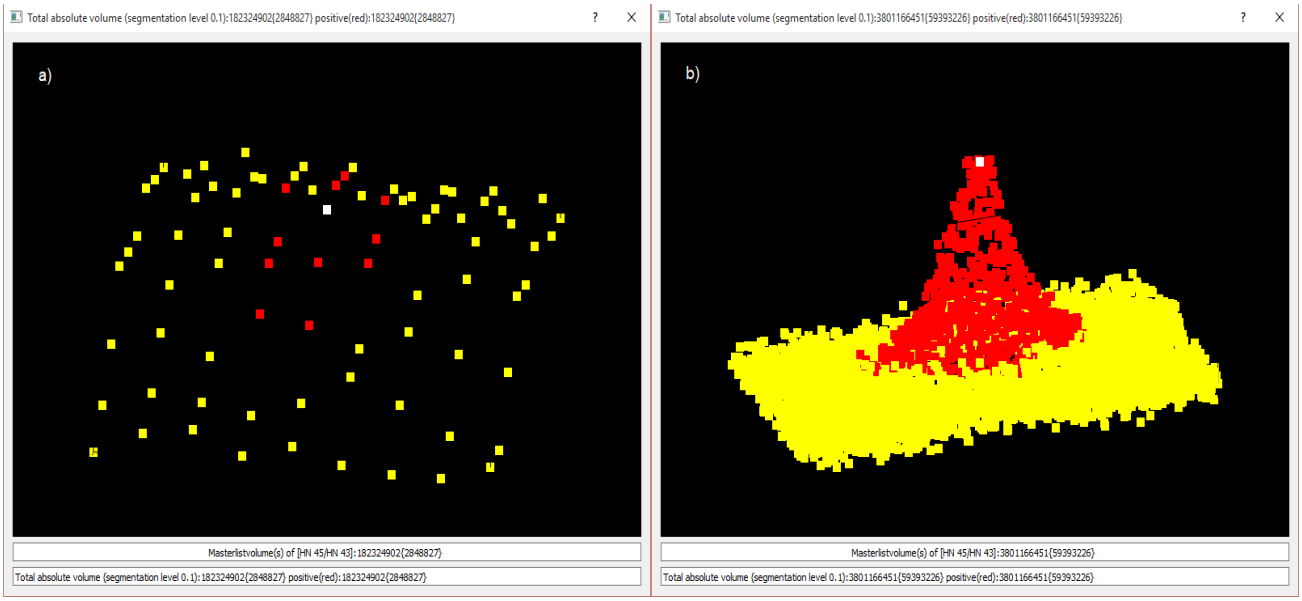

Abb. 15: Vergleich der ursprünglichen Integration mit der erweiterten Integrationsmethode bei einem stark verrauschten NMR-Signal. a) zeigt die Integration ohne Glättung eines verrauschten Signals. Der Algorithmus stoppt zu früh unmittelbar neben dem Startseed durch die lokalen Extrema. b) zeigt die Integration mit Glättung eines verrauschten Signals. Der Algorithmus kann bis zum angegebenen Segmentierungslevel das komplette Volumen segmentieren, ohne durch lokale Extrema (verursacht durch das Rauschen) gestoppt zu werden.

In Abb. 15a ist zu erkennen, dass der ursprüngliche Algorithmus an den durch den Rausch entstandenen lokalen Neben-Extrema stoppt und diese als Abgrenzung der Signalform missinterpretiert. Daher fällt der Intregrationsbereich auch sehr klein aus. In Abb. 15b erkennt man, dass der Algorithmus nun dasselbe Signal bis zur korrekten Abgrenzung der Signalform segmentieren kann, da der Wachstumsalgorithmus nicht durch die rauschbedingten Neben-Extrema gestoppt wird. Dies resultiert in einem ausreichend großen Integrationsbereich, welcher die gesamte Signalform zur Volumenbestimmung erfasst.

3.2.3 Integration mehrerer Signale mit Extremum an derselben Position

Liegen die Extrema zweier Signale an derselben Position am digitalen Raster, stellte dies in der ursprünglichen AUREMOL-Version ein Problem dar, denn es durfte für die korrekte Berechnung nur ein Signal an einer digitalen Position liegen. Ein weiteres Signal musste in den Eigenschaften des Signals in der Peakliste manuell hinzugefügt werden. Dies hatte jedoch nur einen rein informellen Charakter, da diese Information bislang von keinem anderen Modul in AUREMOL berücksichtigt wurde.

Im Falle von zwei Signalen aus der Peakliste an derselben Position hätte die ursprüngliche Integration für beide Signale dasselbe volle Volumen durch die Segmentierung berechnet.

Nun werden die Volumen der Signale relativ zu ihrer Intensität korrigiert und im *Integrations-Hash* entsprechend gekennzeichnet. Zusätzlich greift diese Methode auch für Fälle aus dem nächsten Abschnitt, falls Positionen von Signalen temporär zusammenfallen können (siehe Abb. 18).

3.2.4 Methoden zur Integration von Signalen, deren Position nicht an einem Extremum liegt

Da die ursprüngliche Version der Integration bei zweidimensionalen Spektren (oder höher) die Segmentierung zur Berechnung der Volumen verweigerte, welche keine Position am Extremum aufwiesen, wurden folgende drei Ansätze implementiert, um das Volumen eines Signals bestimmen zu können.

3.2.4.1 Methode 1 – Integration ohne Veränderung der Signalpositionen

Bei dieser Methode werden die Volumen der Signale mittels der angegebenen Position „straight forward“ integriert. Dies hat den Effekt, dass die Segmentierung an dieser Position beginnt und bis zum Segmentierungslevel das Volumen bestimmt. Liegt ein Signal mit nur einer Signalposition vor, wird der Rest des Volumens oberhalb dieser Signal-Position ignoriert, da diese sich ja nicht an einem Extremum befindet. Dies ist dann der Fall, wenn z. B. die Signalposition aus dem Extremum verschoben wurde oder eine Simulation eines Spektrums erstellt wurde, welche Positionen von Signalen festlegt, die nicht an einem Extremum liegen.

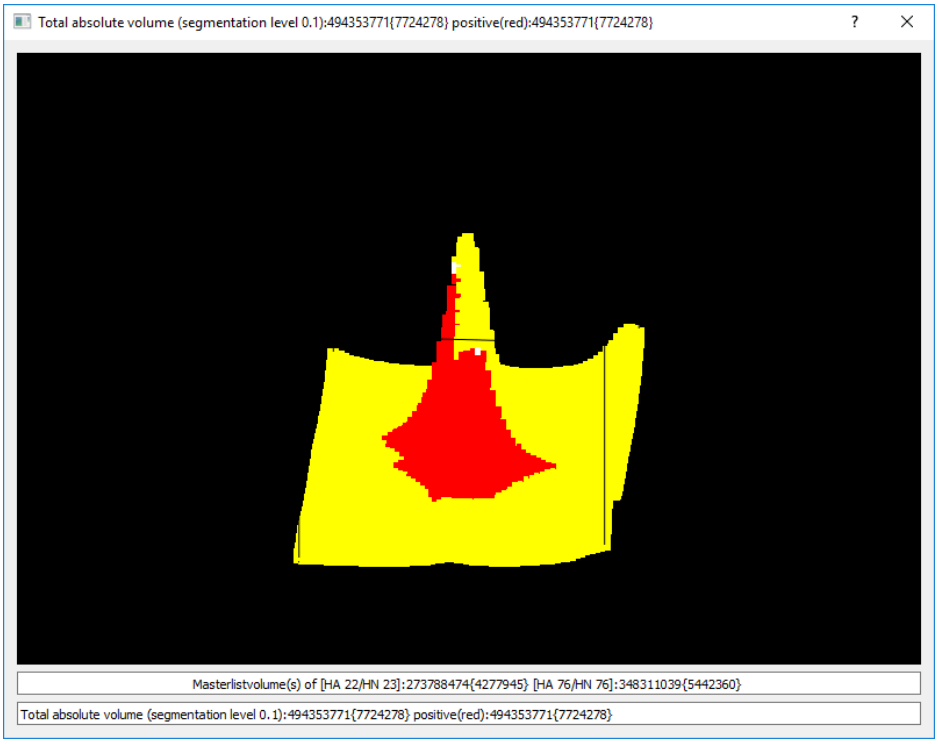

Abb. 16: Volumen (rot) ermittelt durch Methode 1 von zwei Signalen, welche sich eine Signalform teilen und beide nicht an einem Extremum liegen. Die Methode 1 belässt dabei beide Signale an deren Position (weiße Pixel) und beginnt dort jeweils mit der Segmentierung für die Volumen (rot).

In Abb. 16 erkennt man, dass diese Methode im Realfall nicht zu empfehlen ist. Diese wird aber der Vollständigkeit halber mit aufgeführt und soll zeigen, dass die erweiterte Integration auch Signale, deren Positionen nicht an einem Extremum liegt, integrieren kann. Diese Methode stellt somit den Ausgangspunkt für die folgenden Methoden 2 und 3 dar.

3.2.4.2 Methode 2 - Integration mit nur einem erlaubten nächsten Extremum

Hier wird eine Signalposition an das nächste Extremum der Signalform verschoben, falls kein anderes Signal Vorrang für dieses Maximum (siehe 3.1.6) hat oder ein Signal bereits am Maximum gepickt ist.

Alle nicht verschobenen Signale werden dann analog zur Methode 1 separat integriert (siehe Abb. 17). Diese nicht verschobenen Signale werden stets in die Seedlisten aufgenommen (obwohl diese Positionen keine Extrema sind). Dies ermöglicht dem Wachstumsalgorithmus an dieser Stelle zu stoppen um das Signal mit diesem temporär gesetztem Pseudo-Extremum abzugrenzen.

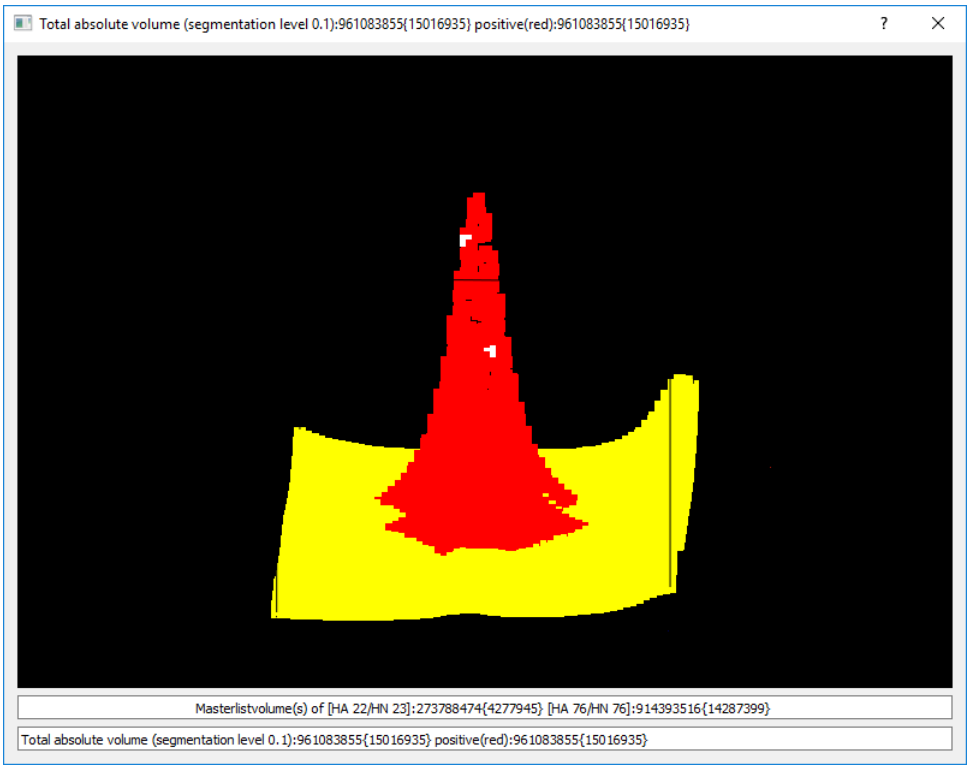

Abb. 17: Volumen (rot) ermittelt durch Methode 2 von zwei Signalen, welche sich eine Signalform teilen und beide nicht an einem Extremum liegen. Die Methode 2 belässt dabei das entferntere Signal vom Extremum an seiner Position (unterer weißer Pixel) und setzt das nächstgelegene Signal (oberer weißer Pixel) an das Extremum.

3.2.4.3 Methode 3 - Integration durch temporäre Verschiebung aller Positionen von Signalen an das Extremum einer gemeinsamen Signalform

Bei dieser Methode werden alle Positionen der Signale temporär an deren nächstes Extremum platziert um zu verhindern, dass alle Positionen abseits eines Extremums einer gemeinsamen Signalform in die Seedliste aufgenommen werden. Deren Volumen wird mit der Methode aus 3.2.3 ermittelt (siehe Abb. 18).

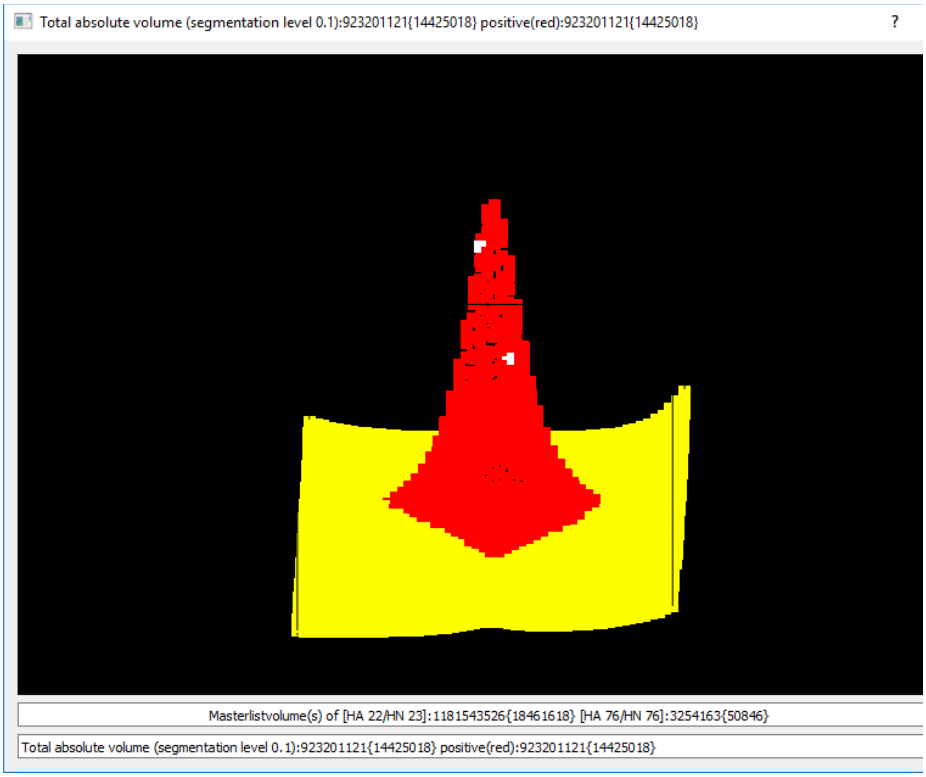

Abb. 18: Volumen (rot) ermittelt durch Methode 3 von zwei Signalen, welche sich eine Signalform teilen und beide nicht an einem Extremum liegen. Die Methode 3 verschiebt dabei beide Positionen der Signale (weiße Pixel) temporär an das Extremum.

3.2.5 Die Integration von Multipletts und zusammengefasster Signale

Liegt ein simuliertes Spektrum vor, welches Multipletts aufweist (Abb. 19) oder manuell zusammengefasste Signale existieren (siehe 3.1.5), werden die Multiplett-Signale aus der Masterliste temporär in Einzelsignale zerlegt und nach der Integration wieder zusammengefügt.

In Abb. 19b sieht man das integrierte Quadruplett, bei dem sich die Position des Signals in der Mitte des Quadrupletts befindet. Abb. 19b stellt das Volumen des Signals aus Abb. 19a dar. Gut zu erkennen ist die Position (weiss) des Signals und dessen Volumen (rot), welche auch die Quadrupletts berücksichtigt hat.

Der Algorithmus der ursprünglichen Integration hätte lediglich den Pixel in das Signalvolumen aufgenommen, an dem sich die Position des Signal befindet, da sich diese Position bereits im lokalen Minimum mittig des Quadrupletts befindet und den Startseed darstellt, welcher nicht wachsen kann.

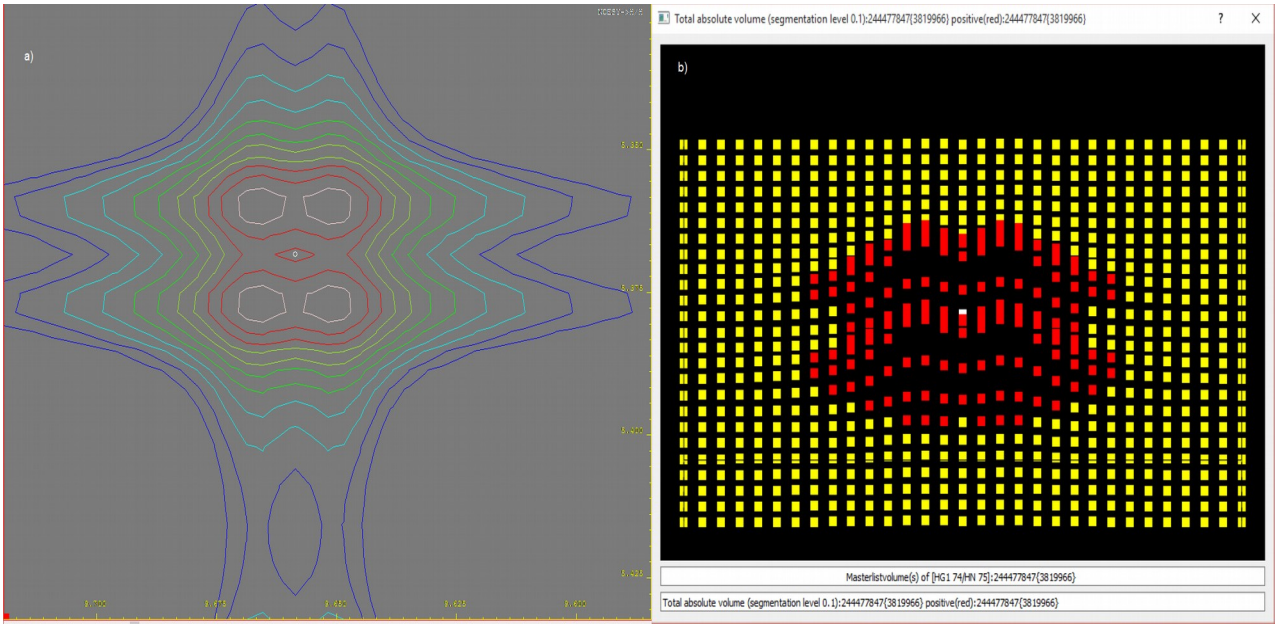

Abb. 19: Screenshots eines Quadruplett-Signals im simulierten zweidimensionalen PfTrx-Spektrum. Screenshot a) zeigt den vergrößerten Bereich um das relevante NMR-Signal aus dem Sprektrum. Screenshot b) zeigt das Volumen (rot) des Quadruplett-Signals mit der Position (weiss) im Zentrum der vier Nebenmaxima im simulierten zweidimensionalen Spektrum von PfTrx.

3.2.6 Vergleich der Integrationsmethoden 1 bis 3 mit der ursprünglichen Integration

Im Folgenden werden die Ergebnisse der Integration separat für Spektren verschiedener Dimensionen dargestellt.

Hierbei wurden alle zu integrierenden Signale in folgenden Typen unterteilt:

- Die Signalposition befindet sich an einem Extremum der Signalform bezüglich des digitalen Rasters
- Die Signalposition befindet sich nicht an einem Extremum der Signalform bezüglich des digitalen Rasters
- Es befinden sich mehrere Positionen von Signalen an einer Position einer gemeinsamen Signalform bezüglich des digitalen Rasters

Über alle Signale eines Typs wird dann sowohl der Median als auch der Mittelwert der Volumenverhältnisse aller Signale eines Typs gebildet. Für das Verhältnis d_v des integrierten Volumens V_i zum theoretischen Volumen V_t oder umgekehrt gilt:

$$d_v = \begin{cases} \frac{V_t}{V_i} - 1, \text{ falls } V_t \geq V_i \\ \frac{V_i}{V_t} - 1, \text{ falls } V_t < V_i \end{cases} \quad (35)$$

Somit wäre das optimale Verhältnis 0, d.h. beide Volumen stimmen überein. Der Grund dafür, dass stets das Verhältnis des größeren zum kleineren Volumen verwendet wird ist der, dass bei der Bildung des Mittelwertes (bzw. des Medians) die Skalierung beibehalten werden muss. Denn das Verhältnis eines kleineren zu einem größeren Verhältnis würde immer zwischen 0 und 1 skalieren. Das würde zur Folge haben, dass solche Werte immer als relativ gut bewertet werden würden.

Die Verwendung des Medians soll das Gewicht der Ausreißer reduzieren und die Aussagekraft der Ergebnisse verbessern. Daher wird zum Vergleich der Methoden ausschließlich der Median als Kriterium für die Bewertung herangezogen. Alle folgenden Daten wurden mit einer Segmentierungstiefe von 0,001 durchgeführt.

Da in simulierten Spektren viele Extremum-Positionen mit bis zu acht Signalen zugeordnet waren und sich auch mehrere Signale eine nicht-Extremum-Position teilten, wurden folgende Signale aus der Auswertung entfernt:

- Signale welche sich mit einem oder mehreren Signalen diese nicht-Extremum-Position teilten
- Signale, welche sich mit mehr als einem anderen Signal eine Extremum-Position teilten

In den Ergebnissen wurden lediglich Positionen betrachtet, welche maximal zwei Signale an einer Position (am Extremum) aufwiesen und nur ein Signal an einer nicht-Extremum-Position aufweisen, da diese im Normalfall nur selten auftreten und der Algorithmus hier an seine Grenzen stößt.

3.2.6.1 Ergebnisse der Integrationen eines eindimensionalen simulierten Spektrums des Proteins *PfTrx*

Da in der AUREMOL-Version von 2009 die Integration für eindimensionale Spektren zwei verschiedene Module aufwies, wurden beide Module (Integration durch die Signalform und Integration durch einen festgelegten Bereich) mit den Methoden 1 bis 3 verglichen.

Tabelle 9: Ergebnisse der Integration eines eindimensionalen simulierten Spektrums (PfTrx) bei den Auflösungen von 4k und 32k mit Rauschen. Fett hervorgehobene Werte stellen dabei die schlechtesten Ergebnisse dar und grau die besten.

Spektrum	Position am Extremum	Anzahl Signale an der selben Position	Methode 1: verschiebe keine Signale		Methode 2: verschiebe Signal mit geringster Distanz zum Extremum		Methode 3: verschiebe alle Signale an deren nächsten Extremum		Methode Version 2009: (Integration durch Bereich)		Methode Version 2009: (Integration durch Peakform)		Anzahl der integrierten Signale
			Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	
4k	j	1	1,492	0,578	1,495	0,578	0,369	0,183	19,494	1,259	25,855	**1,434**	32
4k	j	2	2,077	1,411	2,077	1,411	0,304	0,363	4,897	**6,424**	4,952	3,477	12
4k	n	1	8,745	1,093	8,33	0,979	3,089	0,406	35,967	**4,445**	26,87	2,412	324
32k	j	1	0,497	0,431	0,497	0,431	0,183	0,105	0,747	0,289	1,379	**1,472**	20
32k	j	2	2,58	**3,138**	2,58	**3,138**	0,176	0,296	1,684	1,684	0,09	0,09	2
32k	n	1	8,242	1,159	7,907	1,098	25,418	7,614	79,151	**5,363**	23,723	1,302	405

In Tabelle 9 war im Falle des Spektrums mit der digitalen Auflösung von 4096 Pixeln die Methode 3 die beste Methode (grau hinterlegt). Dabei lieferten beide ursprünglichen Methoden schlechtere Werte und die größten Ausreißer (siehe Spalten der Mittelwerte aus Tabelle 9). Zudem wurden von den 12 zu integrierenden Signalen im Gegensatz zur aktuellen Methode sechs Signale (bei zwei Extrema) nicht integriert und mit einem Volumen von 0 versehen. Beim Spektrum mit 32k Auflösung konnte bei zwei Extrema mit der ursprünglichen Methode nur einer der zwei Signale integriert werden. Alle restlichen Signale wurden von allen Methoden integriert.

Vergleicht man die Methoden 1 bis 3, so lieferte Methode 3 bezüglich der Methode 1 und 2 bessere Resultate. Zudem konnten alle Signale integriert werden.

Relevant für den Vergleich der aktuellen Methoden und der Methoden der ursprünglichen Integration sind lediglich die Werte, bei denen sich ein Signal am Extremum der Signalform befindet. Die restlichen Ergebnisse sollen aufzeigen, dass die aktuelle

Integration bei nahezu allen Signalen die bessere Wahl darstellt. So schneidet die Methode 3 wesentlich besser ab als die ursprünglichen Methoden. Bezüglich den Abweichungen bei Verwendung eines Spektrums mit einer Auflösung von 4k erweist sich Methode 3 als die bessere Methode. Methode 2 dominiert beim Spektrum mit einer Auflösung von 32k, da diese die meisten Signale (405) besser Integrierte. Bei den 20 Signalen, welche die Position am Extremum aufwiesen, war Methode 3 die bessere.

Tabelle 10: Ergebnisse der Integration eines eindimensionalen simulierten Spektrums (PfTrx) bei den Auflösungen von 4k und 32k ***mit dem eingeschränkten Bereich*** *der zu integrierenden Signale von 11,643 ppm bis 5,915 ppm mit Rauschen. Fett hervorgehobene Werte stellen dabei die schlechtesten Ergebnisse dar und grau hinterlegt die besten.*

Spek-trum	Posi-tion am Extre-mum	Anzahl Signale an der selben Position	Methode 1: verschiebe keine Signale		Methode 2: verschiebe Signal mit geringster Distanz zum Extremum		Methode 3: verschiebe alle Signale an deren nächstes Extremum		Methode Version 2009: (Integration durch Bereich)		Methode Version 2009: (Integration durch Peakform)		Anzahl der inte-grierten Signale
			Mittel-w ert	Medi-an	Mittel-w ert	Medi-an	Mittel-w ert	Medi-an	Mittel-w ert	Medi-an	Mittel-w ert	Medi-an	
4k	j	1	0,909	0,446	0,914	0,446	0,27	0,153	1,197	0,351	2,292	**0,915**	19
4k	j	2	0,5	**0,662**	0,5	**0,662**	0,117	0,183	0,461	0,461	0,042	0,042	2
4k	n	1	5,67	0,833	5,342	0,785	4,053	0,375	27,821	**3,689**	8,961	1,387	92
32k	j	1	0,437	0,431	0,437	0,431	0,151	0,105	0,624	0,243	1,307	**1,472**	14
32k	n	1	4,268	0,735	4,204	0,71	12,132	1,243	42,311	**4,995**	7,591	0,919	101

Tabelle 10 zeigt die Ergebnisse analog zur vorherigen Tabelle 9 auf, jedoch mit dem Unterschied, dass die Bereiche der zu integrierenden Signale auf einen Bereich von 11,643 ppm bis 5,915 ppm eingegrenzt wurden, da der Rest des Spektrums starke Überlappungen und enorm viele Positionen von Signalen abseits eines Extremum aufwies. Dies sollte die Bedingungen für die Integration erleichtern, da es sich um eine schwierige Testumgebung handelte, welche im realen Experiment durch manuelles Picken von nicht-Extrema und/oder automatischem Picken nicht gegeben ist. Die Resultate weisen aber auch hier denselben Trend wie zuvor in Tabelle 9 auf.

Die Methode 2 schneidet gegenüber der ursprünglichen Methode „Integration durch Peakform“ deshalb schlechter ab, da diese lediglich eines der beiden Signale integrierte. Alle restlichen Signale wurden von allen Methoden integriert.

Die Abweichungen des Spektrums mit einer Auflösung von 4k weist die Methode 3 als die bessere Methode auf. Wogegen die Methode 2 beim Spektrum mit einer Auflösung von 32k dominiert, da diese die meisten Signale (101) besser integrierte. Bei den 14 Signalen, welche die Position am Extremum aufwiesen, war die Methode 3 wie in Tabelle 9 am effektivsten.

3.2.6.2 Ergebnisse der Integrationen eines zweidimensionalen simulierten Spektrums des Proteins *PfTrx*

Im Unterschied zur Integration eines eindimensionalen Spektrums wartet die ursprüngliche Version von AUREMOL mit nur einem Modul für die zweidimensionale Integration auf. Daher wurden nur die Methoden 1 bis 3 und die Version von 2009 verglichen. Wie bereits erwähnt, kann die ursprüngliche Version keine Signale integrieren, falls sich deren Position nicht an einem Extremum befindet. Daher fehlen in den Tabellen dieses Abschnittes die entsprechenden Werte für diesen Fall (NA).

Tabelle 11: Ergebnisse der Integration von zweidimensionalen simulierten ^{1}H-^{1}H-NOESY-Spektren (PfTrx) bei verschiedenen Auflösungen sowohl mit als auch ohne Rauschen. Fett hervorgehobene Werte stellen dabei die schlechtesten Ergebnisse dar und grau hinterlegt die besten.

Spektrum	Position am Extremum	Anzahl Signale an der selben Position	Methode 1: verschiebe keine Signale		Methode 2: verschiebe Signal mit geringster Distanz zum Extremum		Methode 3: verschiebe alle Signale an deren nächsten Extremum		Methode Version 2009		Anzahl der integrierten Signale
			Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	
2kx1k	j	1	0,402	0,167	0,404	0,167	0,347	0,152	1,114	**0,244**	1118
2kx1k	j	2	0,87	0,321	0,871	0,321	0,367	0,225	13,555	**1,948**	124
2kx1k	n	1	4,752	0,901	4,688	0,8	10,324	1,012	NA	**NA**	1622
2kx1k (Rausch)	j	1	0,393	0,173	0,394	0,174	0,332	0,163	1,113	**0,235**	1109
2kx1k (Rausch)	j	2	0,873	0,336	0,874	0,336	0,369	0,224	13,617	**1,95**	124
2kx1k (Rausch)	n	1	4,991	0,94	4,676	0,794	10,307	0,999	NA	**NA**	1631
2kx2k	j	1	0,404	0,168	0,406	0,167	0,329	0,151	0,738	**0,241**	1069
2kx2k	j	2	1,269	0,386	1,271	0,389	0,368	0,208	16,351	**2,23**	112
2kx2k	n	1	3,915	0,825	3,855	0,74	19,337	1,118	NA	**NA**	1661
2kx2k (Rausch)	j	1	0,404	0,188	0,404	0,183	0,321	0,171	0,723	**0,236**	1043
2kx2k (Rausch)	j	2	1,305	0,402	1,306	0,402	0,378	0,227	16,584	**2,277**	110
2kx2k (Rausch)	n	1	4,079	0,839	3,798	0,722	19,167	1,098	NA	**NA**	1687
4kx4k	j	1	0,38	0,16	0,381	0,16	0,271	0,139	0,458	**0,267**	899
4kx4k	j	2	1,991	0,257	1,994	0,312	0,334	0,202	20,542	**1,494**	78
4kx4k	n	1	3,886	0,753	4,042	0,698	84,61	1,854	NA	**NA**	1885

Vergleicht man die Ergebnisse aus Tabelle 11 und 12, wirkt sich die Reduzierung der zu integrierenden Signale auf einen Bereich stärker aus, als im eindimensionalen Fall, da alle stark überlappenden Signale auf und in der Nähe der Symmetriediagonalen nicht in das Ergebnis in Tabelle 12 eingeflossen sind. Qualitativ ergibt sich für die beiden Fälle jedoch das gleiche Bild.

*Tabelle 12: Ergebnisse der Integration von zweidimensionalen simulierten ^{1}H-^{1}H-NOESY-Spektren (PfTrx) bei verschiedenen Auflösungen sowohl mit als auch ohne Rauschen mit dem **eingeschränkten Bereich** der zu integrierenden Signalen von 5,891 ppm bis -0,445 ppm (w1) und 10,503 ppm bis 6,168 (w2). Fett hervorgehobene Werte stellen dabei die schlechtesten Ergebnisse dar und grau hinterlegt die besten.*

Spektrum	Position am Extremum	Anzahl Signale an der selben Position	Methode 1: verschiebe keine Signale		Methode 2: verschiebe Signal mit geringster Distanz zum Extremum		Methode 3: verschiebe alle Signale an deren nächsten Extremum		Methode Version 2009		Anzahl der integrierten Signale
			Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	Mittelwert	Median	
2kx1k	j	1	0,283	0,152	0,286	0,152	0,268	0,141	0,43	**0,205**	270
2kx1k	j	2	0,361	0,173	0,374	0,173	0,314	0,134	2,225	**1,612**	18
2kx1k	n	1	1,413	0,776	1,277	0,559	4,973	0,371	NA	**NA**	163
2kx1k (Rausch)	j	1	0,29	0,166	0,293	0,166	0,266	0,159	0,418	**0,203**	268
2kx1k (Rausch)	j	2	0,389	0,22	0,398	0,191	0,319	0,14	2,223	**1,616**	18
2kx1k (Rausch)	n	1	1,41	0,804	1,294	0,563	4,915	0,352	NA	**NA**	165
2kx2k	j	1	0,271	0,132	0,272	0,132	0,257	0,123	0,315	**0,201**	248
2kx2k	j	2	0,39	0,237	0,399	0,237	0,291	0,155	1,967	**1,479**	20
2kx2k	n	1	1,227	0,654	1,049	0,379	11,477	0,363	NA	**NA**	187
2kx2k (Rausch)	j	1	0,321	0,149	0,321	0,149	0,306	0,143	0,313	**0,197**	245
2kx2k (Rausch)	j	2	0,433	0,284	0,438	0,28	0,336	0,156	2,09	**1,52**	18
2kx2k (Rausch)	n	1	1,479	0,683	1,092	0,403	11,364	0,389	NA	**NA**	190
4kx4k	j	1	0,247	0,137	0,253	0,135	0,241	0,132	0,309	**0,242**	225
4kx4k	j	2	0,234	0,175	0,234	0,175	0,23	0,136	1,358	**1,124**	12
4kx4k	n	1	0,955	0,54	0,896	0,39	29,53	0,357	NA	**NA**	210

Alle Volumen der Signale, deren Position an einem Extremum anzutreffen waren, konnten mit der Methode 3 am genauesten (grau hinterlegt) berechnet werden. Die ursprüngliche Methode schnitt hier schlechter (fett hervorgehoben) ab. Bei Positionen von zwei Signalen,

deren Position am gleichen Extremum lagen, waren die Werte der ursprünglichen Version sehr schlecht. Dies wurde bereits in 3.2.3 erläutert, da die frühere Version keine Mehrfachzuordnungen integrieren konnte. Somit beschränkt sich der Vergleich für diesen Typ im Folgenden auf die Methoden 1 bis 3.

Methode 3 lieferte für Signale im eingeschränkten Bereich (Tabelle 12) für alle Auflösungen und Typen (Tabellenzeilen) die besten Volumenverhältnisse. Jedoch sind die Abweichungen der Volumen über das gesamte Spektrum aus Tabelle 11 im Falle der Positionen abseits der Extrema zur Tabelle 12 bei der Verwendung der Methode 2 am geringsten und somit am besten.

3.2.6.3 Ergebnisse der Integrationen eines dreidimensionalen simulierten Spektrums des Proteins *PfTrx*

Tabelle 13: Ergebnisse der Integration eines dreidimensionalen simulierten NOESY-HSQC-Spektrums (PfTrx) sowohl mit als auch ohne Rauschen. Fett hervorgehobene Werte stellen dabei die schlechtesten Ergebnisse dar und grau hinterlegt die besten.

Spektrum	Posi-tion am Extre-mum	Anzahl Signale an der selben Position	Methode 1: verschiebe keine Signale		Methode 2: verschiebe Signal mit geringster Distanz zum Extremum		Methode 3: verschiebe alle Signale an deren nächsten Extremum		Methode Version 2009		Anzahl der inte-grierten Signale
			Mittel-w ert	Median	Mittel-w ert	Median	Mittel-w ert	Median	Mittel-w ert	Median	
512x512x 128	j	1	0,34	0,239	0,34	0,239	0,338	0,23	1,368	**0,51**	1314
	j	2	0,452	0,35	0,452	0,35	0,404	0,329	5,093	**2,549**	212
	n	1	1,665	0,74	1,636	0,666	1,584	0,468	NA	**NA**	338
512x512x 128 (Rausch)	j	1	0,342	0,249	0,342	0,249	0,331	0,237	1,368	**0,518**	1312
	j	2	0,466	0,353	0,466	0,353	0,419	0,334	5,111	**2,527**	212
	n	1	1,617	0,572	1,617	0,572	1,595	0,471	NA	**NA**	340

Da bei dreidimensionalen NOESY-HSQC-Spektrum durch die Frequenzdomäne von Stickstoff keine Symmetrieachse mehr zu den Wasserstoffdomänen existiert, wurde auf eine Eingrenzung der zu integrierenden Signale verzichtet. Hier war Methode 3 bei allen Typen die beste (grau hinterlegt) und die ursprüngliche Methode die schwächste (fett hervorgehoben) (siehe Tabelle 13).

Alle Signale, welche ihre Position am Extremum aufwiesen, wurde von allen Methoden integriert. Lediglich die ursprüngliche Methode lieferte in Gegensatz zu den Methoden 1 bis 3 erneut bei Positionen abseits der jeweiligen Extrema ein Volumen von 0.

3.2.6.4 Zusammenfassung der Ergebnisse der Integrationen

Aus den Abweichungen der Volumen aller Methoden vom theoretischen Volumen konnte die Methode 3 als die beste Methode zur Integration von Volumen festgelegt werden. Um die Verbesserung des neuen Integrations-Algorithmus gegenüber der ursprünglichen Version darzustellen, wurde für jedes Spektrum der Median der Volumen-Verhältnisse aufgeführt. Da die ursprüngliche Version weder mehrere Signale an einer Position noch Positionen abseits des Extremums berücksichtigen kann, wurden nachfolgend nur die Ergebnisse des Typs „nur ein Peak an einem Maximum“ aus Tabelle 11 in Abb. 20 aufgeführt.

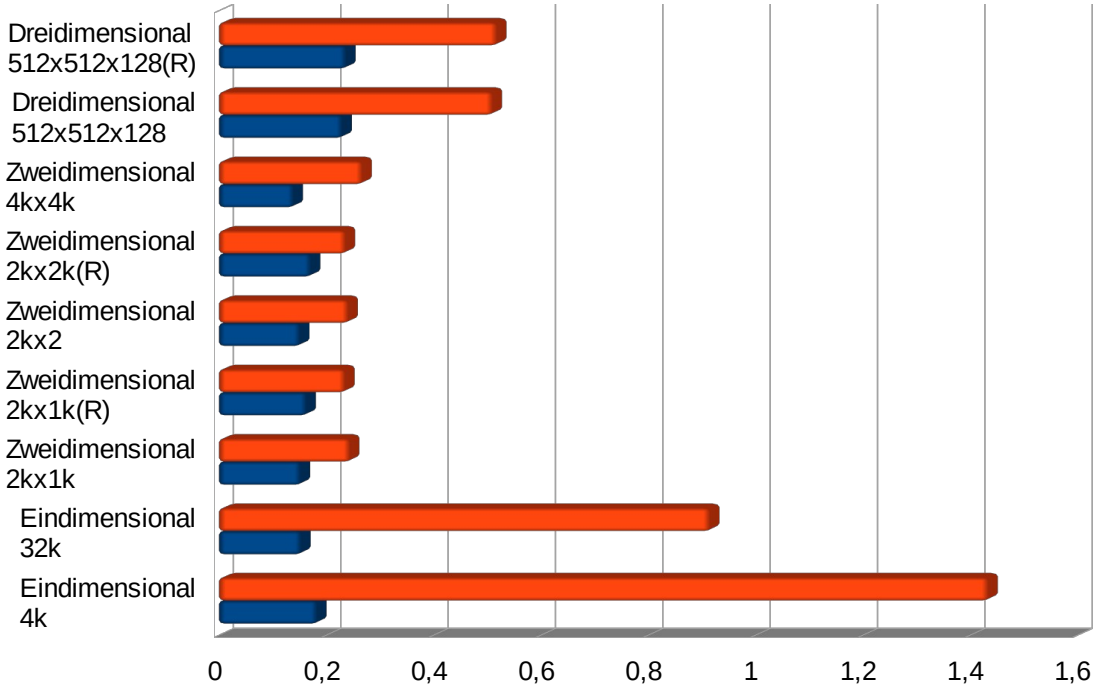

Abb. 20: Vergleich des Median der Verhältnisse von integrierten Volumen und theoretischen Volumen aller Spektren, welche mit der ursprünglichen Integrationsmethode (rot) und mit der Methode 3 (blau) integriert wurden. Für alle Signale (alleinige Position am Extremum) in den jeweiligen Spektren wurde ***keine Einschränkung des Bereiches der zu integrierenden Signale*** *vorgenommen.*

Da für die dreidimensionalen Spektren kein Bereich zur Reduzierung der zu integrierenden Signale festgelegt wurde, sind diese in Abb. 21 nicht enthalten. Es flossen wiederum lediglich die Ergebnisse des Typs „nur ein Peak an einem Extremum“ aus Tabelle 10 und 12 in Abb. 21 ein.

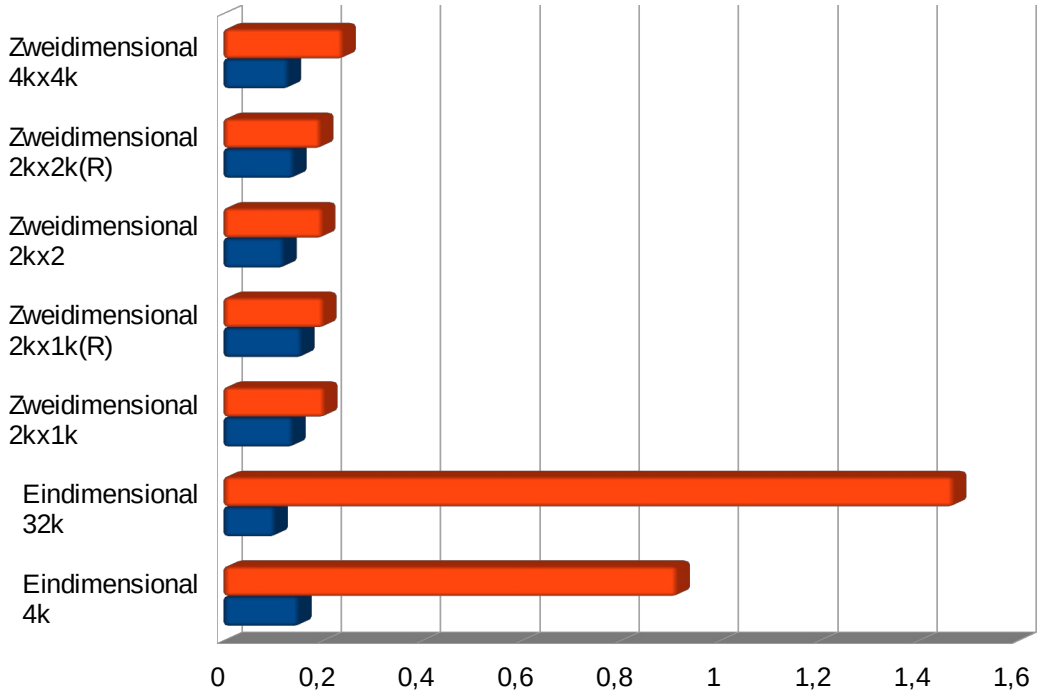

Abb. 21: Vergleich des Median der Verhältnisse von integrierten Volumen und theoretischen Volumen der Spektren, die mit der ursprünglichen Integrationsmethode (rot) und mit der Methode 3 (blau) integriert wurden. Für alle Signale (alleinige Position am Extremum) in den jeweiligen Spektren gilt eine ***Einschränkung des Bereiches*** *der zu integrierenden Signale.*

3.3 Das Modul Schwerpunktbestimmung zur Verbesserung der Positionsbestimmung von NMR-Signalen

Die nach der Integration gespeicherten Volumeninformationen im *Integrations-Hash* dienen als Grundlage für die Berechnung des Schwerpunkts eines Signals basierend auf dessen Volumenform auf dem digitalen Raster. Dazu können alle Intensitäten einer gewünschten Segmentierung, welche einen Beitrag zu dem jeweiligen Signal-Volumen liefern, sehr schnell und ohne Neuberechnung des Volumens aus diesem *Integrations-Hash* extrahiert und bereitgestellt werden (siehe vorangegangenes Kapitel).

3.3.1 Die Berechnung des physikalischen Massenschwerpunkts

Zur Berechnung des Schwerpunkts muss eine gewünschte Segmentierungstiefe festgelegt werden, welche die Größe der Grundfläche des Volumens festlegt. Alle Einzelintensitäten, welche innerhalb dieser Grundfläche liegen gehen dann in Formel (6) als sog. Massenteile ein.

Analog zur Formel (6) zur Bestimmung des Schwerpunkts kann dieser auch mittels der sog. „reduzierten Intensitäten“ durch Formel (7) bestimmt werden. Dazu werden die Intensitäten an jeder Position um den Anteil, welcher unterhalb der Segmentierungsschwelle in Richtung der Intensitätsachse liegt, reduziert (siehe Abb. 23).

In beiden Fällen werden die (reduzierten oder nicht reduzierten) Intensitäten des digitalen Rasters auf die jeweiligen Achsen der Frequenzdomänen projiziert und entsprechen in eindimensionalen Spektren physikalischen Gewichten (Intensitäten) auf einer Stange. Im zweidimensionalen Fall werden die Intensitäten auf eine Fläche und im dreidimensionalen Fall auf einen Kubus projiziert.

Da dieser Schwerpunkt als Gleitkommazahl vorliegt, kann die Position mit sehr geringen Rundungsfehler in die Einheit ppm umgerechnet werden.

In Abb. 22 und Abb. 23 ist das jeweilige Verfahren zur Berechnung der einfließenden Intensitäten aus dem Signalvolumen mit und ohne Reduzierung der Intensitäten durch die Segmentierung dargestellt.

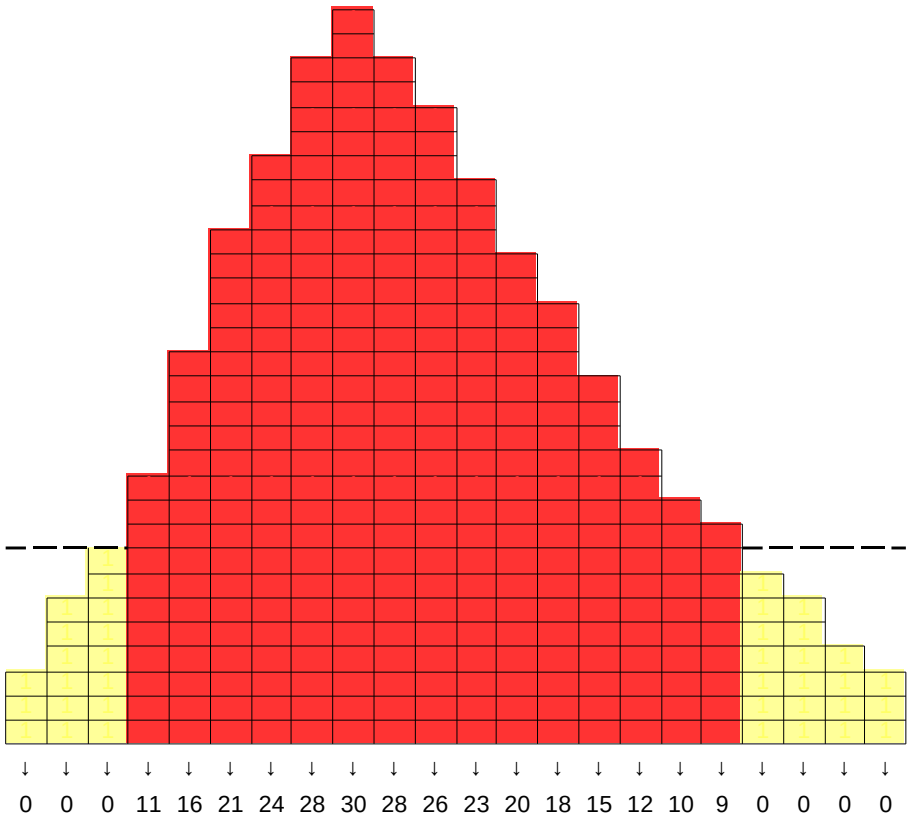

Abb. 22: Schematische Darstellung des Signalvolumens nach Integration. Die ursprüngliche Integration bestimmt durch die Segmentierung ausschließlich die Volumengrundfläche und verändert die anteiligen Intensitäten (rot) zum Volumen nicht. Mit der Segmentierungstiefe (gestrichelte Linie) wird somit lediglich der Bereich festgelegt, welcher Intensitäten vom Volumen ausschließt (gelb) und in das Volumen aufnimmt (rot).

Die Annahme ist nun, dass der Schwerpunkt durch die Verwendung der Formeln (6) und (7) näher an der theoretisch korrekten Koordinate liegt, als die Position am Extremum der Signalform.

Dies erscheint sinnvoll, wenn man Abb. 22 und Abb. 23 näher betrachtet, da sich das Volumen in diesen Abbildungen nicht symmetrisch um die Position des Extremums ausbreitet. Daher erwartet man, dass die Signalform eine bessere Quantifizierung für die Position eines Signals darstellt, da zusätzliche Informationen über das Signal in die Festlegung der Position eines Signals eingeht und z. B. robuster gegen Rauscheinflüsse, schlechter Baseline und Überlappungen ist.

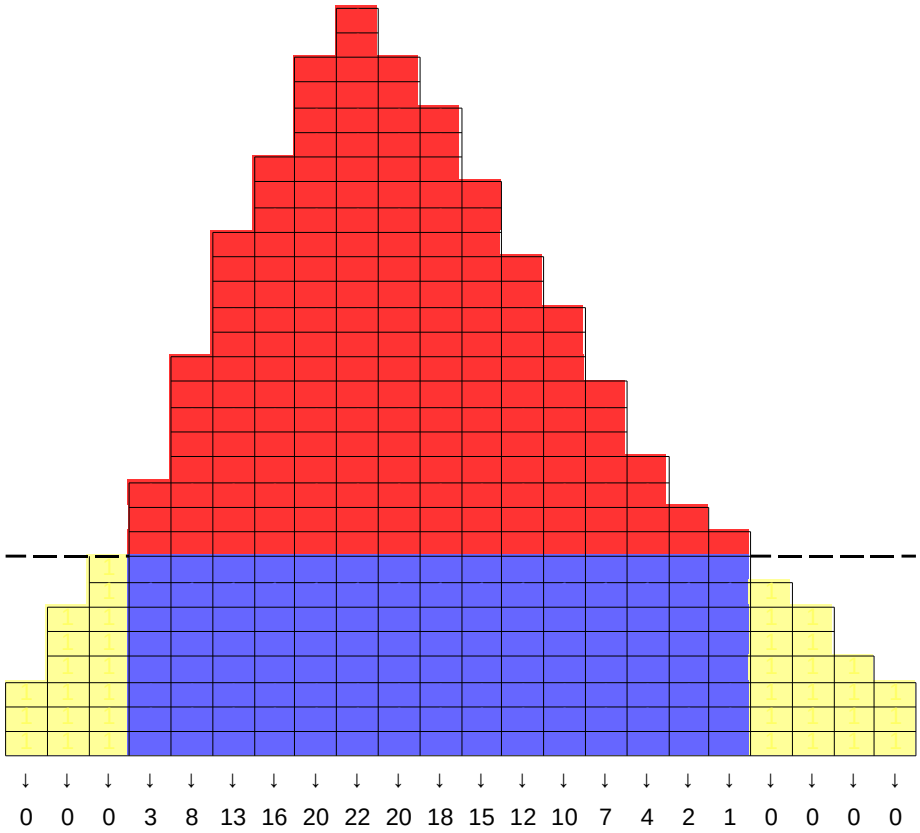

Abb. 23: Schematische Darstellung des Signalvolumens nach Integration bei Reduzierung der Intensitäten. Die verbesserte Integration bestimmt durch die Segmentierung zum einem die Volumengrundfläche und reduziert zum anderen die resultierenden Intensitäten (rot) durch den Intensitätswert unterhalb der Segmentierung (blau). Mit der Segmentierungstiefe (gestrichelte Linie) wird somit analog zu Abb. 22 der Bereich festgelegt, welcher Intensitäten vom Volumen ausschließt (gelb) und in das Volumen als reduzierte Intensitäten aufnimmt (rot).

Ein Anwendungsbeispiel zur Berechnung des Schwerpunkts für die Positionen der Signale 166 und 167 aus Abb. 24 soll exemplarisch ein Ergebnis dieser Methode aufzeigen.

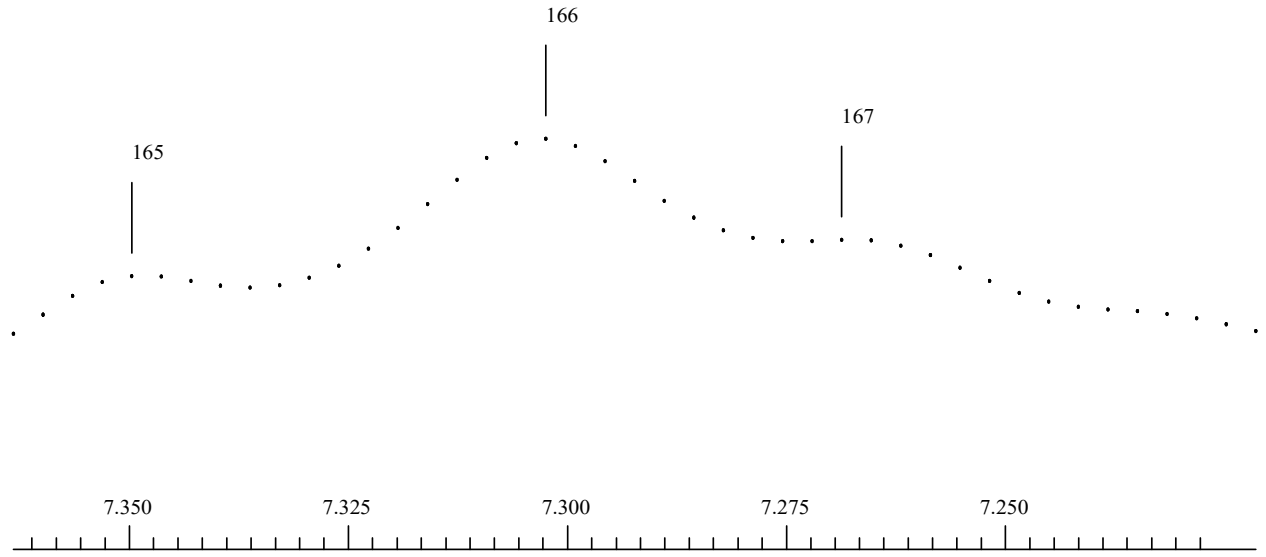

*Abb. 24: Vergrößerung des Bereichs eines simulierten eindimensionalen Spektrums mit einer Auflösung von 4k. Das Spektrum wurde mit dem Modul ‚Pick Peaks' aus AUREMOL gepickt. Zur Veranschaulichung der Berechnung des Schwerpunkts werden die Signale **166** und **167** betrachtet.*

Die Positionen der Extrema der Signalform bezüglich des digitalen Rasters sind die folgenden:

- Signal 166: 7,302 ppm mit der SI-Position bei 1485
- Signal 167: 7,269 ppm mit der SI-Position bei 1495

In Abb. 25 ist das berechnete Volumen dieser beiden Signale und die dazugehörige Position des Schwerpunkts und des Maximums dargestellt.

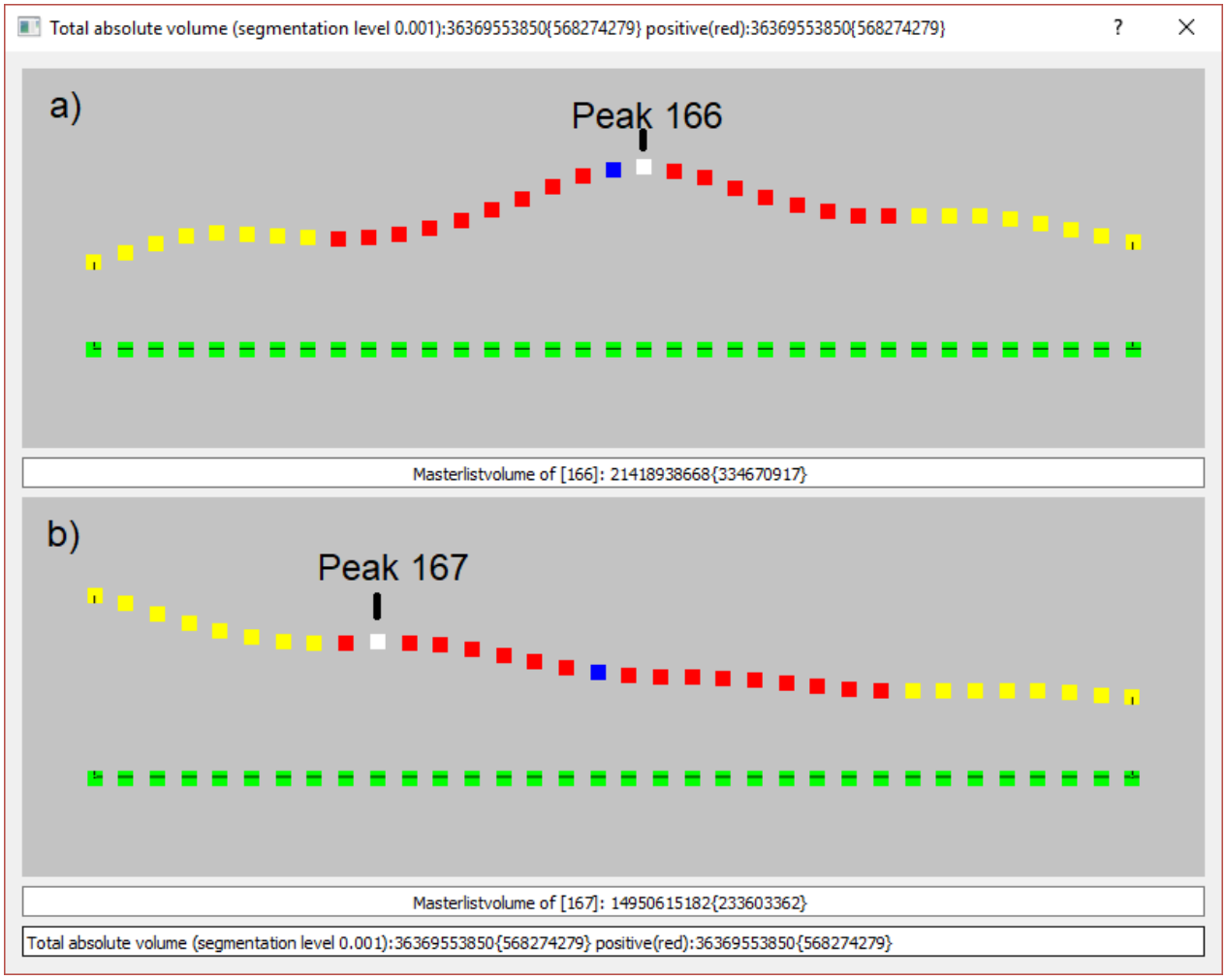

Abb. 25: Screenshot der Volumenplots der beiden Signale 166 und 167 mittels der in dieser Arbeit eingeführten Visualisierung von Signalvolumen. Die Grafiken a) und b) zeigen das Volumen (rot) der jeweiligen Signale. Die Positionen der Extrema (weiß) und der Schwerpunkte (blau) sind auf dem digitalen Raster eingefärbt. Die Nulllinie ist hier grün visualisiert. Die Berechnung des Schwerpunkts erfolgte durch die ***Schwerpunktbestimmung mit Abschneidung an der Segmentierungstiefe*** *bei 0,001.*

Im Falle von Signal „Peak 166“ weist die Position des Schwerpunkts bei 7,305 ppm nur eine geringe Abweichung zur Position am Extremum bei 7,302 ppm auf, da das Volumen hinreichend gleich um das Extremum verteilt ist. Jedoch fällt die Abweichung der Position des Schwerpunkts bei 7,246 ppm von „Signal 167“ zur Position des Extremums bei 7,269 ppm höher aus, da ein Großteil dessen Volumens in das des Nachbarsignals „Peak 166“ durch die Überlagerung eingeflossen ist.

Im folgenden wird gezeigt, wie sich die Lage des Schwerpunkts zu der theoretischen Position verhält, wenn sich die digitale Auflösung verändert.

3.3.2 Die Abhängigkeit der Position des Schwerpunkts von der digitalen Auflösung des Spektrums

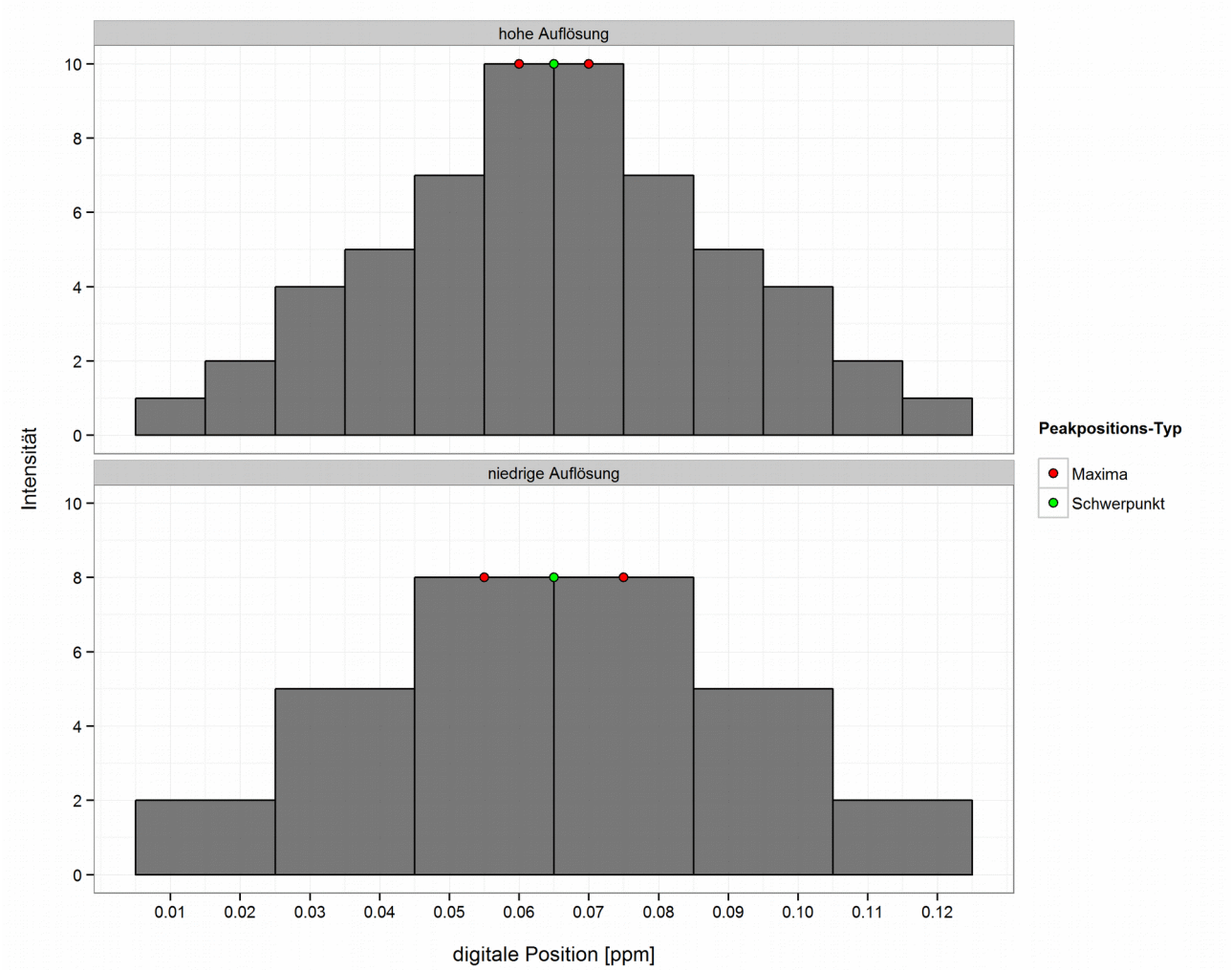

Abb. 26: Vergleich der Signalposition am Extremum (rot) und der Schwerpunktposition (grün) bei einem theoretischen Signal bei 0,065 ppm bei hoher Auflösung (oben) bzw. halbierter Auflösung (unten).

Abb. 26 soll im Fall einer symmetrischen Volumenausbreitung verdeutlichen, wie sich die Position des Schwerpunkts im Vergleich zur Position des Maximums verhält, wenn sich die digitale Auflösung verändert.

Die theoretisch ermittelte Position des Signals liegt in beiden Fällen bei 0,065 ppm. Da in beiden Beispielen das Extremum der Signalform ein Plateau aus zwei Pixeln (also zwei gleich hohe Intensitäten) bildet, kann jede dieser beiden Intensitäten als Extremum in Frage kommen. Vergleicht man nun eines der beiden Extrema (rot) in der jeweiligen

Grafik, kann man erkennen, dass in beiden Fällen die Schwerpunktposition (grün) die theoretische Signalposition bei 0,065 ppm trifft. Auch zeigt sich, dass mit der Halbierung der digitalen Auflösung des Spektrums eine signifikante Abweichung des Maximums bezüglich der theoretischen Position einhergeht.

In diesem Modellbeispiel eines einfachen Signals kann man erkennen, dass die Schwerpunktmethode den theoretischen Wert exakt trifft.

3.3.3 Der Aufbau, die Durchführung und die Auswahl der besten Methode zur Bestimmung der Positionen von NMR-Signalen

In diesem Kapitel sollen drei Methoden zur Festlegung der Position eines NMR-Signals untersucht werden:

1. Berechnung des Schwerpunkts als Position für ein NMR-Signal über das Volumen bei einer festgelegten Segmentierungstiefe (siehe Abb. 22). Hier wird die Intensität eines Signals lediglich dazu benutzt, die Volumengrundfläche zu definieren. Nur die durch diesen (Integrations-)Bereich umfassten Intensitäten gehen unangetastet in Formel (6) ein.

   Diese Methode wird im weiteren Verlauf dieser Arbeit ***Schwerpunktbildung ohne Abschneidung*** genannt.

2. Berechnung des Schwerpunkts als Position für ein NMR-Signal über das Volumen bei einer festgelegten Segmentierungstiefe und zusätzlicher Reduzierung der beteiligten Intensitäten am Volumen durch die Segmentierungstiefe (siehe Abb. 23). Hier werden zusätzlich zur vorherigen Methode 1 die Intensitäten um die angegebene Segmentierungstiefe in Formel (7) reduziert.

   Diese Methode wird im weiteren Verlauf dieser Arbeit als ***Schwerpunktbildung mit Abschneidung*** bezeichnet.

3. Bestimmung der Position über das Extremum einer Signalform am digitalen Raster.

   Diese Methode wird im weiteren Verlauf dieser Arbeit als ***Maximum-Methode*** bezeichnet.

Um diese Methoden an verschiedenen Konstellationen von Signalpositionen zueinander bewerten zu können, wurden zwei Signale aus einem simulierten Spektrum ausgewählt.

Danach wurde ein Signal auf das andere zubewegt, bis beide die gleiche Position aufwiesen. Dabei wurde bei der Erstellung der Simulation darauf geachtet, dass keine Signale existieren, welche auf der Bewegungsbahn liegen. Es wurde nach jedem Bewegungsschritt die Entfernung zwischen den beiden Signalen gegen die Abweichung der Position des Massenschwerpunkts und des üblichen Extremums zur theoretischen Position der Signale untersucht.

Um einen aussagekräftigen Verlauf wiederzugeben, wurden 160 Bewegungsschritte für einen Durchlauf vollzogen und dazu jeweils der Schwerpunkt und das zugehörige Extremum bestimmt. Hierzu wurde im Falle eines eindimensionalen Spektrums „Peak 2“ von der Position bei 4,1513 ppm zur Position von „Peak 1“ bei 3,7830 ppm in 160 Schritten bewegt (Abb. 27).

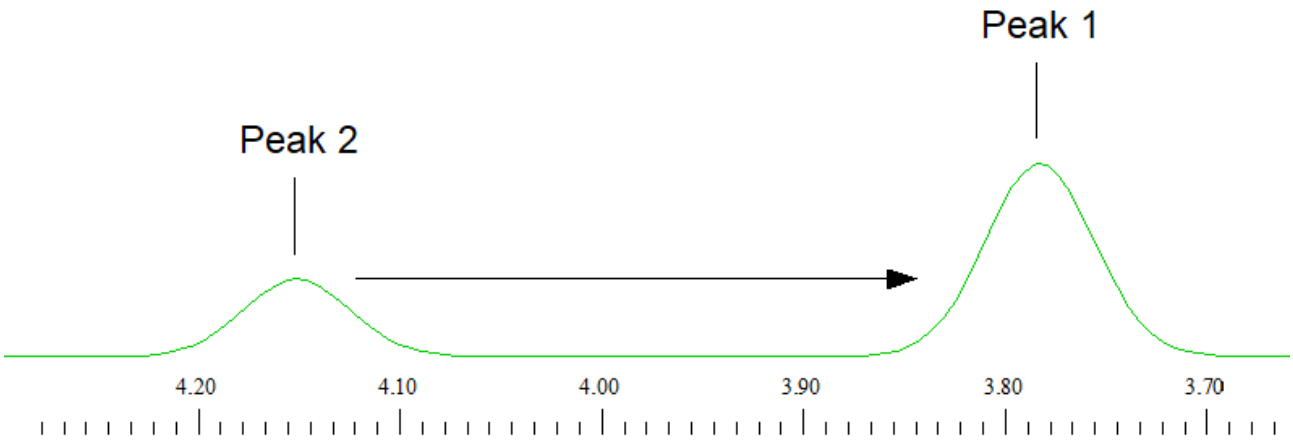

Abb. 27: Verlauf der Bewegung während der 160 Bewegungs-Iterationen. In jeder Iteration wird der Schwerpunkt der Signale „Peak 1“ und „Peak 2“ berechnet und deren jeweilige Abweichung zur theoretisch ermittelten Referenzposition ermittelt.

Bei den zweidimensionalen Spektren werden zwei unterschiedliche Bewegungsrichtungen betrachtet:

1. Ein Signal wurde aus der Position 4,263 ppm(w1)/9,000 ppm(w2) in 160 Iterationen **diagonal** auf die Position des anderen Signals bei 3,783 ppm/9,480 ppm zubewegt.
2. Ein Signal wurde aus der Position 3,783 ppm/9,000 ppm in 160 Iterationen **horizontal** entlang der Frequenzachse der direkten Dimension auf die Position des anderen Signals bei 3,783 ppm/9,480 ppm zubewegt.

3.3.4 Die Erfassung der Bewegungsiterationen in den Bewegungsgraphen

Zu allen Spektren wurde ein Datenpool für die Volumeninformationen (*Integrations-Hash*) generiert, in dem alle zu einem Signal zugehörigen Intensitäten zum Volumen gespeichert wurden. Aus diesem können die Volumen und die relevanten Intensitäten für die Schwerpunktbildung abgerufen werden. Dabei wurden für die Volumina der Signale vier verschiedene Segmentierungstiefen zur Berechnung des Schwerpunkts getestet.

Der Verlauf der Kurven aus aus Abb. 28 und Abb. 29 stellt die Distanzen zwischen „Peak 1“ und „Peak 2“ pro Bewegungsschritt zur Abweichung der theoretischen Positionen (y-Achse) dar. Der gesamte Weg, den das Signal „Peak 2„ zu „Peak 1“ zurücklegt ist somit entlang der x-Achse in 160 Schritten aufgetragen. Dabei ist der Schritt bei 0 die letzte Bewegungsiteration, bei dem beide Signale dieselbe Position einnehmen. **Somit gilt, dass je niedriger der y-Wert ist, desto besser trifft die entsprechende Positionsbestimmung die Referenzposition, da die Abweichung von der theoretischen Position am Geringsten ist.**

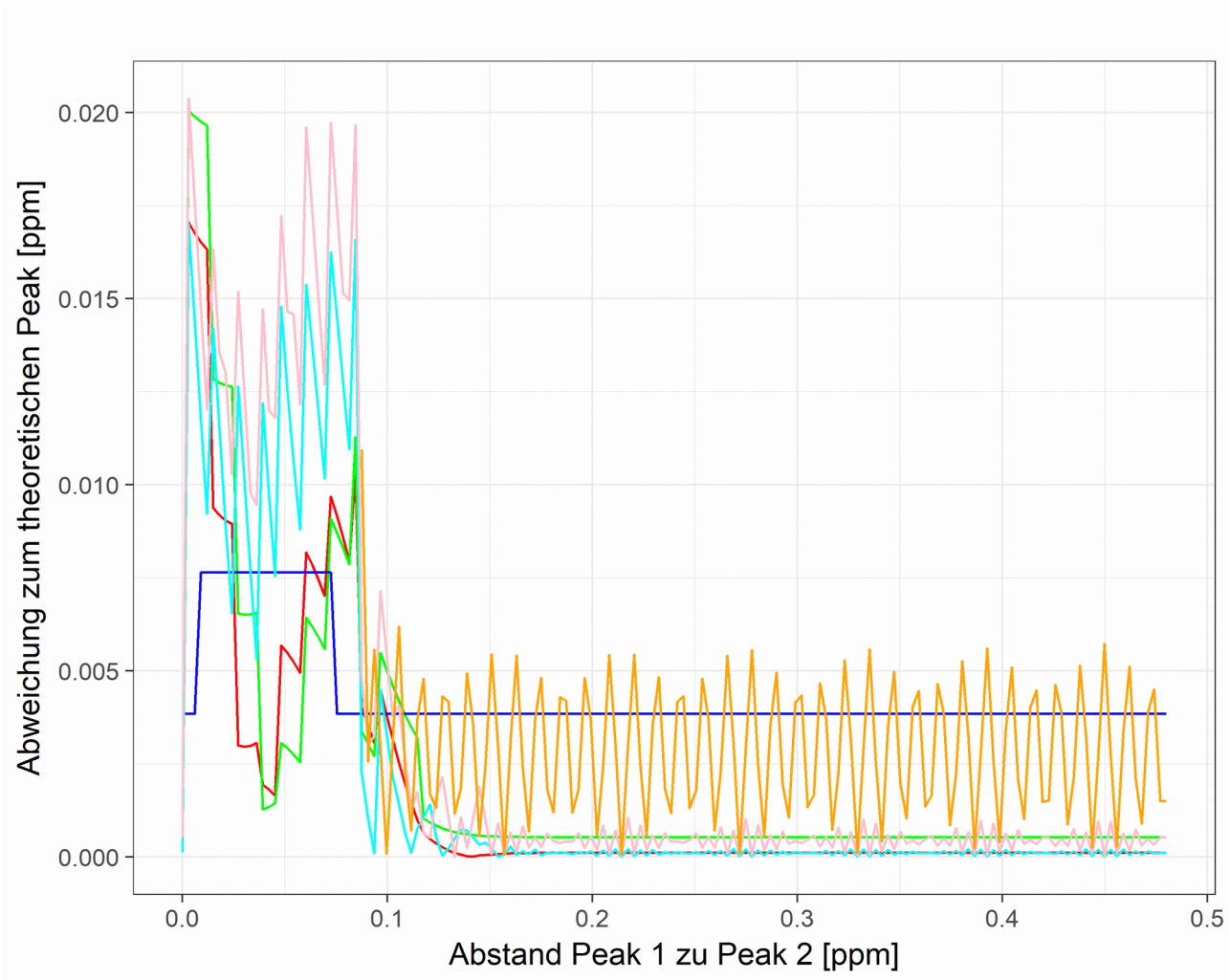

Abb. 28: Abweichungen der Signalpositionen von „Peak 1“ und „Peak 2“ zu deren theoretischen Referenzpositionen bei jeder der 160 Iterationen. Alle für diesen Datensatz verwendeten ***eindimensionalen*** *Spektren weisen eine Auflösung von* ***256*** *Pixeln auf. Die Volumen wurden mit einer Segmentierung von 0,2 bestimmt. „Peak 1“ stellt das Signal dar, welches sich nicht bewegt und „Peak 2“ das sich entlang der Horizontalen bewegende. „Peak 1“ weist zudem ein höheres Volumen auf als „Peak 2“. Die Abstände der berechneten Signale zu den Referenzpositionen in der ppm-Skala während der Annäherung sind für jede Methode der Positionsbestimmung dargestellt. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

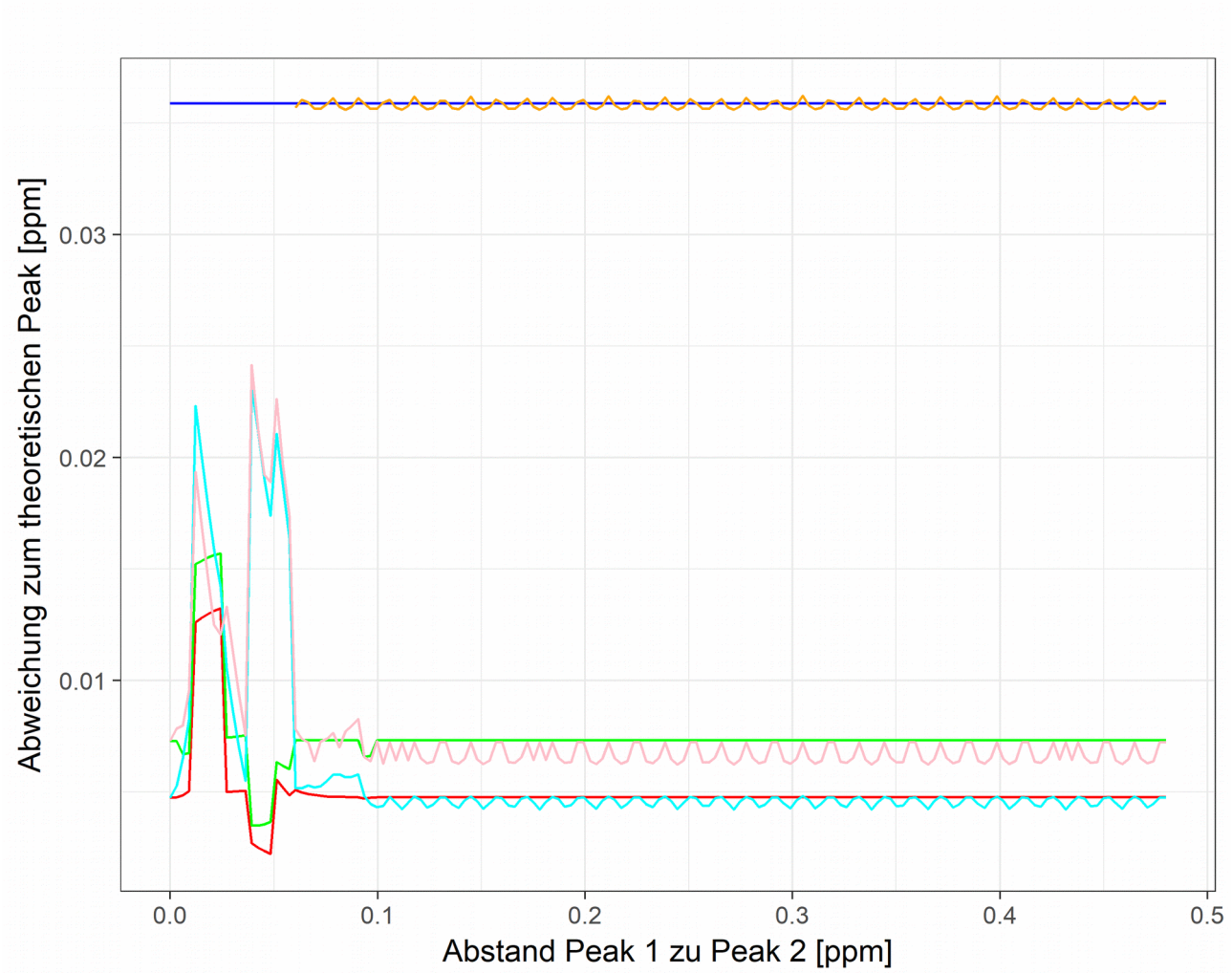

Abb. 29: Abweichungen der Signalpositionen von „Peak 1“ und „Peak 2“ zu deren theoretischen Referenzpositionen bei jeder der 160 Iterationen. Alle für diesen Datensatz verwendeten ***zweidimensionalen*** *Spektren weisen eine Auflösung von* ***64x512*** *Pixeln auf. Die Volumen wurden mit einer Segmentierung von* ***0,1*** *bestimmt. „****Peak 1“*** *stellt das Signal dar, welches sich* ***nicht bewegt*** *und „****Peak 2“*** *das sich entlang der* ***Horizontalen bewegende****. „Peak 1“ weist zudem ein höheres Volumen auf als „Peak 2“. Die Abstände der berechneten Signale zu den Referenzpositionen in der ppm-Skala während der Annäherung sind für jede Methode der Positionsbestimmung dargestellt. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Erläuterungen zum Bewegungsgraphen aus Abb. 28 und Abb. 29:

- Der Verlauf der Kurven bei der die Positionen durch die *Maximum-Methode* von „Peak 1“ (blau) und „Peak 2“ (orange) sind, weisen nach jeder Iteration eine hohe Abweichung zu der jeweiligen theoretischen Referenzposition auf.

- Der Verlauf der Kurven bei der die Positionen durch die *Schwerpunktmethode ohne Abschneidung* von „Peak 1“ (grün) und „Peak 2“ (rosa) ermittelt wurden, weisen eine signifikant geringere Abweichung zur jeweiligen Referenzposition entlang der Bewegung aus.
- Die geringste Abweichungen zu den Referenzpositionen weisen in diesem Fall die Berechnungen mittels der *Schwerpunktmethode mit Abschneidung* von „Peak 1“ (rot) und „Peak 2“ (cyan) auf.

Auffällig in Abb. 28 und Abb. 29 sind die Wellenmuster, welche die Kurve des sich bewegenden Signals „Peak 2“ bei allen Methoden aufweist. Der Grund hierfür ist, dass die theoretisch berechnete Referenzposition nicht vom digitalen Raster abhängig ist. Da die Berechnung des Schwerpunkts jedoch auf Grundlage der diskreten Intensitäten berechnet wird, fließen nur sehr geringfügige Änderungen in die Berechnung ein, solange sich die Referenzposition während der Bewegung nicht aus dem betroffenen Pixel verändert und eine signifikante Repositionierung der Signalform am digitalen Raster induziert.

Zusammengefasst gilt also für den *Bewegungsgraphen*:

- Jeder *Bewegungsgraph* gibt das Durchlaufen aller 160 Iterationsschritte wieder (siehe Abb. 29).
- Die Parameter der verwendeten Spektren bleiben für alle Iterationen gleich. Die Spektren unterscheiden sich lediglich in den Positionen der beiden Signale.
- Es sind stets die Kurven für die *Schwerpunktmethode mit Abschneidung* von „Peak 1“ (rot) und „Peak 2“ (cyan) enthalten.
- Es sind stets die Kurven für die *Schwerpunktmethode ohne Abschneidung* von „Peak 1“ (grün) und „Peak 2“ (rosa) enthalten.
- Es ist stets die Kurve durch die *Maximum-Methode* von „Peak 1“ (blau) enthalten. Jedoch unterbricht die Kurve von „Peak 2“ (orange), sobald sich durch die Überlagerung beider Signale dessen Positionen nicht mehr am Extremum der Signalform befindet.

3.3.5 Die Auftrennung der Bewegungsgraphen

Folgt man dem Verlauf der Kurve der *Maximum-Methode* von „Peak 2“ (orange) aus Abb. 29, so fällt auf, dass die Kurve ab einem bestimmten Abstand zwischen den beiden Signalen verschwindet. Ab hier sind die Signale nicht mehr durch ihre Position am Extremum unterscheidbar. Es lassen sich „Peak 1“ und „Peak 2“ nicht mehr durch das Picken des Extremums separieren und weisen durch Überlappung nur noch ein Extremum auf.

Falls dennoch ein Signal an einer Position, welches **kein Extremum** ist, durch eine Simulation oder durch manuelles Picken in der Peakliste vorhanden ist, kann das Volumen dieses Signals mit der verbesserten Integration wie folgt bestimmt werden:

- Die Integrationsmethode aus 3.2.4.2 (Methode 2 - Integration mit nur einem erlaubten nächsten Extremum)

  Das heißt die Position von „Peak 2“ wird an seiner Stelle belassen und separat segmentiert, da die Position von „Peak 2“ nicht temporär auf „Peak 1“ verschoben werden kann, da diese bereits von „Peak 1“ besetzt ist.

  Die Integrationsmethode aus 3.2.4.1 (Methode 1 – Integration ohne Veränderung der Signalpositionen) könnte für diese Signale genauso angewendet werden, da die Positionen beider Signale nicht verändert werden muss. Dadurch ist das Ergebnis der Integrationsmethode 1 und 2 identisch.

  Die Methode der Schwerpunktbestimmung mittels dieser Integrationsvariante wird in diesem Kapitel als Modul **„getrenntes Volumen“** bezeichnet.

- Die Integrationsmethode aus 3.2.4.3 (Methode 3 - Integration durch temporäre Verschiebung aller Positionen von Signalen an das Extremum einer gemeinsamen Signalform)

  Das heißt die Position von „Peak 2“ wird temporär an die Position von „Peak 1“ versetzt. Beide Signale haben dieselbe Volumeninformation bezüglich der beteiligten Intensitäten.

  Die Methode der Schwerpunktbestimmung mittels dieser Integrationsvariante wird in diesem Kapitel als Modul **„gemeinsames Volumen“** bezeichnet.

Für alle anderen Signale, deren Positionen am Extremum der jeweiligen Signalform sind, wird ausschließlich die Integrationsmethode aus 3.2.4.2 verwendet. Dies wird daher im Folgenden als Modul **„Standardvolumen“** bezeichnet.

In Abb. 30 soll die Anwendung der verschiedenen Integrationsmethoden auf dem Bewegungsgraphen veranschaulicht werden. Der Bereich A stellt im Bewegungsgraphen aus Abb. 30 die Iterationen dar, falls beide Signale nicht überlappen und die Positionen beider Signale je ein Extremum an deren Signalform besitzen.

Der **Bereich C** stellt im Bewegungsgraphen aus Abb. 30 die Iterationen dar, in dem „Peak 1“ mit „Peak 2“ soweit überlappt, dass keine zwei getrennte Extrema mehr auszumachen sind. Dieser Bereich wird hier **nicht mehr** mit dem Modul „Standardvolumen“ ausgewertet, sondern erfolgt mit den Modulen „getrenntes Volumen“ und „gemeinsames Volumen“. Innerhalb des Bereichs B beginnt „Peak 2“ mit „Peak 1“ zu überlappen, dabei weist „Peak 2“ jedoch immer noch ein Extremum auf. Im Bereich C sind beide Signale isoliert und weisen keinerlei Überlappung auf.

Im späteren Verlauf der Arbeit soll untersucht werden, welche der beiden Module („getrenntes Volumen“ oder „gemeinsames Volumen“) für den Fall von nur einem Extremum (also für den Bereich C) die bessere Wahl darstellt. Falls Extrema beider Signale existieren, wird die Schwerpunktbestimmung durch das Modul „Standardvolumen“ mit der Maximum-Methode verglichen.

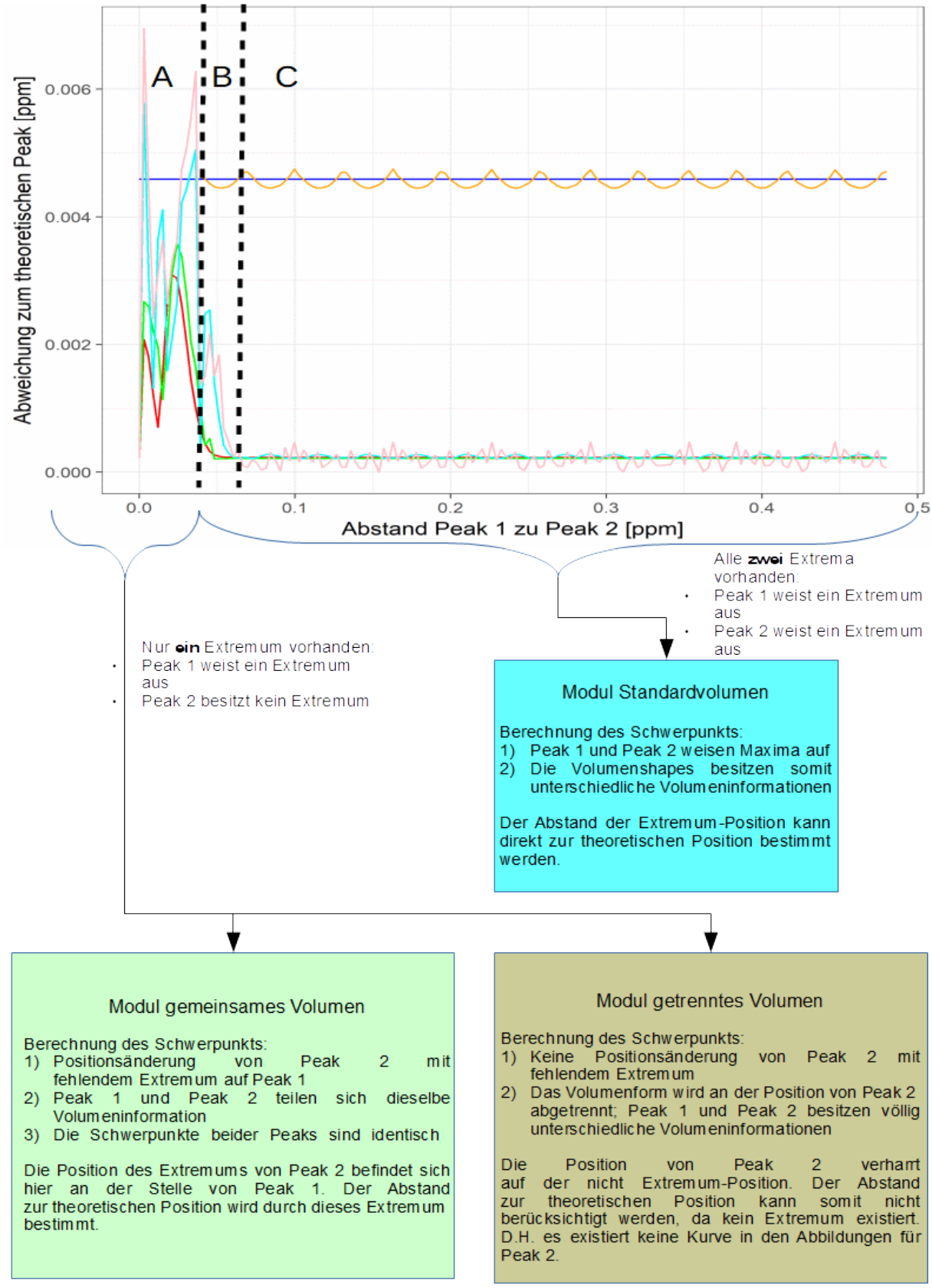

Abb. 30: Unterteilung der Module der Bestimmung der Position eines Signals durch den Schwerpunkt.

Um die Ergebnisse der Methoden vergleichen zu können, wurden verschiedene Variationen der verwendeten Spektren und der Segmentierung zur Volumenbestimmung angewendet und jeweils in einem Bewegungsgraph erfasst und anschließend wie bereits beschrieben (siehe Abb. 30) aufgeteilt:

- Iterationen mit Spektren ohne Rauschen
- Iterationen mit Spektren mit Rauschen
- Iterationen mit Signalen gleichen Volumens in den Spektren
- Iterationen mit Signalen unterschiedlichen Volumens in den Spektren
- Iterationen mit Spektren, deren Signalvolumen mit verschiedenen Segmentierungstiefen [0,0001; 0,1; 0,25; 0,5] bestimmt wurde
- Variation der digitalen Auflösung der Spektren (siehe nachfolgendes Kapitel)

Die Variationen führten zu insgesamt 400 Bewegungsgraphen, wobei 112 davon den eindimensionalen Fall und 288 Bewegungsgraphen die beiden zweidimensionalen Fälle ausmachten.

3.3.6 Definition der gemittelten Bewegungsgraphen als Zusammenfassung der einzelnen Bewegungsgraphen zur Analyse der Abhängigkeit der Positionsbestimmungsmethoden von der Variation der digitalen Auflösung

Da die Analyse der einzelnen Bewegungsgraphen zu unübersichtlich wäre, wurden diese zusammengefasst. Die Ergebnisse zeigen die Mittelwerte der 160 Abweichungen aufgeteilt in die entsprechenden Module („Standardvolumen“, „gemeinsames Volumen“ und „getrenntes Volumen“) zu den theoretischen Signal-Positionen. Diese Mittelwerte wurden (falls vorhanden) für jeden Positionsbestimmungs-Typ (*Maximum-Methode, Schwerpunktbildung ohne Abschneidung, Schwerpunktbildung mit Abschneidung*) zu den jeweiligen Signalen „Peak 1“ und „Peak 2“ bestimmt. Alle jeweiligen y-Werte zu einer Auflösung repräsentieren somit einen Plot aus den gemittelten Abweichungen.

Diese Mittelwerte werden jeweils für folgende Auflösungen des digitalen Rasters ermittelt und aufgetragen:

- 64, 256, 512, 1k, 2k, 4k und 16k für eindimensionale Spektren

- 64x64, 512x64, 64x512, 512x512, 2kx64, 64x2k, 2kx512, 512x2k und 2kx2k für zweidimensionale Spektren

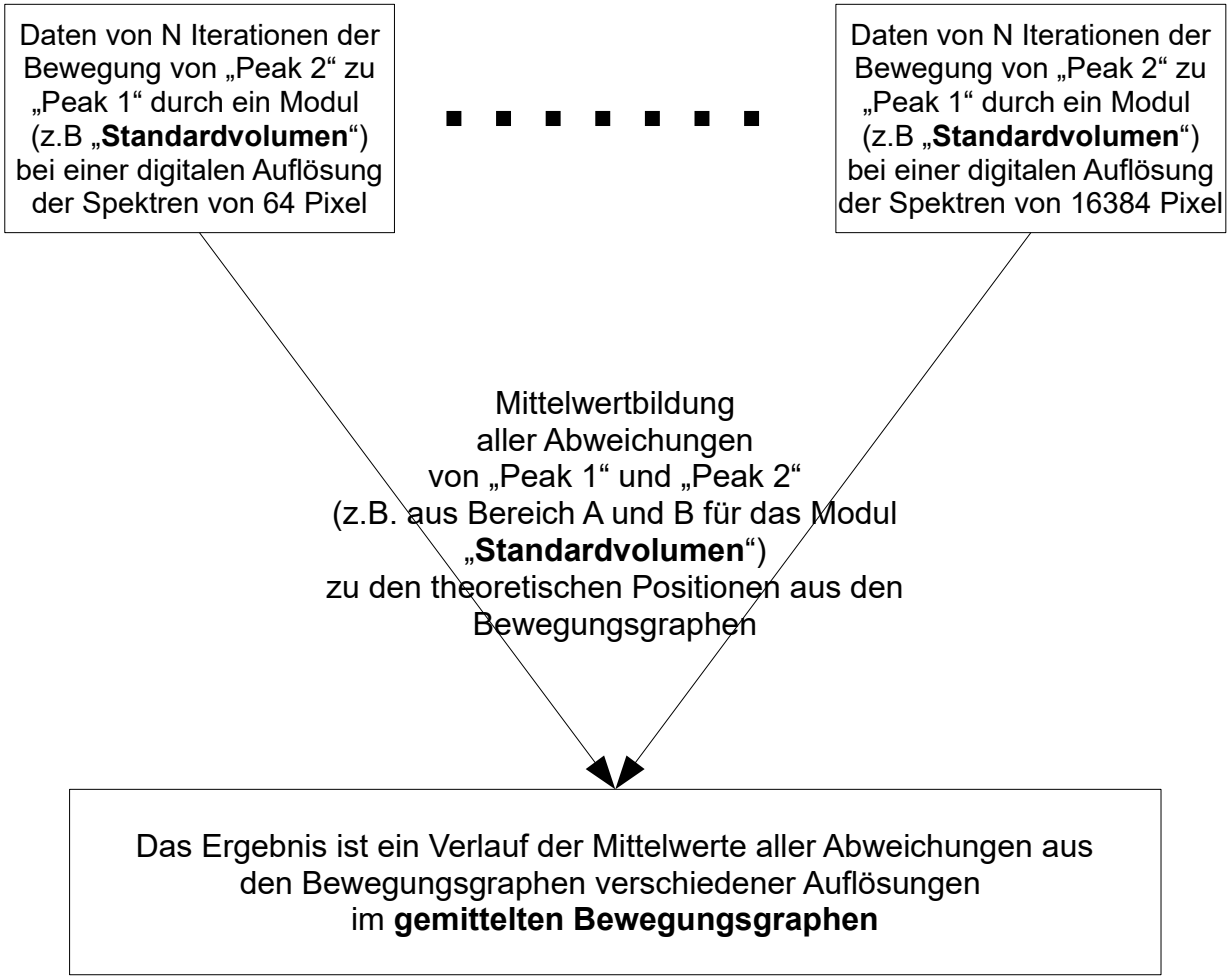

*Abb. 31: Beispiel des Ablaufs zur Zusammenfassung der Daten aus **einem Bewegungsgraphen** von **eindimensionalen** Spektren verschiedener Auflösungen aus einem Modul (hier „**Standardvolumen**“). Dabei werden N Iterationen eines jeden Moduls zu den **gemittelten Bewegungsgraphen** zusammengeführt.*

3.3.7 Ergebnisse des Moduls „Standardvolumen“ - Vergleich der *Schwerpunktbildung mit und ohne Abschneidung* am Segmentierungslevel mit der *Maximum-Methode*

Um einen Vergleich der beiden Schwerpunktmethoden *Schwerpunktbildung ohne Abschneidung* und *Schwerpunktbildung mit Abschneidung* mit der Maximum-Methode durchführen zu können, flossen in diesem Abschnitt nur Daten aus den Spektren ein, bei denen beide Signale „Peak 1“ und „Peak 2“ eine Position am Extremum (während der Bewegung aufeinander zu) der Signalform aufweisen. Daher wird das Modul „Standardvolumen“ verwendet.

Ergebnis des Moduls „Standardvolumen“ bei eindimensionalen Spektren

Die in den folgenden Abbildungen enthaltenen Kurven werden gemäß Abb. 31 durch Bildung der Mittelwerte des Bereiches A und B (siehe Abb. 30) zusammengefasst, so dass dadurch ein Datenpunkt z. B. der grünen Kurve aus Abb. 29 entsteht. Die restlichen Kurven der Positionsbetimmungsmethoden entstehen analog und werden als ***gemittelte Bewegungsgraphen*** bezeichnet.

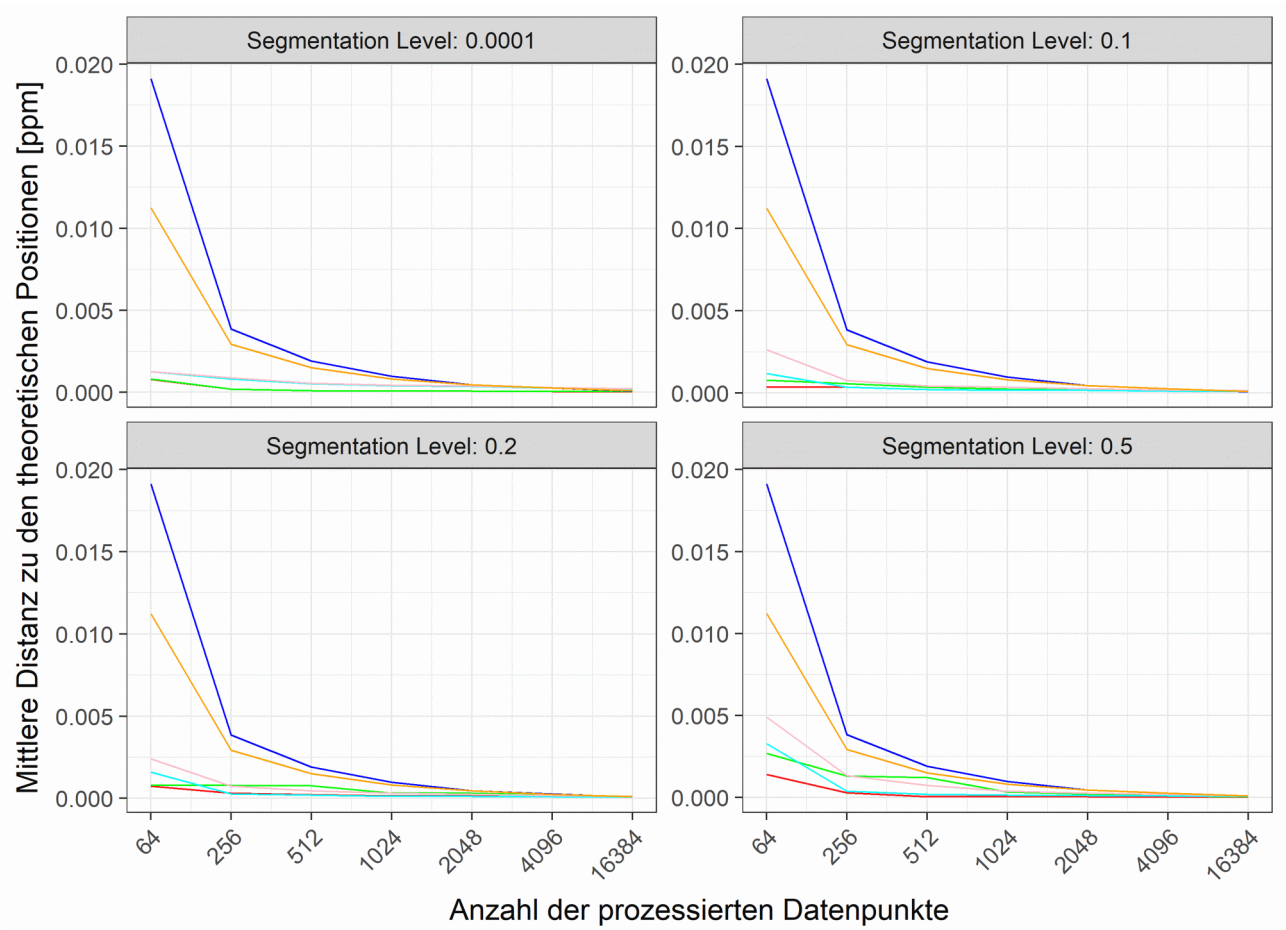

*Abb. 32 : Gemittelter Bewegungsgraph der **eindimensionalen** Spektren bei den digitalen Auflösungen von 64 bis 16384 Pixeln **ohne Rauschen.** Beide Signale weisen dabei **unterschiedliche Volumen** bei den vier Segmentierungstiefen von 0,0001 bis 0,5 und auch eine Position am Extremum der Signalform auf. „Peak 1“ stellt das Signal dar, welches sich nicht bewegt und „Peak 2“ das sich bewegende Signal. Das Signal „Peak 1“ weist zudem ein höheres Volumen auf als „Peak 2“. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Im gemittelten Bewegungsgraphen aus Abb. 32 zeigen sich folgende Aspekte:

- Die Mittelwerte der Abstände des Maximums von „Peak 1“ (blau) und „Peak 2“ (orange) zur Referenzposition verringern sich mit der Erhöhung der Auflösung der Spektren.
- Die mittleren Distanzen zu den theoretischen Positionen aus Abb. 32 , welche durch die „Schwerpunktbildung ohne Abschneidung“ am Segmentierungslevel bestimmt wurden, weisen für die Signale „Peak 1“ (grün) und „Peak 2“ (rosa) einen ähnlichen

Abstand auf, liegen jedoch stets näher an der Referenzposition (also mittlere Distanz 0).

- Ähnlich verhält es sich beim mittleren Abstand der Positionen der Signale „Peak 1“ (rot) und „Peak 2“ (cyan), welche durch die „Schwerpunktbildung mit Abschneidung“ an der Segmentierung bestimmt wurden.
- Aus dem Verlauf der Mittelwerte bezüglich der entsprechenden Auflösung des jeweiligen Spektrums erkennt man, dass sich die Abweichung der Bestimmung der Signalposition durch die Schwerpunktbildung zur theoretischen Position mit abnehmender Auflösung im Vergleich zur Maximum-Methode verbessert.
- Je höher die Auflösung des digitalen Rasters wird, desto mehr tendieren alle drei Methoden zur theoretischen Referenzposition und würden bei unendlich hoher Auflösung mit der theoretischen Position übereinstimmen.

*Tabelle 14: Übersicht der **eindimensionalen** Ergebnisse bei Verwendung von Spektren **verschiedener Auflösungen ohne Rauschen** mit Signalen **unterschiedlichen Volumens**. Die beiden Signale „Peak 1“ und „Peak 2“ weisen jeweils eine Position am Extremum ihrer Signalform auf; zur Schwerpunktbildung wurde das Modul „**Standardvolumen**“ verwendet. Der Schwerpunkt der Signale wurde einmal mit Reduzierung der Intensitäten durch das Segmentierungslevel (Schwerpunktbildung mit Abschneidung) und ohne Reduzierung (Schwerpunktbildung ohne Abschneidung) berechnet.*

Modul „Standardvolumen“ bei Scherpunktbildung **mit** oder **ohne** Abschneidung am Segmentierungs-level und Maximum-Methode	Mittelwerte aller Abstände bei der digitalen Auflösung von 64 Pixeln	Mittelwerte aller Abstände bei der digitalen Auflösung von 256 Pixeln	Mittelwerte aller Abstände bei der digitalen Auflösung von 512 Pixeln	Mittelwerte aller Abstände bei der digitalen Auflösung von 1024 Pixeln	Mittelwerte aller Abstände bei der digitalen Auflösung von 2048 Pixeln	Mittelwerte aller Abstände bei der digitalen Auflösung von 4096 Pixeln	Mittelwerte aller Abstände bei der digitalen Auflösung von 16384 Pixeln	verwendete Sementierungs-tiefe
Peak 1 mit Abschneidung	0,0008	0,0002	0,0001	0,0001	0,0001	0,0000	0,0000	0,0001
Peak 1 ohne Abschneidung	0,0008	0,0002	0,0001	0,0001	0,0001	0,0001	0,0000	0,0001
Peak 2 mit Abschneidung	0,0013	0,0008	0,0005	0,0004	0,0004	0,0003	0,0002	0,0001
Peak 2 ohne Abschneidung	0,0013	0,0009	0,0006	0,0004	0,0004	0,0003	0,0002	0,0001
Peak 1 mit Abschneidung	0,0004	0,0004	0,0002	0,0002	0,0002	0,0001	0,0001	0,1
Peak 1 ohne Abschneidung	0,0008	0,0006	0,0004	0,0002	0,0002	0,0002	0,0001	0,1
Peak 2 mit Abschneidung	0,0012	0,0004	0,0002	0,0002	0,0002	0,0001	0,0001	0,1
Peak 2 ohne Abschneidung	0,0026	0,0008	0,0004	0,0003	0,0003	0,0002	0,0001	0,1
Peak 1 mit Abschneidung	0,0008	0,0003	0,0002	0,0002	0,0001	0,0001	0,0001	0,2
Peak 1 ohne Abschneidung	0,0008	0,0008	0,0008	0,0003	0,0003	0,0002	0,0001	0,2
Peak 2 mit Abschneidung	0,0016	0,0003	0,0002	0,0001	0,0001	0,0001	0,0001	0,2
Peak 2 ohne Abschneidung	0,0024	0,0007	0,0004	0,0003	0,0003	0,0002	0,0001	0,2
Peak 1 mit Abschneidung	0,0014	0,0003	0,0001	0,0001	0,0000	0,0000	0,0000	0,5
Peak 1 ohne Abschneidung	0,0027	0,0013	0,0012	0,0003	0,0002	0,0001	0,0000	0,5
Peak 2 mit Abschneidung	0,0033	0,0004	0,0002	0,0002	0,0001	0,0001	0,0001	0,5
Peak 2 ohne Abschneidung	0,0049	0,0013	0,0007	0,0004	0,0003	0,0002	0,0001	0,5
Peak 1 Maximum-Methode	**0,0191**	**0,0038**	**0,0019**	**0,0010**	**0,0005**	**0,0003**	**0,0001**	**-**

Betrachtet man nun die Ergebnisse der Methode der Schwerpunktbildung aus Tabelle 14 bei der Verwendung verschiedener Segmentierungen für die Volumenberechnungen, erkennt man, dass die *Schwerpunktbildung mit Abschneidung* des Volumens bei einer Segmentierungstiefe von 0,1 (grau hinterlegt) im Mittel über alle Auflösungen die besten Ergebnisse liefert. Alle Volumen wurden durch das Modul Standardvolumen bestimmt, da beide Signale „Peak 1“ und „Peak 2“ eine Position am Extremum der jeweiligen Signalform aufweisen.

In Tabelle 16 wurden die Mittelwerte der drei Arten der Positionsbestimmung beider Signale über alle Auflösungen hinweg (aus Tabelle 14) gebildet und erfasst.

*Tabelle 15: Ergebnisse der **eindimensionalen** Ergebnisse bei Spektren **ohne Rauschen*** mit Signalen ***unterschiedlichen Volumens.*** *Zur Berechnung der Volumen wurden die Segmentierungslevel von 0,001 bis 0,5 verwendet.*

Modul „Standardvolumen“ bei Schwerpunktbildung mit oder ohne Abschneidung am Segmentierungslevel und Maximum-Methode	Mittelwert aller Abstände von der Referenz-position beider Signale über alle Auflösungen	verwendete Sementierungs-tiefe
Peak 1 und Peak 2 mit Abschneidung	0,0004	0,0001
Peak 1 und Peak 2 ohne Abschneidung	0,0004	0,0001
Peak 1 und Peak 2 mit Abschneidung	0,0003	0,1
Peak 1 und Peak 2 ohne Abschneidung	0,0005	0,1
Peak 1 und Peak 2 mit Abschneidung	0,0003	0,2
Peak 1 und Peak 2 ohne Abschneidung	0,0006	0,2
Peak 1 und Peak 2 mit Abschneidung	0,0004	0,5
Peak 1 und Peak 2 ohne Abschneidung	0,0010	0,5
Peak 1 und Peak 2 Maximum	**0,0031**	**-**

Die Ermittlung des optimalen Segmentierungslevels wurde ebenso durch den niedrigsten Mittelwert aller Methoden bestimmt. In Tabelle 15 lag dieser bei einem Segmentierungslevel von 0,1 mit Abschneidung (grau hintergelt). Die herkömmliche Methode der Bestimmung der Position durch das Maximum lieferte hier schlechtere Werte (fett hervorgehoben).

Zur Wahrung der Übersichtlichkeit wird im Folgenden nur noch das Ergebnis der besten Methode grafisch angeführt.

Um das Verhalten der Methoden der Positionsbestimmung zu testen, wurden auch folgende Variationen untersucht:

- beide zu untersuchenden Signale haben dasselbe Volumen bei Spektren ohne Rauschen
- beide zu untersuchenden Signale haben unterschiedliches Volumen bei Spektren mit Rauschen
- beide zu untersuchenden Signale haben dasselbe Volumen bei Spektren mit Rauschen

Die gemittelten Bewegungsgraphen mit Erläuterung finden sich im Anhang 8.2 wieder.

Abschließend kann für den eindimensionalen Fall bei der Anwendung des Moduls „Standardvolumen“ festgestellt werden:

- Die beiden Schwerpunktbildungen mit und ohne Abschneidung sind bei einer niedrigen Auflösung stets besser und weisen den Trend auf, dass sich bei immer höher werdender Auflösung die Position des Schwerpunkts immer mehr mit der Maximum-Position deckt.
- Die Variation des Segmentierungslevels wirkt sich im eindimensionalen Fall eher gering auf die Abweichung der Ergebnisse aus.
- Ist das Volumen des Signals „Peak 2“ gleich dem des Volumen von „Peak 1“ (Abb. 52) oder unterschiedlich (Abb. 32), ist der Unterschied kaum feststellbar, da die Überlappung später als bei Signalen unterschiedlichen Volumens eintritt und damit lediglich mehr Iterationen bei vorhandenen Positionen an den Extrema der beiden Signale in die Mittelwert-Berechnung eingehen.
- Ist Rauschen im Spektrum vorhanden (Abb. 54 oder Abb. 53), ist dieses auch nicht signifikant für eine Ergebnisänderung, da die Glättung der Integration versucht, eventuell vorhandene Zerklüftungen der Signale auszugleichen. Daher soll im Folgenden auf die gesonderte Abbildung für Spektren mit anteiligem Rauschen verzichtet werden und in der Übersichts-Tabelle 16 aufgeführt werden.

Ergebnisse des Moduls „Standardvolumen“ bei zweidimensionalen Spektren

Um den Einfluss der Auflösung auf das Signal „Peak 2“ darzustellen, wurde sowohl die Bewegung horizontal als auch diagonal von „Peak 2“ hin zu „Peak 1“ untersucht. Im Anhang 8.1 sind zwei Bewegungsgraphen mit Erläuterung aufgeführt.

Im Folgenden wurden nun die Mittelwerte (wie in Abb. 31 erläutert) hinsichtlich ihrer mittleren Abweichungen analog zum eindimensionalen Experiment erfasst (Abb. 33). Theoretisch würden Pixelmittelpunkt und die Position des theoretischen Peaks dann zusammenfallen, falls sich „Peak 2“ exakt am vertikalen Mittelpunkt bewegen würde und wenn die Iterationsschritte (unendlich) hoch wären.

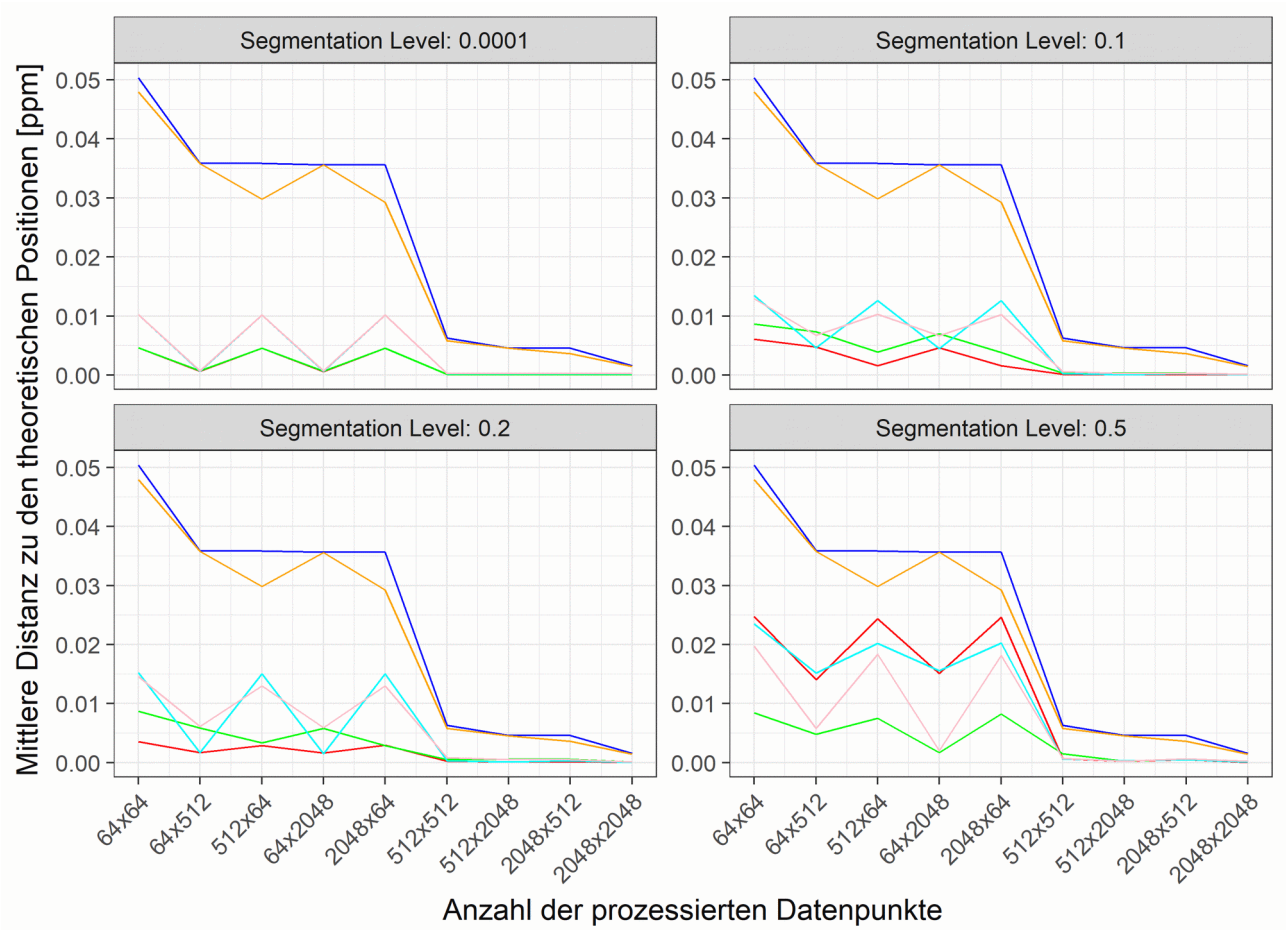

Abb. 33: Anwendung des Moduls "Standardvolumen" bei der Bewegung ***horizontal*** *in* ***zweidimensionalen*** *Spektren* ***ohne Rauschen.*** *Die Signale haben* ***unterschiedliches Volumen****, ansonsten analog zu Abb. 32; der kleinste mittlere Abstand zur theoretischen Position wurde durch die Schwerpunktbestimmung mit Abschneidung am Segmentierungslevel bei 0,0001 (links oben) mit 0 ppm erreicht. Die mittlere Abweichung der Maximum-Methode liegt bei 0,022 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

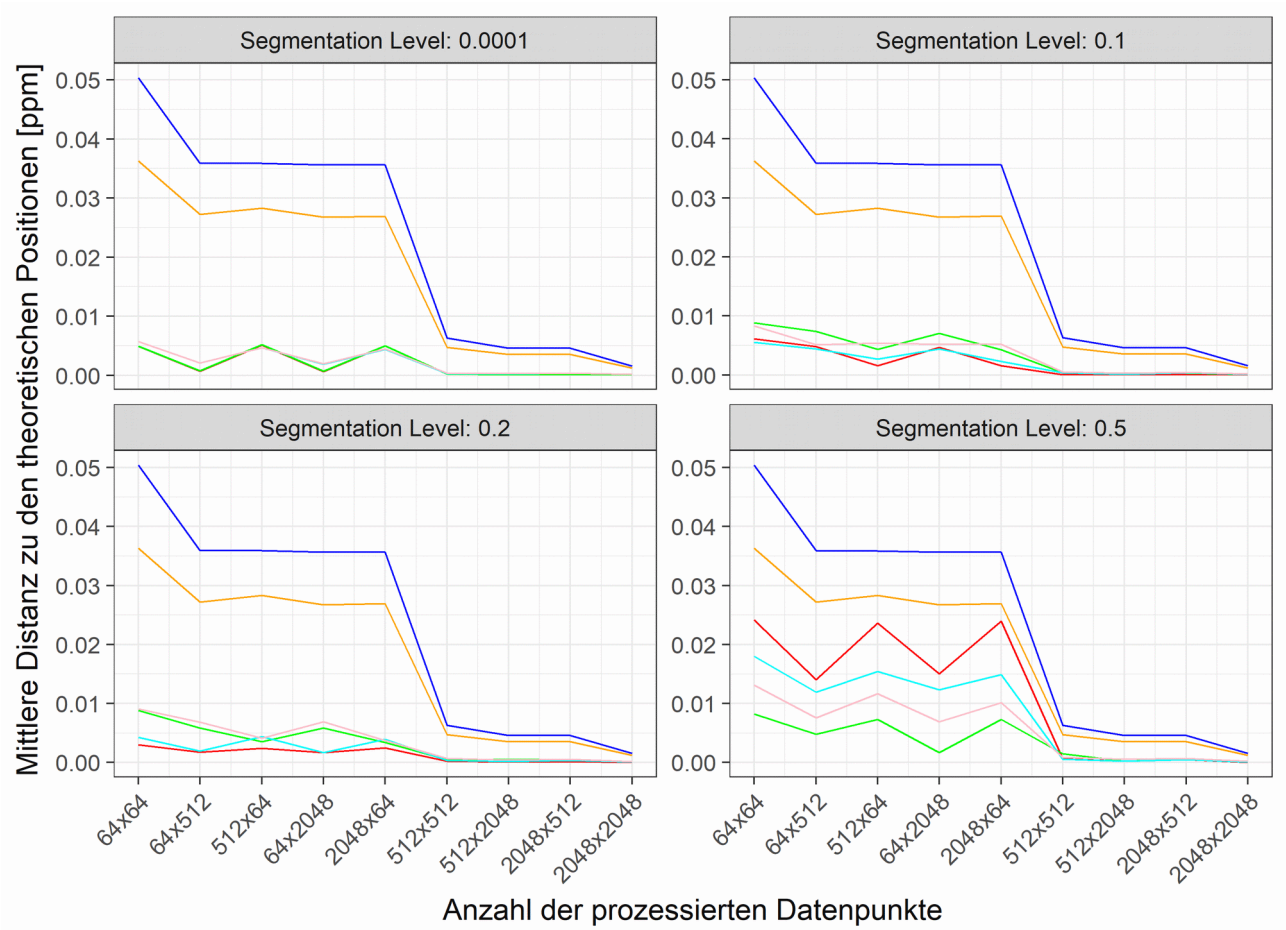

*Abb. 34: Anwendung des Moduls "Standardvolumen" bei der Bewegung **diagonal** in **zweidimensionalen** Spektren **ohne Rauschen.** Die Signale haben **unterschiedliche Volumen**, ansonsten analog zu Abb. 32; Der kleinste **mittlere** Abstand zur theoretischen Position wurde durch die Schwerpunktbestimmung mit Abschneidung am Segmentierungslevel bei 0,2 (links unten) mit 0 ppm erreicht. Die mittlere Abweichung der Maximum-Methode beträgt 0,02 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Wiederum kann hier für den zweidimensionalen Fall bei der Anwendung des Moduls „Standardvolumen“, welche je ein Extremum an den Signalpositionen aufweisen, festgestellt werden:

- Die Schwerpunktbildung ist bei niedrigeren Auflösungen stets besser und weist einen ähnlichen Trend auf wie im eindimensionalen Fall.
- Die Segmentierung wirkt sich im zweidimensionalen Fall stärker aus, als im eindimensionalen Fall. Im Falle der sehr tiefen Segmentierung der Volumen bei

0,0001 weichen die Kurven bei „Schwerpunktbildung mit Abtrennung“ (grün und cyan) und „Schwerpunktbildung ohne Abtrennung“ (grün und rosa) kaum voneinander ab.

- Ist das Volumen des Signals „Peak 2“ gleich dem Volumen von „Peak 1“ oder unterschiedlich, ist die Auswirkung wie im eindimensionalen Fall nicht hinreichend signifikant, daher wurde auf die grafische Darstellung dieser Fälle verzichtet. Die Ergebnisse sind jedoch in Tabelle 16 aufgeführt.

Im Falle von vorhandenem Rauschen wird auf die grafische Darstellung verzichtet, da sich die Ergebnisse kaum von obiger Methode ohne Rauschen unterscheiden.

Zusammenfassung der Ergebnisse zur Berechnung der Position des Schwerpunkts durch das Modul „Standardvolumen“ - Vergleich der Schwerpunktbildung mit der *Maximum-Methode*

Um die vorangegangenen Ergebnisse bewerten zu können wurden diese in Tabelle 16 in einer Übersicht zusammengefasst.

Hier wurde der Mittelwert zu jeder Bewegung einer Methode (*Schwerpunktbestimmung mit Abschneidung, Schwerpunktbestimmung ohne Abschneidung* und *Maximum-Methode*) und über beide Signale („Peak 1“ und „Peak 2“) gebildet. Für eindimensionale Spektren lieferte eine Segmentierungtiefe von 0,1 im Mittel die besten Ergebnisse, während für zweidimensionale Spektren eine Segmentierungstiefe von 0,2 die besten Ergebnisse lieferte. Der Unterschied der Ergebnisse bei der Verwendung von Spektren mit und ohne Rauschen sowie unterschiedliche Volumen der Signale „Peak 1“ und „Peak 2“ waren nicht signifikant.

*Tabelle 16: Ergebnisse zur Berechnung der Position des Schwerpunkts durch das Modul **"Standardvolumen"** bei optimalem Segmentierungslevel (letzte Spalte) und der Maximum-Methode, falls **beide** Signale ein **Extremum** an der Signalform aufweisen. Dabei stellen die Mittelwerte (MW) die mittlere Abweichung beider Signale „Peak 1" und „Peak 2" (von den theoretischen Referenzpositionen) einer Methode der Positionsbestimmung (Spalte 2-4) aus einer Gruppierung nach der digitalen Auflösung der verwendeten Spektren aus allen generierten Bewegungsgraphen dar.*

verwendete Spektren und Bewegung	MW *Schwerpunkt-bestimmung **ohne** Abschneidung;* verwendetes Modul „Standardvolumen"	MW *Schwerpunkt-bestimmung **mit** Abschneidung;* verwendetes Modul „Standardvolumen"	MW *Maximum-Methode*	Volumen der Signale „Peak 1" und „Peak 2" unterschiedlich?	Rauschen vorhanden?	verwendetes Segmentierungs-level
eindimensional	0,001	0,000	0,003	ja	nein	0,1
	0,001	0,000	0,003	nein	nein	0,1
	0,001	0,000	0,004	nein	ja	0,1
	0,001	0,000	0,004	ja	ja	0,2
zweidimensional – Bewegung horizontal	0,003	0,003	0,022	ja	nein	0,0001
	0,003	0,003	0,022	nein	nein	0,0001
	0,006	0,004	0,023	nein	ja	0,2
	0,004	0,003	0,023	ja	ja	0,2
Zweidimensional - Bewegung diagonal	0,003	0,002	0,020	ja	nein	0,2
	0,003	0,002	0,021	nein	nein	0,2
	0,003	0,002	0,021	nein	ja	0,2
	0,003	0,001	0,021	ja	ja	0,2

Abschließend kommt man zu dem Ergebnis, dass die Methode zur Berechnung der Signalposition durch den Schwerpunkt und Reduzierung der Intensitäten am Segmentierungslevel (Spalte 3 aus Tabelle 16) die besseren Resultate lieferte. Daher ist die *Schwerpunktbestimmung mit Abschneidung* der *Maximum-Methode* vorzuziehen. Die Wahl der Integrationsmethode zur Bestimmung der Signalvolumen ist irrelevant, da sich die Positionen beider Signale jeweils am Extremum deren Signalform befinden.

3.3.8 Die Positionsbestimmung durch die Festlegung des Schwerpunkts bei Existenz von nur einem Extremum – Anwendung der Module „gemeinsames Volumen“ und „getrenntes Volumen“

Falls sich zwei Signale so stark überlagern, dass keine unterscheidbaren Extrema der beiden Signalformen mehr vorhanden sind, nimmt die Integration eine Annäherung aufgrund des vorhandenen Extremums des Signals „Peak 1“ und der Position (festgelegt durch die Simulation) von „Peak 2“ ohne Extremum vor.

Es liegen damit zwei Signale mit folgenden Positionen vor:

- Position eines Signals, welches ein Extremum am digitalen Raster belegt
- Position eines Signals, welches **kein** Extremum am digitalen Raster belegt (z. B. durch manuelles Picken oder durch eine Simulation)

Für diesen Fall sollen nun folgende in Abschnitt 3.3.5 definiert Module angewandt werden:

- Modul „gemeinsames Volumen“
- Modul „getrenntes Volumen“

Im Folgenden soll gezeigt werden, welches der beiden Module die besseren Ergebnisse hinsichtlich der Abweichungen zu den theoretischen Positionen bestimmt. Dazu werden zunächst die Ergebnisse des Moduls „getrenntes Volumen“ (mittels der *gemittelten Bewegungsgraphen)* und anschließend die Ergebnisse des Moduls „getrenntes Volumen“ dargestellt. Die Darstellung und der Aufbau der gemittelten Bewegungsgraphen geschieht analog zum vorhergehenden Kapitel für das Modul „Standardvolumen“ (siehe 3.3.7).

3.3.8.1 Berechnung der Position des Schwerpunkts durch das Modul „getrennte Volumen“

Hier wurde ein Signal ohne ein Extremum (im Folgenden legt diese Position stets „Peak 2“ fest) bei der Integration manuell zur Seedliste hinzugefügt, was die Konsequenz hatte, dass das vorerst gemeinsame Volumen der beiden Signale aufgetrennt wurde. Da sich die Position des Nicht-Extremums direkt an der Integrationsgrenze der Signalform befindet, wird mit dieser Methode im Prinzip nur die vom dominierenden „Peak 1“ abgewandte Flanke integriert und „Peak 2“ zugeordnet (siehe dazu Kapitel 3.2.4).

Ergebnisse des Moduls „getrennte Volumen“ bei eindimensionalen Spektren

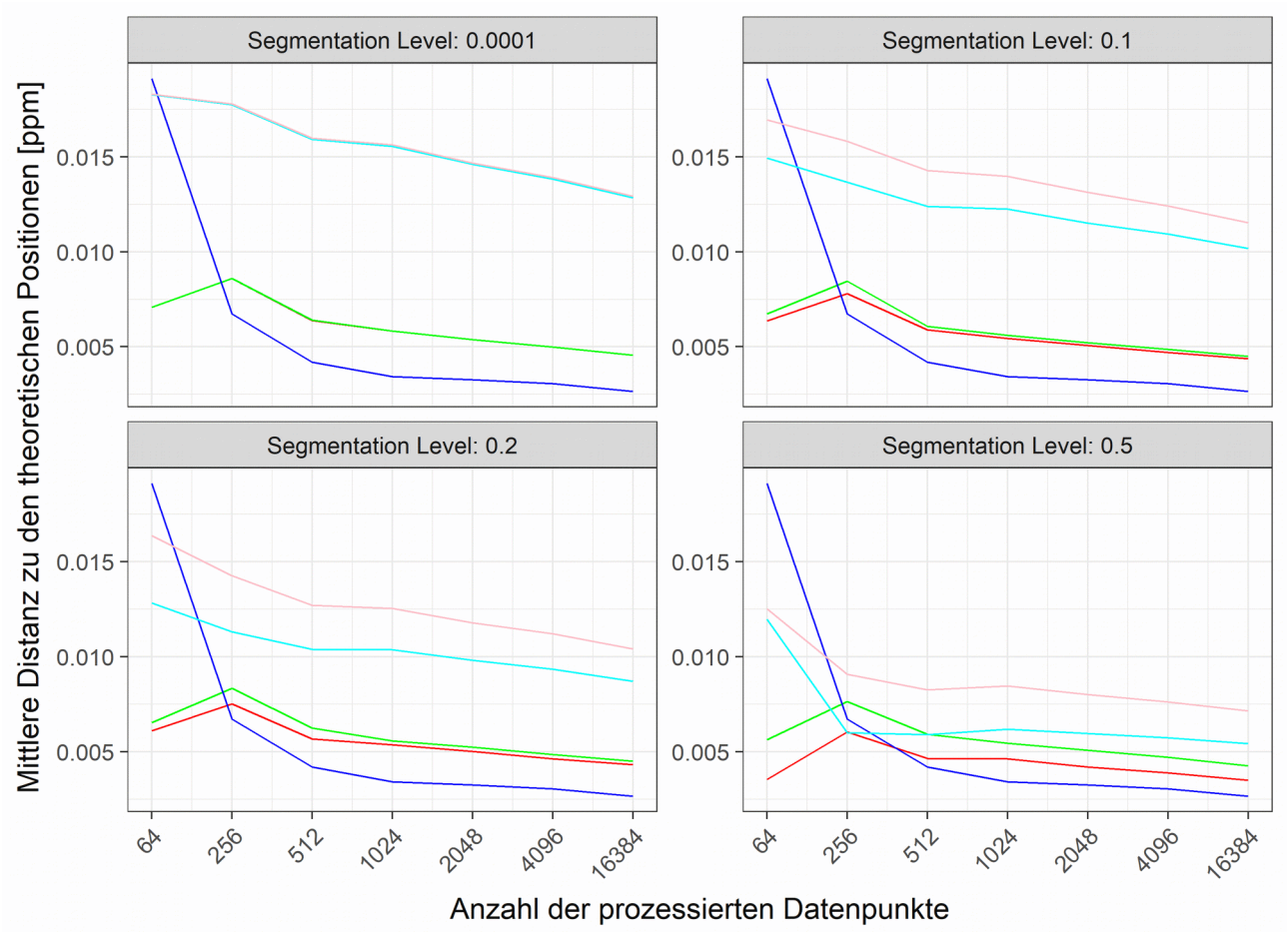

*Abb. 35: Anwendung des Moduls „**getrennte Volumen**“ bei der Bewegung in **eindimensionalen** Spektren **ohne Rauschen** analog zu Abb. 32. Beide Signale besitzen **unterschiedliche Volumen**. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die Methode Abschneiden am Segmentierungslevel bei 0,5 mit 0,0054 ppm erreicht. Die mittlere Abweichung der Maximum-Methode lag bei 0,006 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Die Kurve der Maximum-Methode für das Signal „Peak 2“ fehlt in Abb. 35 (in den vorhergegangenen Abbildungen gelb), da kein Maximum für dieses Signal existiert. Die Position von „Peak 2“ (welche keine Position am digitalen Raster darstellt) wurde jedoch durch die Simulation festgelegt. Die restlichen Farbcodes sind analog zum vorherigen Kapitel definiert.

Es kann für den eindimensionalen Fall bei der Anwendung des **Moduls getrennte Volumen** festgestellt werden:

- Die Schwerpunktmethoden sind bei der niedrigsten Auflösung besser, jedoch weisen sie nicht den Trend auf, dass sich bei der Erhöhung der digitalen Auflösung die Position des Schwerpunkts immer mehr mit der Position am Extremum deckt. Der Grund hierfür ist die Abtrennung des Volumens von „Peak 2“ direkt an dessen Position. Dies resultiert in ein zu kleines Volumen von „Peak 2“ und ein zu großes von „Peak 1“.
- Die Variation der Segmentierungstiefe im eindimensionalen Fall wirkt sich hier stark auf die Ergebnisse aus. Die höhere Segmentierung bei 0,5 (in etwa bei halber Höhe des Signals) verbessert die Positionierung.
- Die Variationen „Gleiches Volumen“ der beiden Signale und „Rauschen“ werden am Ende dieses Kapitels in Tabelle 17 aufgeführt.

Der direkte Vergleich der Schwerpunkt-Methoden stellt sich in diesem Fall als schwierig dar, denn der Maximum-Peak-Pick-Algorithmus kann diese Position nicht festlegen, da „Peak 2“ keine Position am Extremum der Signalform aufweist. Daher sollte die Methode bezüglich „Peak 2“ lediglich eine Annäherung darstellen. Schlussendlich sollte ein Ergebnis erzielt werden, bei dem unter minimaler Verschlechterung für „Peak 1“ überhaupt eine Position für „Peak 2“ berechnet werden kann. Dies war der Grund die im nächsten Unterkapitel beschriebene Methode „gemeinsames Volumen“ zu implementieren um zu testen, ob dies die Position für „Peak 1“ verbessert.

Ergebnisse des Moduls „getrennte Volumen“ bei zweidimensionalen Spektren

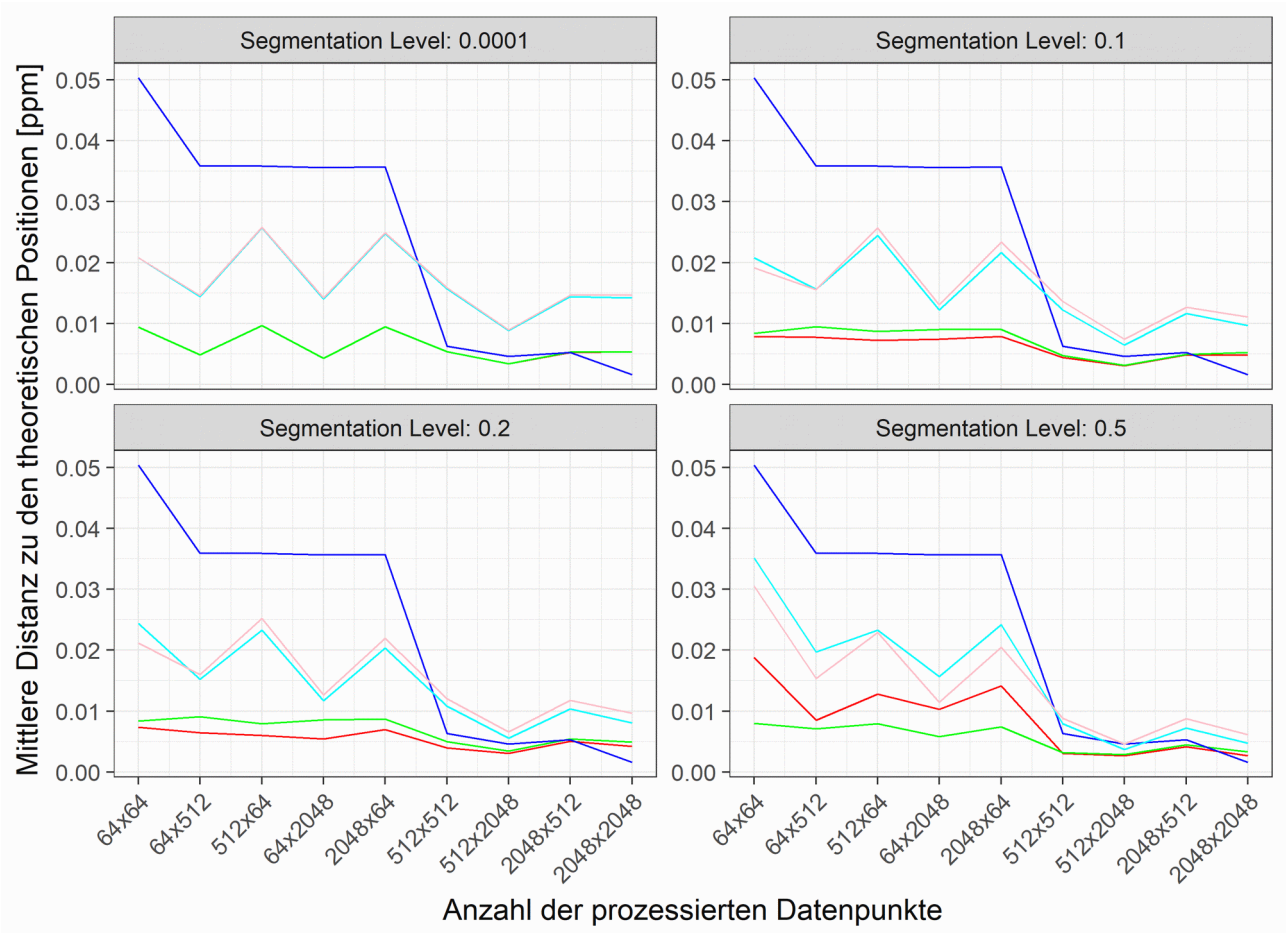

*Abb. 36: Anwendung des Moduls **getrennte Volumen** bei der Bewegung in **zweidimensionalen** Spektren **ohne Rauschen** analog zu Abb. 32. Peak 2 bewegt sich **diagonal** auf Peak 1 zu; Peak 1 und Peak 2 besitzen **unterschiedliche Volumen**. Die Kurve des Extremum von Peak 2 fehlt hier, da es kein Maximum für diesen Peak gibt. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die Methode Abschneiden am Segmentierungslevel bei 0,2 (links unten) mit 0,001 ppm (Tabelle 17) erreicht. Die mittlere Abweichung der Maximum-Methode lag bei 0,023 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Für den zweidimensionalen Fall bei der Anwendung des Moduls „getrennte Volumen“ konnte festgestellt werden:

- Beide *Schwerpunktmethoden* sind bei einer beteiligten niedrigen Auflösung besser als die Maximum Methode, jedoch weisen sie nicht den Trend auf, dass sich bei immer höher werdender Auflösung die Position des Schwerpunkts immer mehr mit der Position am Extremum deckt. Je höher die Spektren aufgelöst (digitale

Auflösung) sind, desto mehr schwindet der Vorteil der Schwerpunktmethode, jedoch konnte die Position für das Signal „Peak 2“ ausreichend genau bestimmt werden.

- „Gleiches Volumen“ der beiden Signale und „Rauschen“ werden in Tabelle 17 aufgeführt.
- Die Kurve der Mittelwerte der *Maximum-Methode* (blaue Kurve) ist zwar in den verwendeten Abbildungen vorhanden, wird aber erst im letzten Abschnitt dieses Kapitels betrachtet, wenn es darum geht, zu bewerten, ob die *Schwerpunktmethode* eine alternative zur *Maximum-Methode* darstellt.

*Tabelle 17: Übersicht Ergebnisse des Moduls „**getrennte Volumen“** bei den optimalen Segmentierungstiefen unter Anwendung der Schwerpunktbestimmung ohne Abschneidung und der Schwerpunktbestimmung mit Abscheidung am optimalen Segmentierungslevel (analog zur Tabelle 16).*

verwendete Spektren und Bewegung	MW *Schwerpunkt-bestimmung **ohne** Abschneidung;* verwendetes Modul „getrennte Volumen“	MW *Schwerpunkt-bestimmung **mit** Abschneidung;* verwendetes Modul „getrennte Volumen“	Volumen der Signale „Peak 1“ und „Peak 2“ unterschiedlich?	Rauschen vorhanden?	verwendetes Segmentierungs-level
eindimensional	0,007	0,006	ja	nein	0,5
	0,008	0,008	nein	nein	0,5
	0,010	0,009	nein	ja	0,2
	0,007	0,005	ja	ja	0,5
zweidimensional – Bewegung horizontal	0,012	0,011	ja	nein	0,2
	0,017	0,017	nein	nein	0,0001
	0,019	0,019	nein	ja	0,5
	0,013	0,016	ja	ja	0,5
zweidimensional – Bewegung diagonal	0,011	0,010	ja	nein	0,2
	0,013	0,012	nein	nein	0,2
	0,014	0,014	nein	ja	0,5
	0,010	0,012	ja	ja	0,5

Betrachtet man die Ergebnisse aus Tabelle 17, lässt sich bei nahezu allen mittleren Abständen der beiden Schwerpunkt-Methoden ablesen, dass zur Bestimmung der Signal-Position mit dem Modul „getrennte Volumen“ unter Verwendung der Methode der Schwerpunktbestimmung durch Reduzierung der Intensitäten am Segmentierungslevel die besseren Ergebnisse erzielt werden konnten.

3.3.8.2 Berechnung der Position des Schwerpunkts durch das Modul „gemeinsames Volumen“

Bei der Volumenzuweisung in die Peakliste wird jedes Signalvolumen bezüglich der Segmentierung und relativ zu deren Intensität an dessen Position der Ausgangsposition vor dem Verschieben korrigiert. D.h. das Signal „Peak 2“ hat durch sein fehlendes Extremum eine geringere Intensität als „Peak 1“. Die Signale teilen sich zwar die beteiligten Pixel am Volumen, aber das resultierende gleiche Volumen beider Signale wird relativ zu den Signalintensitäten aufgeteilt (siehe Kapitel 3.2.4.3). Dies hat jedoch für die Bildung des Schwerpunkts keinerlei Auswirkungen, da es hierbei egal ist, wie die Intensitäten skaliert sind (siehe Formel 7).

Ergebnisse des Moduls „gemeinsames Volumen“ für eindimensionale Spektren

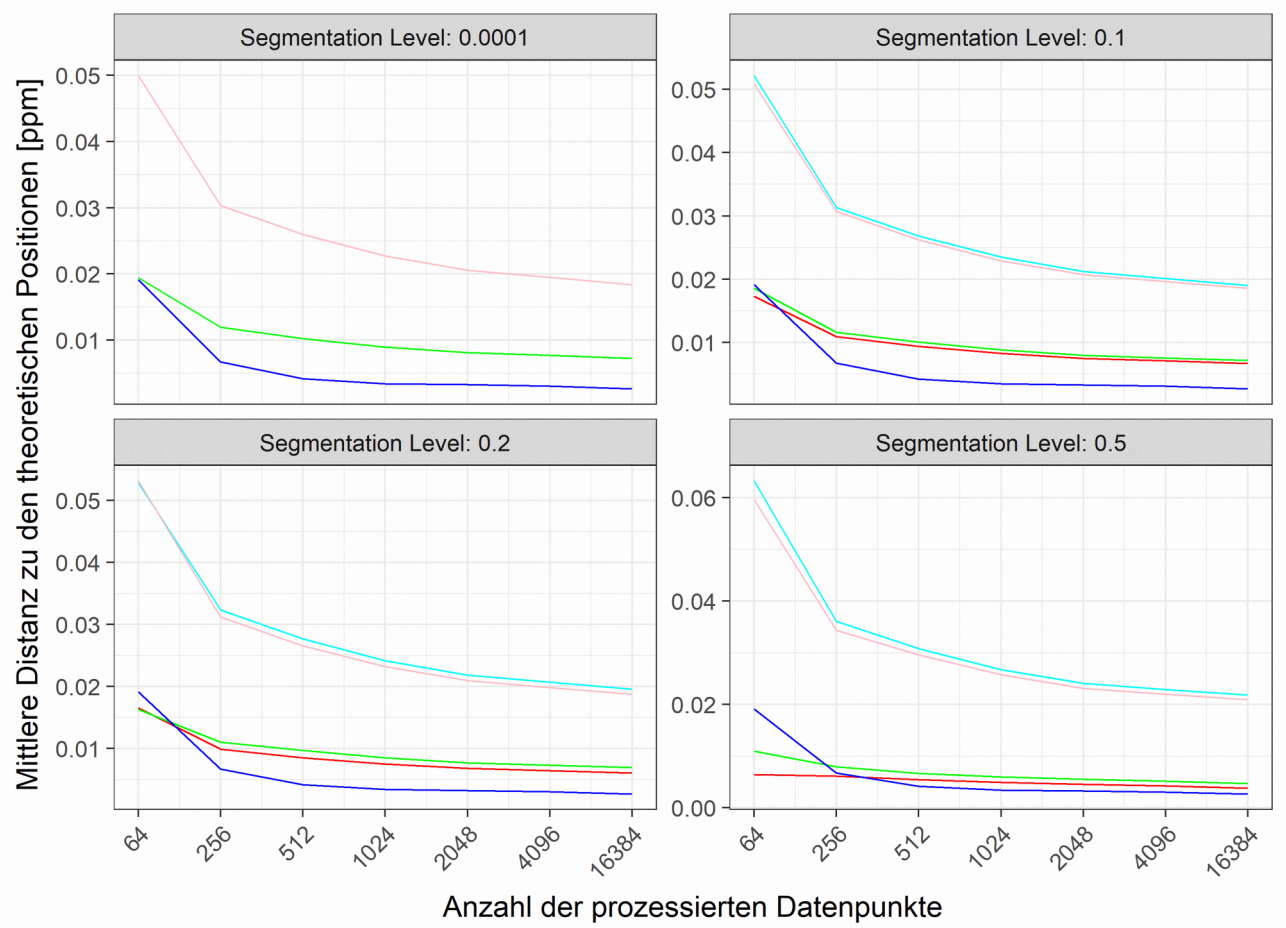

*Abb. 37: Anwendung des Moduls „**gemeinsames Volumen**„ bei **eindimensionalen** Spektren **ohne Rauschen** analog zu Abb. 32. Beide Signale haben **unterschiedliche Volumen**. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die Positionsbestimmung mit Abschneidung am Segmentierungslevel bei 0,5 (rechts unten) mit 0,005 ppm erreicht. Die mittlere Abweichung der Maximum-Methode lag bei 0,006 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Ergebnisse des Moduls „gemeinsames Volumen“ für zweidimensionale Spektren

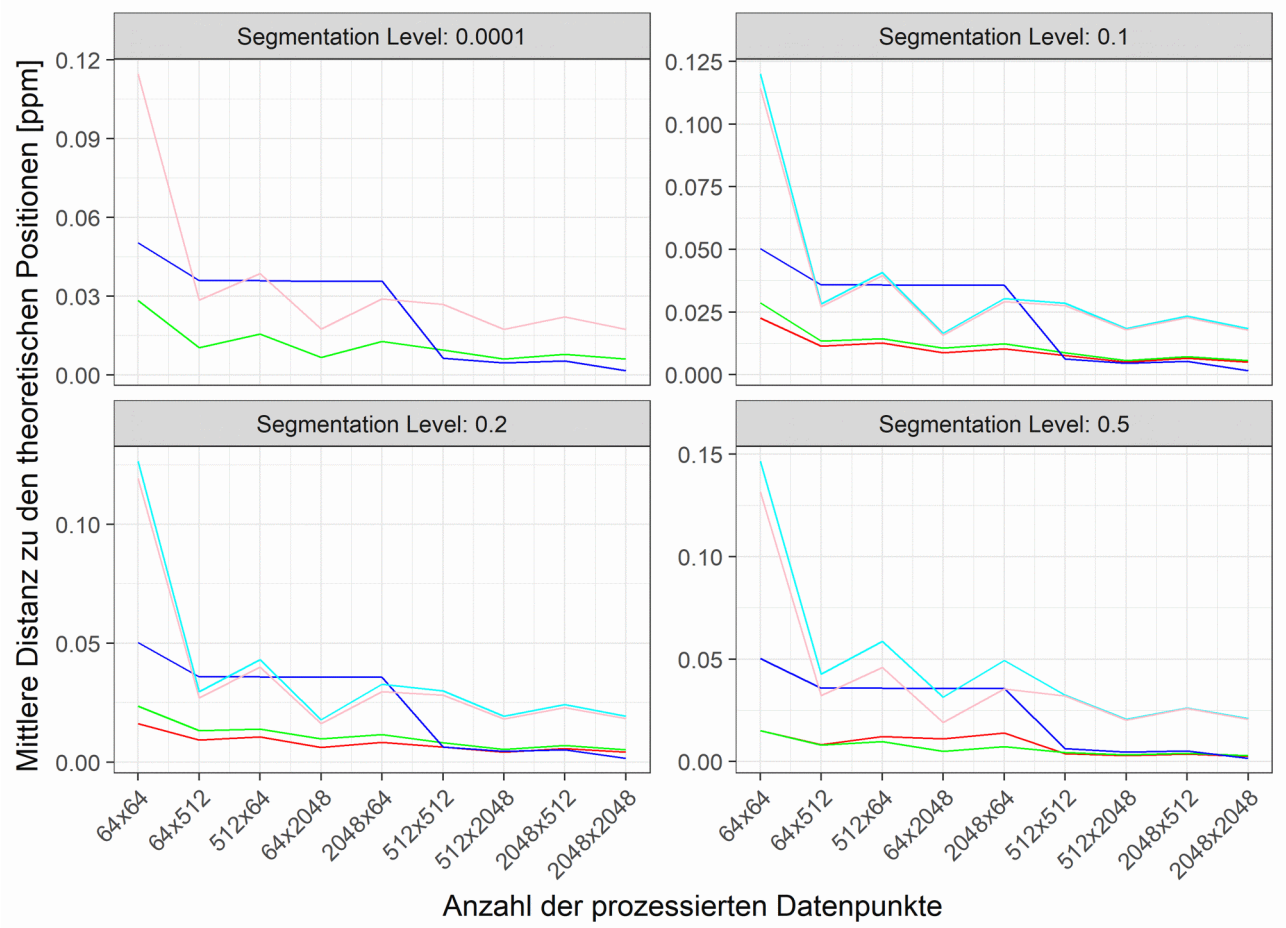

*Abb. 38: Anwendung des Moduls „**gemeinsames Volumen**“ bei **zweidimensionalen** Spektren **ohne Rauschen**. Das Signal „Peak 2“ bewegt sich diagonal auf „Peak 1“ zu; Beide Signale haben **unterschiedliche Volumen**. Die Kurve des Extremum von „Peak 2“ fehlt hier, da es kein Extremum für dieses Signal an der Signalform gibt. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die Methode Abschneiden am Segmentierungslevel bei 0,2 (links unten) mit 0,01 ppm erreicht. Die mittlere Abweichung der Maximum-Methode (nur „Peak 1“) betrug 0,023 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

*Tabelle 18: Übersicht Ergebnisse des Moduls **„gemeinsames Volumen“** bei den optimalen Segmentierungstiefen unter Anwendung beider Schwerpunkt-Methoden (mit und ohne Reduzierung der Intensitäten im Volumen) analog zur Tabelle 16.*

verwendete Spektren und Bewegung	MW *Schwerpunkt-bestimmung **ohne** Abschneidung;* verwendetes Modul „gemeinsames Volumen“	MW *Schwerpunkt-bestimmung **mit** Abschneidung;* verwendetes Modul „gemeinsames Volumen“	Volumen der Signale „Peak 1“ und „Peak 2“ unterschiedlich?	Rauschen vorhanden?	verwendetes Segmentierungs-level
eindimensional	0,019	0,019	ja	nein	0,0001
	0,014	0,014	nein	nein	0,0001
	0,016	0,016	nein	ja	0,1
	0,02	0,02	ja	ja	0,2
zweidimensional – Bewegung horizontal	0,027	0,027	ja	nein	0,0001
	0,02	0,02	nein	nein	0,0001
	0,034	0,034	nein	ja	0,0001
	0,039	0,039	ja	ja	0,0001
zweidimensional – Bewegung diagonal	0,023	0,023	ja	nein	0,2
	0,015	0,015	nein	nein	0,2
	0,026	0,026	nein	ja	0,2
	0,032	0,032	ja	ja	0,2

Vergleicht man die Ergebnisse des Moduls „getrennte Volumen“ aus Tabelle 17 mit den Resultaten aus Tabelle 18, lässt sich feststellen, dass das Modul „getrennte Volumen“ über das Modul „gemeinsames Volumen“ dominiert. Daher werden im Folgenden Kapitel die Volumen zur Schwerpunktbildung durch das Modul „getrennte Volumen“ berechnet und die *Schwerpunktbestimmung mit oder ohne Abschneidung* mit der *Maximum-Methode* verglichen.

3.3.9 Vergleich der *Schwerpunktbestimmung* durch das Modul „getrennte Volumen“ mit der *Maximum-Methode* im Falle von nur einem Extremum

Wie bereits in Abschnitt 3.3.7 gezeigt wurde, lieferte das Modul "Standardvolumen" bei der *Schwerpunktbildung mit* oder *ohne Abschneidung* am Segmentierungslevel bessere Ergebnisse als die *Maximum-Methode*. Falls jedoch das sich bewegende Signal „Peak 2“ kein Extremum mehr durch Überlappung mit dem Signal „Peak 1“ aufweist, soll hier gezeigt werden, welche Methode zur Bestimmung der Signalposition im Mittel die geringste Abweichung zur theoretischen Referenzposition liefert.

Tabelle 19: Vergleich der beiden Schwerpunktbestimmungen mit der Maximum-Methode, falls nur noch ein Extremum bei „Peak 1“ existiert. Zur Bestimmung der Schwerpunkte wurde das Modul „getrennte Volumen“ verwendet; „Peak 2“ kann in diesem Fall nicht verglichen werden, da dieser kein Extremum aufweist.

verwendete Spektren und Bewegung	MW *Schwerpunkt-bestimmung mit Abschneidung der Position des Signals* „Peak 1“; verwendetes Modul „getrennte Volumen“	MW *Maximum-Methode* für Signal „Peak 1“	Volumen der Signale „Peak 1“ und „Peak 2“ unterschiedlich?	Rauschen vorhanden?	Segmen-tierungstiefe
eindimensional	0,004	0,006	ja	nein	0,5
	0,007	0,015	nein	nein	0,5
	0,007	0,006	nein	ja	0,5
	0,004	0,006	ja	ja	0,5
zweidimensional – Bewegung horizontal	0,005	0,023	ja	nein	0,2
	0,015	0,035	nein	nein	0,2
	0,018	0,026	nein	ja	0,2
	0,007	0,023	ja	ja	0,2
zweidimensional – Bewegung diagonal	0,005	0,023	ja	nein	0,2
	0,01	0,031	nein	nein	0,2
	0,012	0,025	nein	ja	0,2
	0,006	0,024	ja	ja	0,2

In Tabelle 19 wurden analog zu vorherigen Abschnitten die Abweichungen im Mittel zur Referenzposition über alle sieben Variationen der Auflösung der Spektren aufgeführt. Dazu wurde beim Modul „getrenntes Volumen“ lediglich die Variante *Positionsbestimmung mit Abschneidung* am Segmentierungslevel (Abb. 23) betrachtet, da diese identische

Ergebnisse aufwies. Da das Signal „Peak 2“ hier nicht ausgewertet wurde, war das beste Ergebnis (also allein auf „Peak 1“ beschränkt) die *Schwerpunktbildung mit Abschneidung am Segmentierungslevel*. Zudem war in fast allen Fällen das Segmentierungslevel 0,2 im zweidimensionalen Fall und 0,5 im eindimensionalen Fall die optimale Schwelle.

Es konnte gezeigt werden, dass nahezu alle Methoden (Tabelle 19) der Bestimmung der Signalposition durch den Schwerpunkt bessere Ergebnisse lieferten als die Maximum-Methode, falls eine Überlappung so stark ist, dass das andere überlappte Signal kein Extremum mehr aufweist. Dabei war das Modul „getrennte Volumen“ die Methode, die gegenüber den anderen Methoden am besten abgeschnitten hat.

3.4 Signalidentifizierung durch die Bestimmung der Bayesschen Wahrscheinlichkeit

3.4.1 Die Bestimmung der optimalen Parameter zur Berechnung der Verteilungen der Eigenschaften

Die Bestimmung der Eigenschaften kann im einfachsten Fall direkt aus der Peakliste entnommen werden oder im komplexeren Fall muss die gewünschte Eigenschaft aus zusätzlichem Wissen über das NMR-Signal (aus dem Rohspektrum und der Peakliste) ermittelt werden.

Die Eigenschaften von NMR-Signalen werden in vier **Eigenschaftstypen** gruppiert:

- nicht volumenbasierte Eigenschaften
- volumenbasierte Eigenschaften
- gaußsche Signalwahrscheinlichkeit (Wahrscheinlichkeit basierend auf dem lokalen Rauschen)
- Symmetrie-Eigenschaften

Die Berechnungen aller Eigenschaften wurden in zwei Schritten durchgeführt (siehe auch Abschnitt 2.6.5.4):

1. Eigenschaften, welche sich durch verschiedene Berechnungsmethoden variieren lassen. Das ist z. B. bei volumenbasierten Eigenschaften die variierende Segmentierungstiefe oder bei der äußeren Symmetrie die Suchmethode des symmetrischen Partners. Die Methoden zur Berechnung haben je nach Eigenschaftstyp verschieden viele freie Parameter.

   Eigenschaften wie die Intensität oder Linienbreite werden direkt aus der Peakliste ausgelesen und benötigen keine Berechnung.

2. Jeder Datensatz einer Eigenschaft (also die Ansammlung einer Eigenschaft aller NMR-Signale) kann nachträglich reskaliert werden, falls dies nötig ist. So wäre z. B. die Bildung des Absolutwerts sinnlos, wenn die Daten keine negativen Werte aufweisen.

Daraus ergeben sich für Schritt 1 mehrere Variationen, wie die Parameter zur Berechnung der Datensätze ermittelt werden können. Liegt der Datensatz einer Eigenschaft vor, muss unterschieden werden, welcher Typ von Verteilungen (*theoretische Verteilung* durch Simulated Annealing oder *geglättete Verteilung*) verwendet werden soll, um eine spätere Diskriminierung durchführen zu können. Soll die Wahrscheinlichkeitsdichteverteilung durch die *geglätteten Verteilungen* generiert werden, so kommt eine zusätzliche Kombinationsmöglichkeit hinzu, nämlich die Größe des „Glättungsfaktors“ mit dem die Verteilung geglättet werden soll.

Um die optimalen Parameter zur Berechnung der Eigenschaften auch noch in ihrer Güte bewerten zu können, wurde eine sog. *Hitliste* zu jeder Eigenschaft generiert. Diese *Hitliste* beinhaltet die Ergebnisse einer Diskriminierung von Signal und Störsignalen durch Bildung der Bayesschen Wahrscheinlichkeiten aller NMR-Signale bei Variation der Verteilungsgenerierung aller verwendeten Klassen (*Rauschen*, *Signal* bzw. *Wasser*). Zusätzlich wurde die Klasse *Signal* bereinigt, indem die durch eine bereits existierende Zuordnung des verwendeten Spektrums bekannten Störsignale aus dieser Klasse entfernt wurden.

Danach wurde der Schwellwert der Wahrscheinlichkeit bestimmt, bei dem die höchste prozentuale korrekte Wiedererkennung der Nutzsignale bei **gleichzeitig** höchster Wiedererkennung der Störsignale des gesamten Spektrums erreicht wird. Dabei wurde für jede Variation des Parametersatzes der Mittelwert aus den Prozentwerten der korrekten Wiedererkennung (richtig positiv und richtig negativ) gebildet.

Alle Resultate wurden dann nach diesem Mittelwert absteigend sortiert, so dass die erste Zeile der Hitliste die optimale Berechnungsgrundlage dieser Eigenschaft darstellte:

- optimale **Berechnungsparameter** zu Erstellung des Datensatzes der Eigenschaft für alle NMR-Signale aus der Peakliste
- optimale **Skalierung** des Datensatzes für die Bildung der Verteilungen

Da auch die Variation der Klassenanzahl untersucht werden sollte, wurde für jede Klassenkombination ebenfalls eine Hitliste erstellt, welche sich durch die unterschiedliche Generierung der Störsignalklassen (Klasse *Rauschen* und *Wasser*) hinsichtlich der Klassenbereiche auszeichnet.

Da die Methode der Generierung der Hitlisten für alle Eigenschaften gleichartig war, wurde im Folgenden nur die Generierung der Hitliste der Eigenschaft der Intensität (also die Intensität der der Position des Signals) exemplarisch dargestellt. Für dieses Fall wurden die beiden Klassen *Signal* und *Rauschen* mit den jeweiligen Bereichen aus Tabelle 6 und 7 verwendet. Für die Generierung der Datenbasis der folgenden Abschnitte lag sowohl das experimentelle als auch das simulierte Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).

*Tabelle 20: Auszug aus der **Hitliste** als Zwischenergebnis der Eigenschaft **Intensität** bestehend aus den 40 Diskriminierungen durch die Bayessche Wahrscheinlichkeit bei Verwendung der Klassen Signal und Rausch durch die Variation der **Berechnungsparameter** und **Skalierungen.***

Die erste Zeile (fett hervorgehoben) zeigt die beste Diskriminierung mit 98,71 % korrekt wiedererkannten Signalen und 90,91 % korrekt wiedererkannten Störsignalen. Dabei wurde die Schwelle für die Wahrscheinlichkeit von 0,32 verwendet. Es wurde keine andere Skalierung für den Datensatz verwendet und der optimale Glättungsfaktor zur Erstellung der zugrundeliegenden geglätteten Verteilungen war ein adaptiver Glättungsfaktor von -1 (also nur ein erlaubtes Extremum in der Wahrscheinlichkeitsdichteverteilung).

0|1.89646616029302|0.987088156723063|0.909378003569957|46389|6651|39738|0.318211225357798	**0_C[S_N]Intensity[a[0]sc[no scaling]]smf[-1].bayescalculator**
24|1.8940888785478|0.985900860789552|0.90818801775825|46329|6643|39686|0.0542852820892094	2_C[S_N]Intensity[a[0]sc[logarithmic scale]]smf[10].bayescalculator
21|1.89390478491428|0.981300089047195|0.912604695867088|46491|6612|39879|0.268718416995493	2_C[S_N]Intensity[a[0]sc[logarithmic scale]]smf[-2].bayescalculator
22|1.89363084527344|0.981003265063817|0.912627580209621|46490|6610|39880|0.266353094396507	2_C[S_N]Intensity[a[0]sc[logarithmic scale]]smf[-3].bayescalculator
23|1.89363084527344|0.981003265063817|0.912627580209621|46490|6610|39880|0.268951808632594	2_C[S_N]Intensity[a[0]sc[logarithmic scale]]smf[-4].bayescalculator
.	
.	
.	
17|1.42577863924523|0.983081032947462|0.442697606297771|25969|6624|19345|0.197142245278408	1_C[S_N]Intensity[a[1]sc[no scaling]]smf[3].bayescalculator
16|1.30504423076758|0.99168892846542|0.313355302302165|20375|6682|13693|0.00160469056204875	1_C[S_N]Intensity[a[1]sc[no scaling]]smf[2].bayescalculator
35|1.29491186715979|0.999406352033244|0.29550551512655|19647|6734|12913|8.42935296598679e-05	3_C[S_N]Intensity[a[1]sc[logarithmic scale]]smf[1].bayescalculator
15|1.18426975940711|0.992727812407243|0.191541946999863|15059|6689|8370|0.0665353757663573	1_C[S_N]Intensity[a[1]sc[no scaling]]smf[1].bayescalculator

Das Ergebnis der ersten Zeile der Hitliste spiegelt die Diskriminierung wieder, welche durch die Wahl der optimalen Parameter ermöglicht wurde. Diese ermöglichte die maximale Anzahl an Signalen (98,71 %) und Störsignalen (90,01 %) des gesamten Spektrums wiederzuerkennen. Dabei wurde eine Wahrscheinlichkeitsschwelle von 0,318 verwendet.

Die Bestimmung dieser Wahrscheinlichkeitsschwelle beruht darauf, dass die höchste Wiedererkennung von allen Signalen und die höchste Wiedererkennung von Störsignalen relativ der Zugehörigkeit ihrer Klasse mit einfließt, da ansonsten bei Verwendung der Gesamtzahl die Störsignale dominieren könnten. D. h. falls ein Spektrum z. B. 100.000 gepickte NMR-Signale aufweist und davon lediglich 5000 wirklich Nutzsignale wären,

könnte dies dazu führen dass zwar nahezu alle richtig wiedererkannt werden, jedoch die meisten als richtig negativ auf Kosten der richtig positiven dominieren würden.

Es wurden insgesamt 46389 von 50436 (91,98 %) NMR-Signale richtig erkannt und 6651 der 6738 (also 98,7 %) Signale korrekt als Nutzsignale erkannt (richtig positiv). Im Falle der Störsignale wurden jedoch nur 39738 von 43698 (90,93 %) als solche wieder erkannt (richtig negativ). Nach der Berechnung des Datensatzes wurde auf eine abweichende Skalierung verzichtet und die Werte auch nicht als Absolutwerte genommen. Der Glättungsfaktor der Verteilung wurde so gewählt (-1), dass dieser automatisch adaptiert wurde, bis nur noch ein Maximum in den Verteilungen auszumachen war. In der zweiten Spalte aus Tabelle 20 findet sich der Mittelwert der richtig positiven und richtig negativen Wiedererkennungen wieder. Diese liegt hier nicht als Mittelwert vor, da diese Liste nur ein Zwischenergebnis in dem vom Autor erstellten Auswerte-Tools darstellt. Tabelle 20 soll daher nur einen Eindruck verschaffen, wie die Generierung der optimalen Datensätze erfolgt und hat **lediglich einen schematischen** Charakter.

Beide Kurven aus Abb. 39 wurden mit der adaptiven Glättung des Moduls *geglättete Verteilungen* erstellt und der *Glättungsfaktor* solange variiert, bis nur noch ein Maximum vorhanden war. Dabei wurde zur Glättung der Klasse *Signal* ein Glättungsfaktor von 63 und bei der Klasse *Rauschen* ein *Glättungsfaktor* von 711 gefunden.

Zur besseren Visualisierung der signifikanten Bereiche der Verteilungen in Abb. 39 wurden nur Intensitäten zwischen -40.000 und 40.000 aufgetragen, da die Ausläufer der Verteilungen sehr niedrige relative Häufigkeiten aufweisen. Zudem sind die Verteilungen nicht auf 1 normiert, da die Normierung durch die Bayessche Formel ohnehin garantiert wird. Dadurch entsteht der geringe Offset an den Ausläufern der Verteilungen, da bei einer Normierung der beiden Dichtefunktionen, die Kurve der Klasse *Rauschen* relativ zur Kurve der Klasse *Signal* weiter an 0 (an der x-Achse) liegen würde.

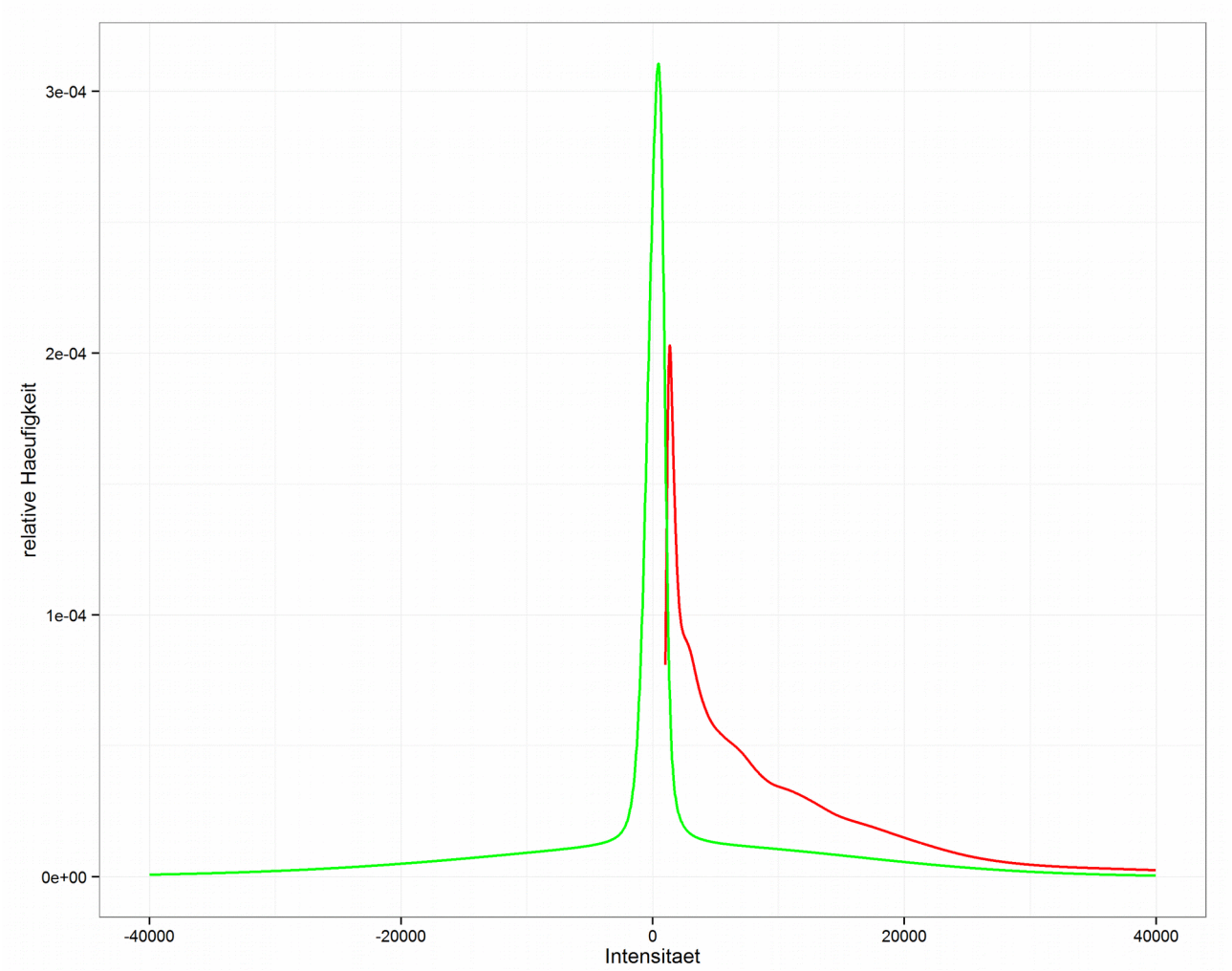

Abb. 39: Die optimalen Wahrscheinlichkeitsdichteverteilungen der Eigenschaft Intensität der Klassen Signal (rot) und Rauschen (grün) mit den Bereichen aus Tabelle 6 und 7. Für die Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).

Bei der shwächsten Diskriminierung (letzter Eintrag) der Hitliste in Tabelle 20 wurden zwar 6689 der 6738 (also 99,27 %) Signale korrekt wieder erkannt, jedoch wurden im Falle der Störsignale nur 8370 von 43698 (19,15 %) korrekt bestimmt. Das heißt, dass 35328 NMR-Signale fälschlicherweise als Nutzsignale interpretiert wurden. Man kann in Abb. 40 erkennen, dass durch das Nehmen der Absolutwerte der Intensitäten dazu führt, dass die relativen Häufigkeiten der negativen Intensitäten verschwinden und ihren Beitrag zusätzlich bei den positiven relativen Häufigkeiten einbringen und dadurch die Verteilung des Rauschens anhebt und damit den Abstand zur Signal-Verteilung verringert. Dies führt zu einer schwächeren Diskriminierung als in Abb. 39, da sich die Verteilungen zu wenig

unterscheiden und der Unterschied beider Kurven vor allen im negativen Bereich der Intensität wegfällt.

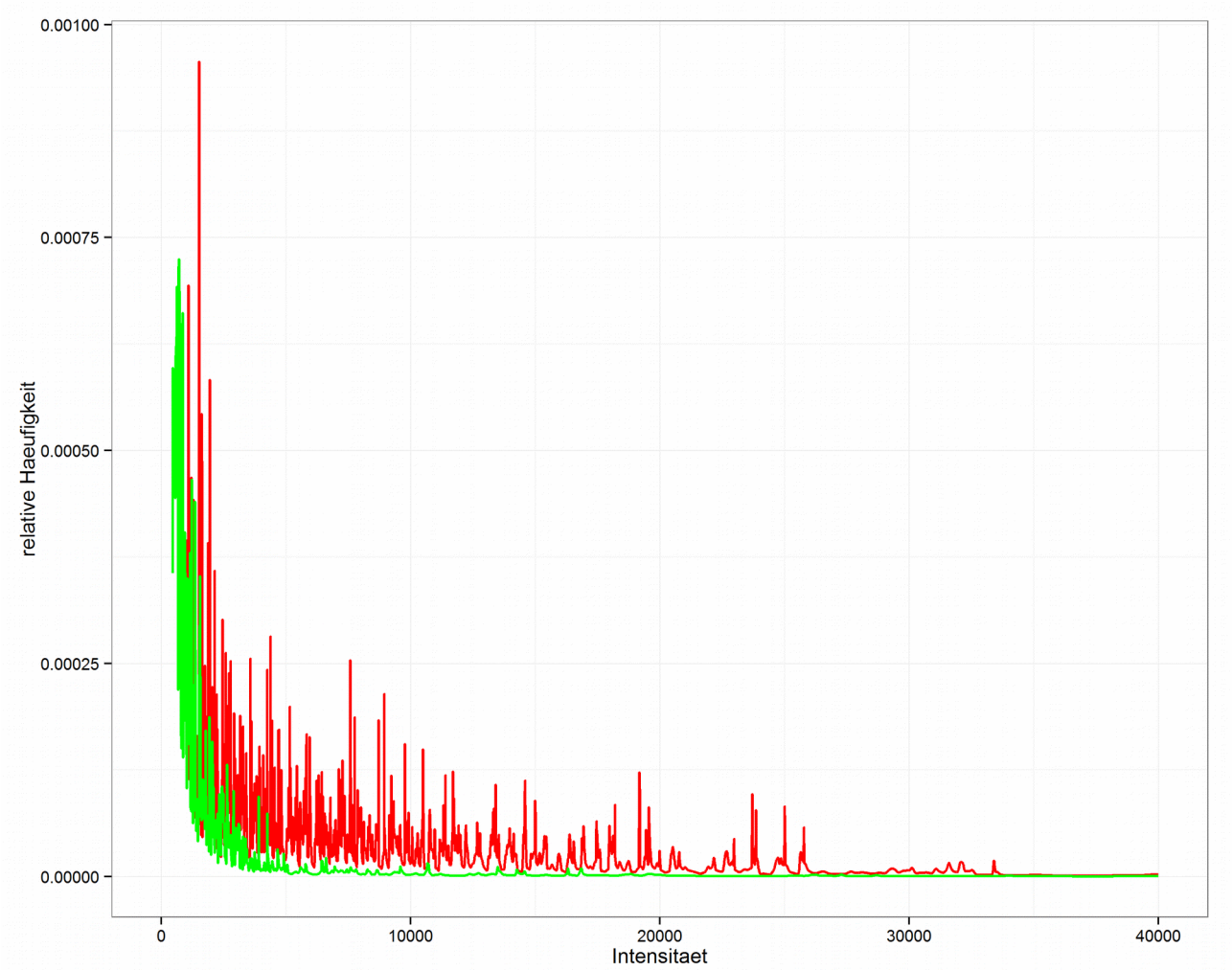

*Abb. 40: Die am schlechtesten diskriminierenden Wahrscheinlichkeitsdichteverteilungen der Eigenschaft **Intensität** der Klassen Signal (rot) und Rauschen (grün) mit den Bereichen aus Tabelle 6 und 7. Beide wurden mit dem kleinsten möglichen Glättungsfaktor von 1 geglättet. Die beiden Wahrscheinlichkeitsdichteverteilungen weisen eine starke Überlappung auf. Zur besseren Visualisierung der signifikanten Bereiche der Verteilungen wurden nur Intensitäten zwischen 0 und 40.000 dargestellt, da die Ausläufer der Verteilungen sehr nahe an 0 liegen. Für die Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).*

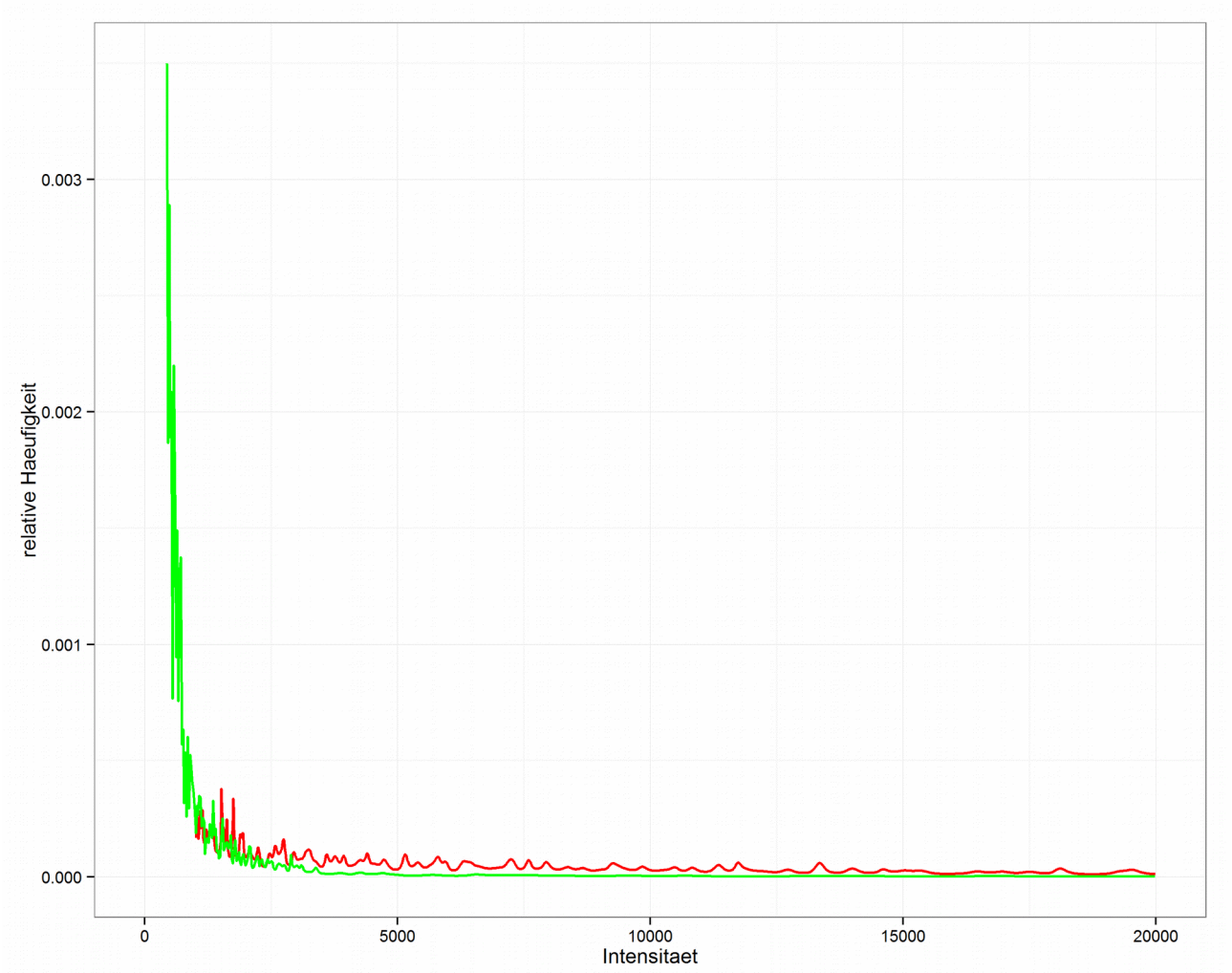

*Abb. 41: Die Wahrscheinlichkeitsdichteverteilungen der Eigenschaft **Intensität** des ursprünglichen Bayes-Picking der Klassen Signal (rot) und Rauschen (grün) mit den Bereichen aus Tabelle 6 und 7. Beide wurden mit dem nicht variierbaren Glättungsfaktor von 5 geglättet. Die Wahrscheinlichkeitsdichteverteilungen weisen eine starke Überlappung der beiden Klassen auf. Zur besseren Visualisierung der signifikanten Bereiche der Verteilungen wurden nur Intensitäten zwischen 0 und 40.000 dargestellt, da die Ausläufer der Verteilungen sehr niedrige relative Häufigkeiten aufweisen. Für die Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).*

Betrachtet man die Ergebnisse der Diskriminierung der Eigenschaft Intensität der ursprünglichen Version des Bayesschen Peak-Picking, muss beachtet werden:

- die Daten aus den Rohdatensatz wurde stets als Absolutwerte genommen
- keine Reskalierung der Rohdaten
- fester *Glättungsfaktor* von 5 bei den geglätteten Verteilungen

In der *Hitliste* belegte diese Parameterkonfiguration Platz 29 von den 40 Diskriminierungen und konnte 6623 von 6738 Signalen und 26874 von den 43698 Störsignalen korrekt wiedererkennen. Alle weiteren Eigenschaften werden analog behandelt.

*Tabelle 21: Ergebnisse der Diskriminierung aller **einzelnen** Eigenschaften unter Verwendung der **geglätteten Verteilungen** mit den optimalen Berechnungsparametern, Skalierungen und Glättungsfilter bei der Verwendung der Klassen **Signal und Rauschen**. Die Ergebnisse wurden anhand der ersten Einträge der Hitlisten je Eigenschaft zusammengefasst. Das Ergebnis der Eigenschaft Intensität aus vorhergegangenen Beispiel findet sich in dieser Tabelle ebenfalls wieder.*

Eigenschaft	Mittelwert in Prozent aus richtig positiv und richtig negativ	Prozent der wieder-erkannten **Signale** richtig positiv (%)	Prozent der wieder-erkannten **Störsignale** richtig negativ (%)	Summe richtig positiv und richtig negativ	Anzahl der wieder-erkannten **Signale** richtig positiv	Anzahl der wieder-erkannten **Störsignale** richtig negativ
Verhältnis Intensität zu lokalen Rauschen	95,83	95,44	96,22	48477	6431	42046
Intensität	94,82	98,71	90,94	46389	6651	39738
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rausch	94,56	93,9	95,23	47939	6327	41612
Volumen	92,81	94,81	90,81	46068	6388	39680
Verhältnis Intensität zur aufsummierten Linienbreite	91,44	96,08	86,79	44401	6474	37927
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w1	91,08	94,76	87,4	44578	6385	38193
Verhältnis der Peakintensität zur Linienbreite w1	91,07	95,79	86,34	44185	6454	37731
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w2	89,18	93,16	85,19	43505	6277	37228
Verhältnis der Peakintensität zur Linienbreite w2	88,84	91,67	86,01	43762	6177	37585
Pseudodistanz	84,79	86,21	83,37	42239	5809	36430
Linienbreite w2	84,27	83,53	85,01	42774	5628	37146
Volumengrundfläche	84,04	79,53	88,55	44053	5359	38694
Linienbreite Summe (w1+w2)	83,89	91,85	75,93	39369	6189	33180
Verhältnis der Peakintensität zur Volumengrundfläche	83,77	87,87	79,67	40736	5921	34815
multivariates diskriminiertes Volumen (Bayes früher)	83,16	79,52	86,79	43285	5358	37927
Breite der Volumengrundfläche w2	82,08	76,82	87,34	43344	5176	38168
Linienbreite w1	81,72	91,76	71,68	37507	6183	31324
Äußere Symmetrie	81,16	77,55	84,77	42268	5225	37043
Kreuz-Rauschen	80,33	80,68	79,98	40385	5436	34949
Breite der Volumengrundfläche w1	76,75	69,95	83,56	41228	4713	36515
Volumenfehler	71,12	51,91	90,32	42966	3498	39468
Innere Symmetrie	68,96	77	60,92	31809	5188	26621
Abstand Schwerpunkt zum Maximum	55,02	64,62	45,42	24203	4354	19849

*Tabelle 22: Ergebnisse der Diskriminierung aller **einzelnen** Eigenschaften unter Verwendung der **geglätteten Verteilungen** mit den optimalen Berechnungsparametern, Skalierungen und Glättungsfilter bei der Verwendung der **drei Klassen Signal, Rauschen und Wasser**. Die Ergebnisse wurden anhand der ersten Einträge der Hitlisten je Eigenschaft zusammengefasst.*

Eigenschaft	Mittelwert in Prozent aus richtig positiv und richtig negativ	Prozent der wieder-erkannten **Signale** richtig positiv (%)	Prozent der wieder-erkannten **Störsignale** richtig negativ (%)	Summe richtig positiv und richtig negativ	Anzahl der wieder-erkannten **Signale** richtig positiv	Anzahl der wieder-erkannten **Störsignale** richtig negativ
Verhältnis Intensität zu lokalen Rauschen	95,82	95,5	96,13	48441	6435	42006
Intensität	95,33	99,99	90,67	46357	6737	39620
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rausch	94,67	93,6	95,73	48141	6307	41834
Volumen	92,81	94,81	90,81	46071	6388	39683
Verhältnis Intensität zur aufsummierten Linienbreite	91,51	95,21	87,82	44791	6415	38376
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w1	91,12	94,57	87,67	44682	6372	38310
Verhältnis der Peakintensität zur Linienbreite w1	91,15	94,88	87,42	44593	6393	38200
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w2	89,28	93,68	84,88	43401	6312	37089
Verhältnis der Peakintensität zur Linienbreite w2	88,81	91,04	86,59	43971	6134	37837
Pseudodistanz	84,82	85,9	83,75	42384	5788	36596
Linienbreite w2	84,98	88,33	81,62	41619	5952	35667
Volumengrundfläche	84,04	79,53	88,55	44054	5359	38695
Linienbreite Summe (w1+w2)	83,89	91,85	75,93	39369	6189	33180
Verhältnis der Peakintensität zur Volumengrundfläche	83,77	87,87	79,67	40735	5921	34814
multivariate diskriminiertes Volumen (Bayes früher)	83,35	78,52	88,17	43818	5291	38527
Breite der Volumengrundfläche w2	82,08	76,82	87,34	43344	5176	38168
Linienbreite w1	82,13	94,57	69,7	36830	6372	30458
Äußere Symmetrie	81,17	77,28	85,05	42373	5207	37166
Kreuz-Rauschen	80,34	81,14	79,54	40225	5467	34758
Breite der Volumengrundfläche w1	76,76	69,95	83,56	41229	4713	36516
Volumenfehler	71,06	51,99	90,14	42892	3503	39389
Innere Symmetrie	68,93	76,31	61,55	32037	5142	26895
Abstand Schwerpunkt zum Maximum	55,05	61,96	48,13	25209	4175	21034

In Tabelle 21 sind alle Diskriminierungen aufgeführt, welche nur zwei Klassen (*Signal* und *Rauschen*) verwendeten und in Tabelle 22 die Diskriminierungen, bei deren drei Klassen verwendet wurden. Dazu wurden alle Bayesschen Wahrscheinlichkeiten aller NMR-Signale (also Signale und Störsignale) berechnet. Da das Vorwissen durch eine bereits erfolgte Zuordnung existierte, welche NMR-Signale wirklich Signale aus dem Protein waren, konnte evaluiert werden, wie gut die Wiedererkennung erfolgte.

Da jeder Datensatz einer Eigenschaft durch die Variation der *Skalierung* und der Variation der *Parameter der Berechnungsmethode* verschiedene Ergebnisse der Diskriminierung lieferten, wurde diese mit dem höchsten Mittelwert aus prozentualer Wiedererkennung der Signale und der Störsignale absteigend sortiert (siehe Spalte 2 aus Tabelle 21 und 22). Die höchste prozentuale mittlere Wiedererkennung legte die zu verwendenden *Skalierungen* und *Parameter der Berechnungsmethoden* der Eigenschaften fest und wurde in Tabelle 23 zusammengefasst.

Da die Berechnung des Symmetrie-Kriteriums erheblich mehr Variationen der Berechnungsmöglichkeiten aufwies, wurden die zusätzlich verwendeten Parameter separat in Tabelle 24 aufgeführt. Im Falle der *inneren Symmetrie* gibt es keinen zu variierenden Suchradius, da kein diagonaler symmetrischer Partner gesucht werden musste.

Auch für die Diskriminierung durch die Eigenschaft *gaußsche Signalwahrscheinlichkeit* wurden zusätzliche Berechnungsvariationen durchgeführt, welche nicht in Tabelle 23 aufgeführt wurden. Hier wurde zusätzlich zur Wahl einer festen Größe zur Einbeziehung der Nachbarschaftspixel der Position des Signals die Verwendung einer von der Volumengrundfläche abhängigen Nachbarschaft gewählt. Diese erreichte aber nicht die stärkste Diskriminierung. Das beste Ergebnis des prozentualen Mittelwerts der korrekt wiedererkannten Signale und Störsignale wurde bei dieser Eigenschaft eine feste Größe (wie bisher) der Nachbarschaftsbereiche von 3x3 zur Berechnung verwendet und nicht die Idee der Abhängigkeit der Nachbarschaftsbereiche von der digitalen Auflösung und von dem volumengrundflächenabhängigen Gaußfilter aus Formel 18. Jedoch wurde nicht die vormals verwendete **normale kumulative Dichtefunktion**, sondern die **logarithmische kumulative Dichtefunktion** zur Berechnung der gaußschen Wahrscheinlichkeit (17) angewendet.

Tabelle 23: Übersicht der verwendeten Berechnungsmethoden und Skalierungen der besten Diskriminierung bei der Verwendung nur einer Eigenschaft im Falle von zwei Klassen (Signal und Rauschen) und von drei Klassen (Signal, Rauschen, Wasser).

	Segmentierungslevel		Absolut-werte	Reskalierung		verwendeter Glättungsfaktor	
Eigenschaft	Signal und Rauschen	Signal, Rauschen und Wasser		Signal und Rauschen	Signal, Rauschen und Wasser	Signal und Rauschen	Signal, Rauschen und Wasser
Verhältnis Intensität zu lokalen Rauschen	-	-	n	logarithmisch	logarithmisch	2	1
Intensität	-	-	n	keine	logarithmisch	1 Maximum	1
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rausch	-	-	n	keine	keine	1 Maximum	4
Volumen	0,001	0,001	n	keine	keine	4 Maxima	4 Maxima
Verhältnis Intensität zur aufsummierten Linienbreite	-	-	n	logarithmisch	logarithmisch	10	3
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w1	0,5	0,5	n	logarithmisch	logarithmisch	2 Maxima	4
Verhältnis der Peakintensität zur Linienbreite w1	-	-	n	logarithmisch	logarithmisch	10	5
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w2	0,5	0,5	n	logarithmisch	logarithmisch	10	2
Verhältnis der Peakintensität zur Linienbreite w2	-	-	n	logarithmisch	logarithmisch	2 Maxima	2 Maxima
Pseudodistanz	0,001	0,001	n	keine	keine	1 Maximum	10
Linienbreite w2	-	-	n	logarithmisch	logarithmisch	3	3 Maxima
Volumengrundfläche	0,001	0,001	n	logarithmisch	logarithmisch	3 Maxima	4 Maxima
Linienbreite Summe (w1+w2)	-	-	n	logarithmisch	logarithmisch	3 Maxima	4 Maxima
Verhältnis der Peakintensität zur Volumengrundfläche	0,001	0,001	n	keine	keine	4 Maxima	4 Maxima
multivariates diskriminiertes Volumen (Bayes früher)	0,1	0,1	n	logarithmisch	logarithmisch	3 Maxima	3 Maxima
Breite der Volumengrundfläche w2	0,001	0,001	n	logarithmisch	logarithmisch	1 Maximum	4 Maxima
Linienbreite w1	-	-	n	logarithmisch	logarithmisch	4 Maxima	3 Maxima
Äußere Symmetrie	0,5	0,5	n	keine	keine	1 Maximum	1 Maximum
Kreuz-Rauschen	0,001	0,001	n	logarithmisch	logarithmisch	10	5
Breite der Volumengrundfläche w1	0,001	0,001	n	logarithmisch	logarithmisch	2 Maxima	2 Maxima
Volumenfehler	0,001	0,1	n	logarithmisch	logarithmisch	4 Maxima	3 Maxima
Innere Symmetrie	0,001	0,2	n	keine	keine	4 Maxima	4 Maxima
Abstand Schwerpunkt zum Maximum	0,001	0,001	n	logarithmisch	logarithmisch	2 Maxima	2 Maxima

Die Definitionen der jeweiligen Eigenschaften sind in Abschnitt 2.6.5.2 detailliert aufgeführt.

Tabelle 24: Die verwendeten Berechnungsmethoden für das Symmetrie-Kriterium der inneren Symmetrie und der äußeren Symmetrie, welche die beste Diskriminierung erlangten.

	Äußere Symmetrie		Innere Symmetrie	
Berechnungsmethode	Signal und Rauschen	Signal, Rauschen und Wasser	Signal und Rauschen	Signal, Rauschen und Wasser
Patternkorrektur	keine	keine	Differenz Pixelintensität und lokales Rauschen	Differenz Pixelintensität und lokales Rauschen
Symmetrie-Suchradius	0	1	-	-
Spline-Typ	größte gemeinsame Größe	größte gemeinsame Größe	kleinste gemeinsame Größe	kleinste gemeinsame Größe
Verwendetes Kosinuskriterium	zwei Vektoren	zwei Vektoren	Mitterlwert der Cosinuswerte aller Spalten und Zeilen	Mitterlwert der Cosinuswerte aller Spalten und Zeilen
Ausschneidung der Volumengrundfläche bei Segmentierung	0,5	0,5	0,001	0,2
Patternausrichtung bezüglich der Peakposition	Vergrößerung des Patterns, bis Peak im Zentrum	Vergrößerung des Patterns, bis Peak im Zentrum	Unveränderter Integrations-bereich als Patterngröße	Unveränderter Integrations-bereich als Patterngröße
Minimale Patterngröße, falls die Grundfläche zu klein wird	5	5	3	3

Die in diesem Abschnitt evaluierten Parameter zur Bestimmung der Datensätze und die Skalierung stellt die Basis für die folgenden Abschnitte dieses Kapitels dar. Die Einstellung der *Parameter* und *Skalierungen* wurden gespeichert und können für andere Spektren von AUREMOL geladen und verwendet werden. Diese Einstellungen werden im Folgenden als *optimaler Parametersatz* bezeichnet.

3.4.2 Die Erzeugung der theoretischen Verteilungen basierend auf dem *optimalen Parametersatz* zur Berechnung der Datensätze aller Eigenschaften

Zur Erzeugung der theoretischen Verteilungen benötigt man den optimalen Glättungsfaktor nicht. Daher wird für die Berechnung der Datenbasis zur Generierung der theoretischen Verteilungen lediglich der *optimale Parametersatz*, welche auch den optimalen Glättungsfaktor aufweist, verwendet (siehe Abschnitt 3.4.1).

Da es nicht möglich war, aus den *geglätteten Verteilungen* eine Funktion zu erhalten, wurde versucht, Verteilungen zu finden, welche die Definition einer Funktion und deren Parameter erlauben, welche die *geglätteten Verteilungen* möglichst genau wieder geben. Diese Funktionen können dann auf ein anderes noch nicht zugeordnetes Spektrum gleichen Typs übertragen und angewendet werden. Dies wurde mit der Methode des *Simulated Annealing* realisiert, welche für jede Verteilung einer Klasse die Zielfunktionen einer Normalverteilung (N) und/oder einer logarithmischen Normalverteilung (LOGN) zu optimieren versucht (siehe Kapitel 2.7.1).

Die Ergebnisse aller Verteilungen der Eingenschaften mit deren *optimalen Parametersatz* wurden in eine Konfigurationsdatei (Auszug aus der Datei „Simualted Annealing.ini“ siehe Tabelle 20) im Verzeichnis des verwendeten Spektrums abgespeichert, so dass diese jederzeit zur Berechnung der Wahrscheinlichkeiten ausgelesen und verwendet werden konnte.

Da jede Klasse eine andere Verteilungsfunktion verwenden kann, mussten alle möglichen Kombinationen aus den Verteilungen für eine Eigenschaft getestet werden. Im Falle der *Intensität* lagen somit 36 Kombinationen aus N und LOGN in einer weiteren *Hitliste* vor. Jeder Eintrag dieser *Hitliste* stellt wieder eine komplette Diskriminierung analog zu vorangegangenen *Hitlisten* aus Abschnitt 3.4.1 dar. Im Folgenden soll der Vorgang **anhand des Beispiels der Eigenschaft *Intensität*** veranschaulicht werden. Der Ablauf ist für alle Eigenschaften analog.

*Tabelle 25: Die Übersicht aller möglichen Kombinationen der Verteilungsfunktionen N und LOGN (hier als Beispiel der Eigenschaft **Intensität**) zur Diskriminierung bei Verwendung des optimalen Parametersatzes.*

Mittelwert in Prozent aus richtig positiv und richtig negativ	Prozent der wieder-erkannten Signale richtig positiv (%)	Prozent der wieder-erkannten Störsignale richtig negativ (%)	Summe richtig positiv und richtig negativ	Anzahl der wieder-erkannten Signale richtig positiv	Anzahl der wieder-erkannten Störsignale richtig negativ	verwendeter Schwellwert der Wahrschein-lichkeit	verwendete Zielfunktion der Klasse Signal	verwendete Zielfunktion der Klasse Rauschen
95,23	100	90,47	46271	6738	39533	0,253	LOGNLOGN	NN
95,23	100	90,47	46271	6738	39533	0,238	NLOGN	NN
95,23	100	90,47	46271	6738	39533	0,229	LOGN	NN
95,23	100	90,47	46271	6738	39533	0,239	LOGNN	NN
95,23	100	90,46	46266	6738	39528	0,199	LOGN	LOGNN
95,23	100	90,46	46266	6738	39528	0,218	NLOGN	NLOGN
95,23	100	90,46	46266	6738	39528	0,221	LOGNLOGN	LOGNN
95,23	100	90,46	46266	6738	39528	0,219	LOGNN	NLOGN
95,23	100	90,46	46266	6738	39528	0,207	NLOGN	LOGNN
95,23	100	90,46	46266	6738	39528	0,232	LOGNLOGN	NLOGN
95,23	100	90,46	46266	6738	39528	0,208	LOGNN	LOGNN
95,23	100	90,46	46266	6738	39528	0,209	LOGN	NLOGN
93,46	100	86,91	44716	6738	37978	0,505	LOGNLOGN	N
92,21	93,19	91,22	46142	6279	39863	0,493	N	N
90,68	88,79	92,57	46434	5983	40451	0,5	LOGNLOGN	LOGNLOGN
90,41	96,35	84,46	43401	6492	36909	0,159	NN	NN
89,32	85,6	93,03	46419	5768	40651	0,501	LOGN	LOGNLOGN
82,94	70,96	94,92	46258	4781	41477	0,5	LOGN	LOGN
82,62	95,52	69,72	36902	6436	30466	0,56	NLOGN	N
82,55	96,08	69,02	36636	6474	30162	0,557	LOGNN	N
81,93	95,49	68,37	36312	6434	29878	0,094	N	NN
81,59	68,24	94,94	46084	4598	41486	0,676	N	NLOGN
81,3	70,45	92,14	45010	4747	40263	0,635	NN	N
81,24	67,42	95,06	46081	4543	41538	0,678	N	LOGNN
80,86	100	61,71	33706	6738	26968	0,471	LOGN	N
80,12	64,5	95,75	46185	4346	41839	0,5	LOGNLOGN	LOGN
79,82	65,14	94,5	45683	4389	41294	0,68	NN	NLOGN
79,12	63,58	94,65	45646	4284	41362	0,682	NN	LOGNN
64,48	100	28,96	19393	6738	12655	0,111	NLOGN	LOGNLOGN
64,48	100	28,96	19391	6738	12653	0,112	LOGNN	LOGNLOGN
64,36	100	28,72	19290	6738	12552	0,162	LOGNN	LOGN
64,36	100	28,72	19290	6738	12552	0,162	NLOGN	LOGN
63,83	98,14	29,51	19508	6613	12895	0,104	NN	LOGN
63,43	97,28	29,57	19477	6555	12922	0,06	N	LOGN
63,41	97,24	29,57	19475	6552	12923	0,07	NN	LOGNLOGN
62,99	96,36	29,62	19435	6493	12942	0,04	N	LOGNLOGN

Die beste Kombination der Verteilungsfunktionen zur Diskriminierung der Eigenschaft *Intensität* aus Tabelle 25 für die Klasse *Rauschen* war hier durch die Anwendung der zuvor generierten NN-Verteilung gegeben (Also die Kombination zweier N-Verteilungen). Im Falle der Klasse *Signal* diskriminierten die Verteilungen LOGNLOGN, LOGN, NLOGN und LOGNN am besten.

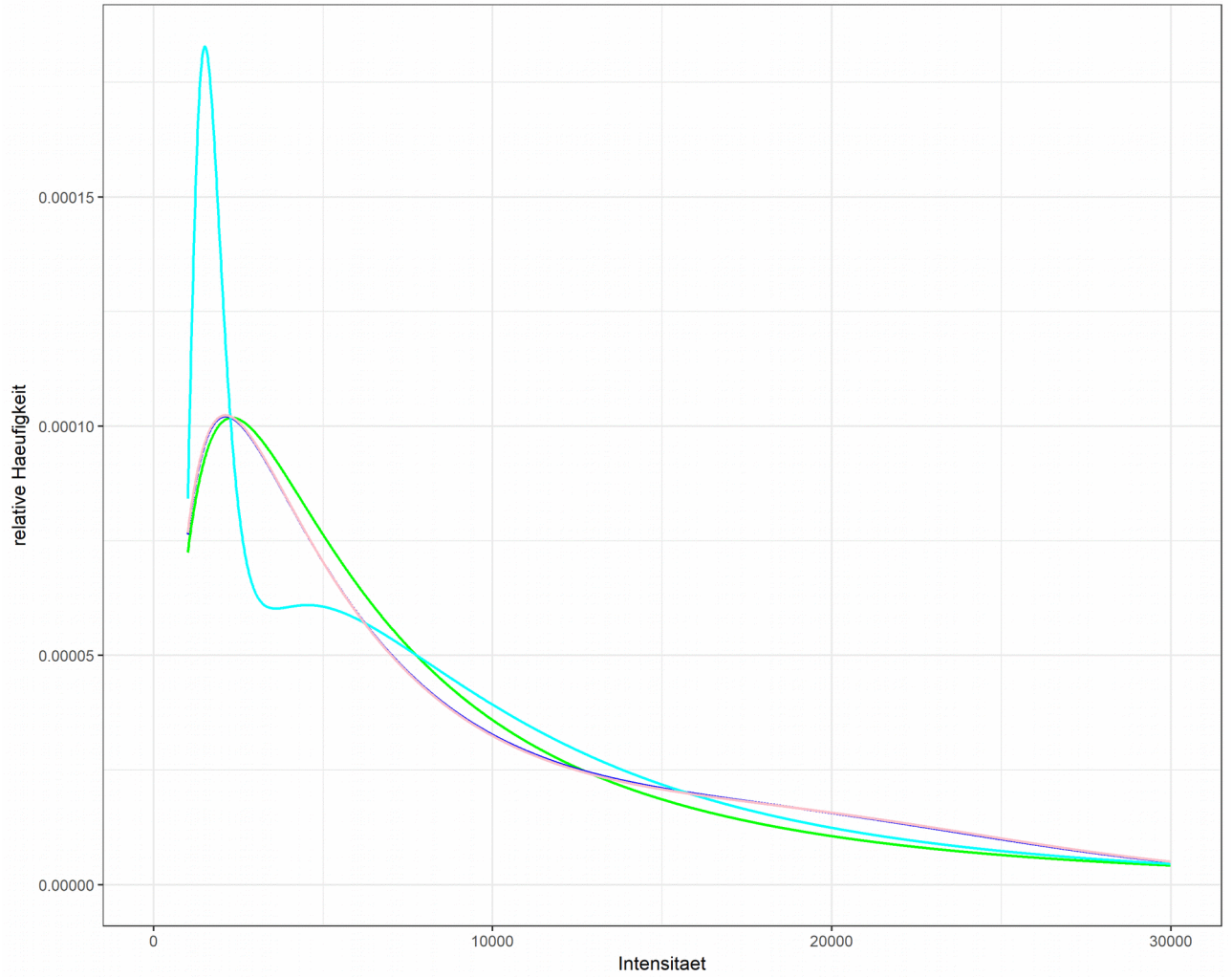

Abb. 42: Wahrscheinlichkeitsdichteverteilungen der besten optimierten Zielfunktionen LOGNLOGN (cyan), LOGN (grün), NLOGN (blau) und LOGNN (pink) der Klasse Signal für die Eigenschaft Intensität. ***Die Verteilung NLOGN (blau) wird durch die Verteilung LOGNN (pink) stark überdeckt, da sie nahezu denselben Verlauf hat****. Zur Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx aus Abschnitt 2.2 zugrunde.*

Der *optimierte Parametersatz* wurde zur Darstellung der Verteilungen der Intensität bezüglich der Klasse *Signal* in Abb. 42 aus der erstellten Konfigurationsdatei (siehe Tabelle 26 und Tabelle 27) ausgelesen. In dieser Datei befinden sich sämtliche Parameter aller optimierten Zielfunktionen.

Abb. 43 zeigt die optimierte Zielfunktion LOGNLOGN im Vergleich zu der geglätteten Verteilung mit dem kleinsten möglichen Glättungsfaktor von 1 noch einmal, um darzustellen, wie Nahe diese Zielfunktion an der geglätteten Verteilung liegt.

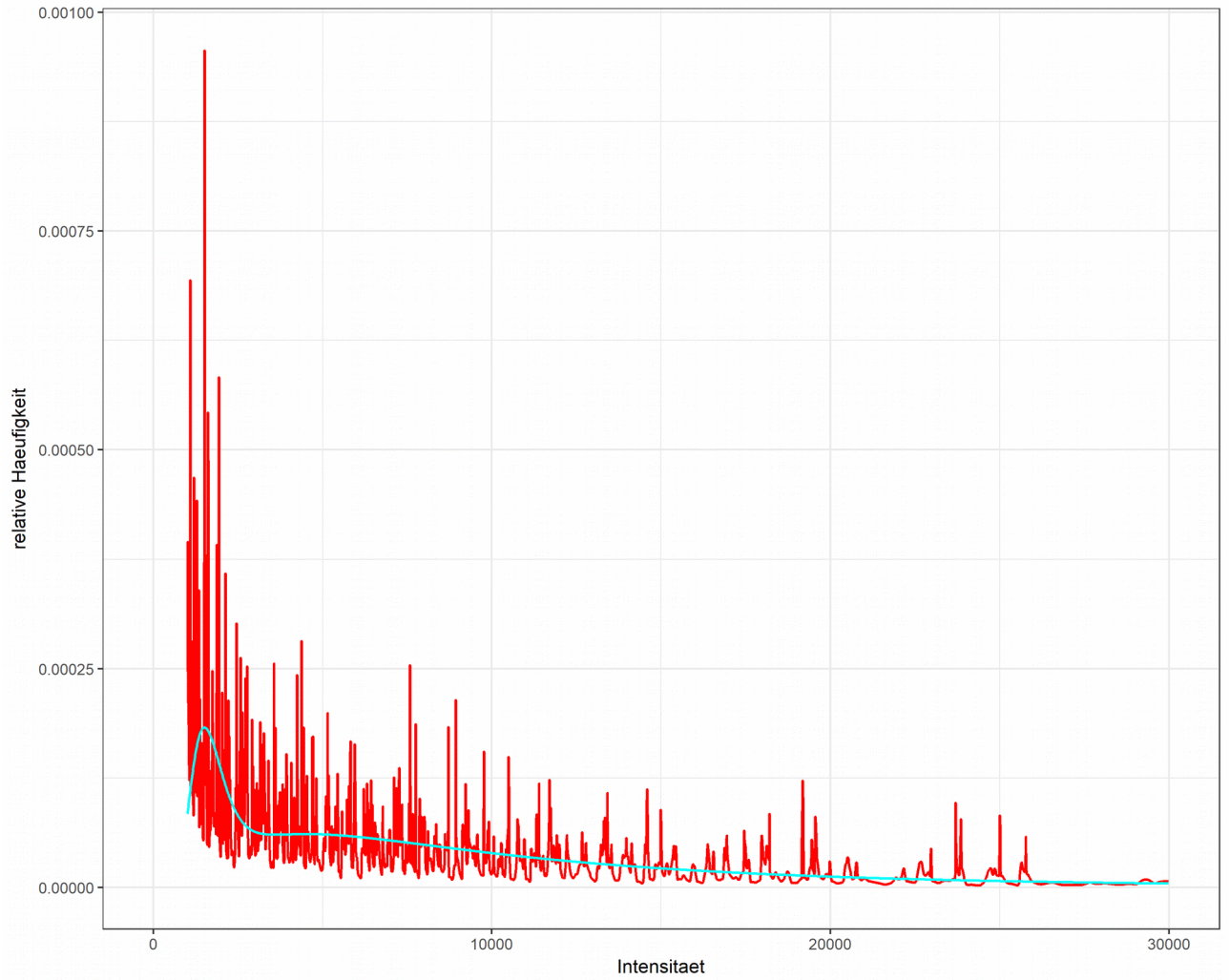

Abb. 43: Wahrscheinlichkeitsdichteverteilungen der besten optimierten Zielfunktion LOGNLOGN (cyan) der Klasse Signal für die Eigenschaft Intensität. Als Vergleich wurde die geglättete Wahrscheinlichkeitsdichteverteilung mit dem kleinsten möglichen Glättungsfaktor 1 (rot) dargestellt. Für die Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).

Da die zur Optimierung verwendeten Daten wie bereits im Kapitel zuvor erläutert (siehe Tabelle 23 und 24) berechnet wurden, mussten diese Parameter zur Erstellung der Rohdaten in dieser Konfigurationsdatei festgehalten werden, damit die Rohdaten zu der jeweiligen Verteilung passten. Wurde z. B. für die Klasse *Rauschen* die *Skalierung* so verändert, dass alle Werte ins positive verschoben werden, jedoch die bereits optimierte Funktion durch das nicht verschieben der Daten erstellt worden sein, kommt es zu dem Effekt, dass die theoretische Verteilung nicht mehr zu den Daten passen würde.

Für die Wiedergabe der optimalen Verteilungen in Abb. 42 wurde folgende zuvor mit Simulated Annealing erstellte Konfigurationsdatei (Tabelle 26) verwendet.

Tabelle 26: Auszug der Konfigurationsdatei der optimalen theoretischen Verteilungen der Klasse Signal für die Eingenschaft Intensität.

```
[SignalLOGNIntensity]
Data="8.77336194591818;1.01211469095607;0.999983366811644"
usedminmaxvalue="1006;102423"
rescaletype=0
PropertyName=Intensity
absolute=false
Classname=Signal
reducetopercent=0
ComboTypeID=1
minimizationType=-1

[SignalLOGNLOGNIntensity]
Data="7.39138089469885;9.11366114649422;0.307876568511333;0.816199252635033;0.190853850727634"
usedminmaxvalue="1006;102423"
rescaletype=0
PropertyName=Intensity
absolute=false
Classname=Signal
reducetopercent=0
ComboTypeID=3
minimizationType=-1

[SignalLOGNNIntensity]
Data="8.61776992293932;19171.021157746;0.980028095956409;6817.68191526295;0.857121604461484"
usedminmaxvalue="1006;102423"
rescaletype=0
PropertyName=Intensity
absolute=false
Classname=Signal
reducetopercent=0
ComboTypeID=5
minimizationType=-1

[SignalNLOGNIntensity]
Data="18852.4295167416;8.61692778797394;6939.04730507344;0.978859438557447;0.146217484061191"
usedminmaxvalue="1006;102423"
rescaletype=0
PropertyName=Intensity
absolute=false
Classname=Signal
reducetopercent=0
ComboTypeID=2
minimizationType=-1
```

Die Werte des Schlüssels „Data“ innerhalb einer Sektion, welche die gesamte Verteilung definiert, spiegelt die optimalen Parametersätze wieder. Der Schlüssel ComboTypeID gibt die verwendete Kombination aus den Verteilungstypen N und NLOGN an. Wir betrachten nachfolgend den Fall der LOGNLOGN-Verteilung für die Eigenschaft *Intensität* der Klasse *Rauschen*, welche später auch verwendet wird, da diese an der ersten Stelle in der Hitliste anzutreffen ist (siehe Tabelle 25). Die restlichen Einträge werden von der Datenverarbeitung des Auswerte-Algorithmus benötigt um die Daten unter anderem wieder in die passende Skalierung zu überführen.

Daher gelten für die $f_{LOGNLOGN}$-Formel 29 die optimierten Parameter (Sektion *[Signal**LOGNN**Intensity]* aus Tabelle 26) μ_1=7,3914, σ_1=0,3079, μ_2=9,1137, σ_2=0,8162 und

λ=0,1909. Der Parameter x verbleibt als freier Parameter in diesem Fall für die *Intensitäten*. Betrachtet man in Tabelle 26 die Parameter der Verteilung NLOGN und die der Verteilung LOGNN, fällt auf, dass sich deren μ und σ in erster Näherung jeweils vertauschen und dass das λ von LOGNN annähernd (1- λ) dem von NLOGN entspricht. Die beiden Verteilungen weisen in Abb. 42 einen nahezu identischen Verlauf aus.

Für die Bestimmung der Verteilungen der Klasse *Rauschen* wird analog vorgegangen. In diesem Fall war bei den besten Diskriminierungen zusammen mit denen der Klasse *Signal* ausschließlich die theoretische NN-Verteilung beteiligt.

Tabelle 27: Auszug der Konfigurationsdatei der optimalen Verteilung für die Eigenschaft Intensität der Klasse Rauschen.

```
[NoiseNNIntensity]
Data="875.763593628275;183.466754422639;50124.5870350426;1150.91460919378;0.0996936825336175"
usedminmaxvalue="-229900;146909"
rescaletype=0
PropertyName=Intensity
absolute=false
Classname=Noise
reducetopercent=0
ComboTypeID=4
minimizationType=-1
```

Somit findet für die Klasse *Rauschen* die Formel F_{NN} (26) Anwendung. Die optimierten Parameter dieser Funktion waren dabei μ_1=875,7636, σ_1=50124,5870, μ_2=183,4668, σ_2=1150,9146 und λ=0,0997. Die Abb. 44 zeigt den Verlauf dieser NN-Verteilung (grün) mit den Parametern aus der entsprechenden Konfigurationsdatei (Tabelle 27).

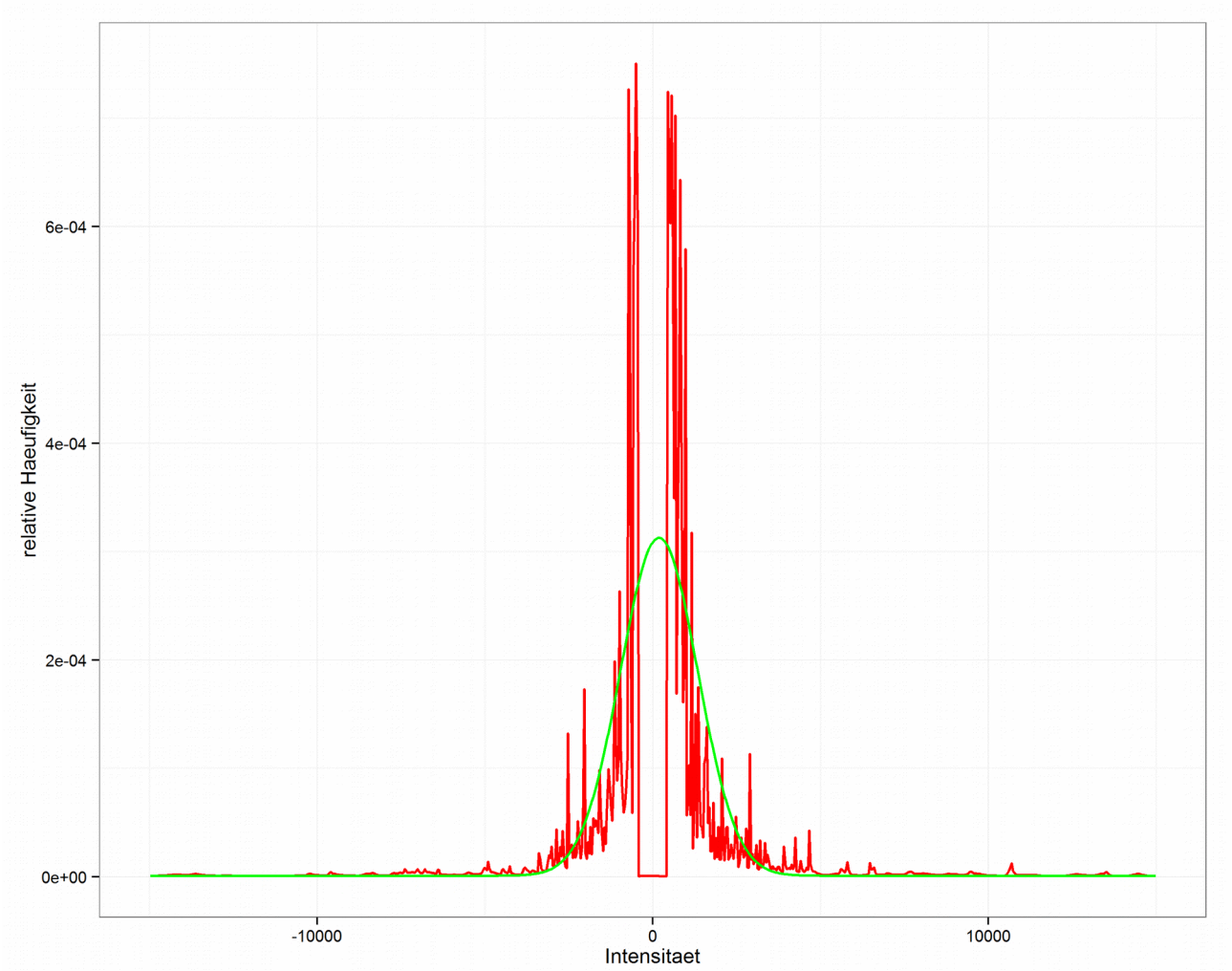

*Abb. 44: Wahrscheinlichkeitsdichteverteilungen der optimierten Zielfunktion **NN** der Klasse **Rauschen** (grün) für die Eigenschaft Intensität. Als Vergleich wurde die geglättete Wahrscheinlichkeitsdichteverteilung mit dem kleinsten möglichen Glättungsfaktor 1 (rot) dargestellt. Für die Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).*

Die Fehlstelle um die Intensität 0 in Abb. 44 wurde durch die Optimierung der SA-Parameter μ_1, σ_1, μ_2, σ_2 und λ der NN-Verteilung „herausgefittet“. Das gleiche Bild lieferte auch die Anwendung der geglätteten Verteilungen aus Abb. 39 (grüne Kurve). Diese Fehlstelle entstand durch den verwendeten Schwellwert der Intensität durch das Peak-Picking, bei dem NMR-Signale (also Signale und Störsignale) mit einer Intensität unter (bzw.bei negativen Intensitäten über) diesem Schwellwert nicht in die Peakliste mit aufgenommen wurden.

Die Auswirkungen sind dabei die folgenden:

- Die grüne Kurve tendiert zu einer geringeren relativen Häufigkeit um die 0-Stelle durch das Fitten über die Abbruchstellen. Dies hat zur Folge, dass Intensitäten eine geringere Wahrscheinlichkeit liefern, dass sie Störsignale sind.
- Würde man den Schwellwert beim Peak-Picken auf 0 setzen, würden alle Extrema in die Peakliste mit aufgenommen werden, was die Anzahl der Störsignale enorm erhöhen würde und dadurch die Kurve besser der Realität angleicht. Dies hat aber den Effekt, dass die Datenmenge für das Softwarepaket derart erhöht werden würde und so der Speicher zum einen nicht ausreichen würde und zum anderen zusätzlich benötigte Rechenzeit für den Optimierungslauf und die Erstellung der geglätteten Verteilungen enorm ansteigen würde.
- Da in der Regel keine NMR-Signale mit Intensität 0 vorkommen, würde die 0-Stelle in der Verteilung zwar wieder eine sehr hohe relative Häufigkeit aufweisen, jedoch würde dieser Wert ohnehin nie abgefragt werden. Bei einer breiten Verteilung käme dies auch kaum zu tragen, da die Nachbarhäufigkeiten keinen signifikanten Unterschied zu dem der ‚falschen' 0-Stelle haben. Lediglich im Extremfall, bei dem die Verbreiterung der Verteilung sehr gering ist und dadurch die Kurve ein sehr schmales Maximum aufweist, ist der Unterschied zu den Nachbarhäufigkeiten der 0-Stelle hoch.

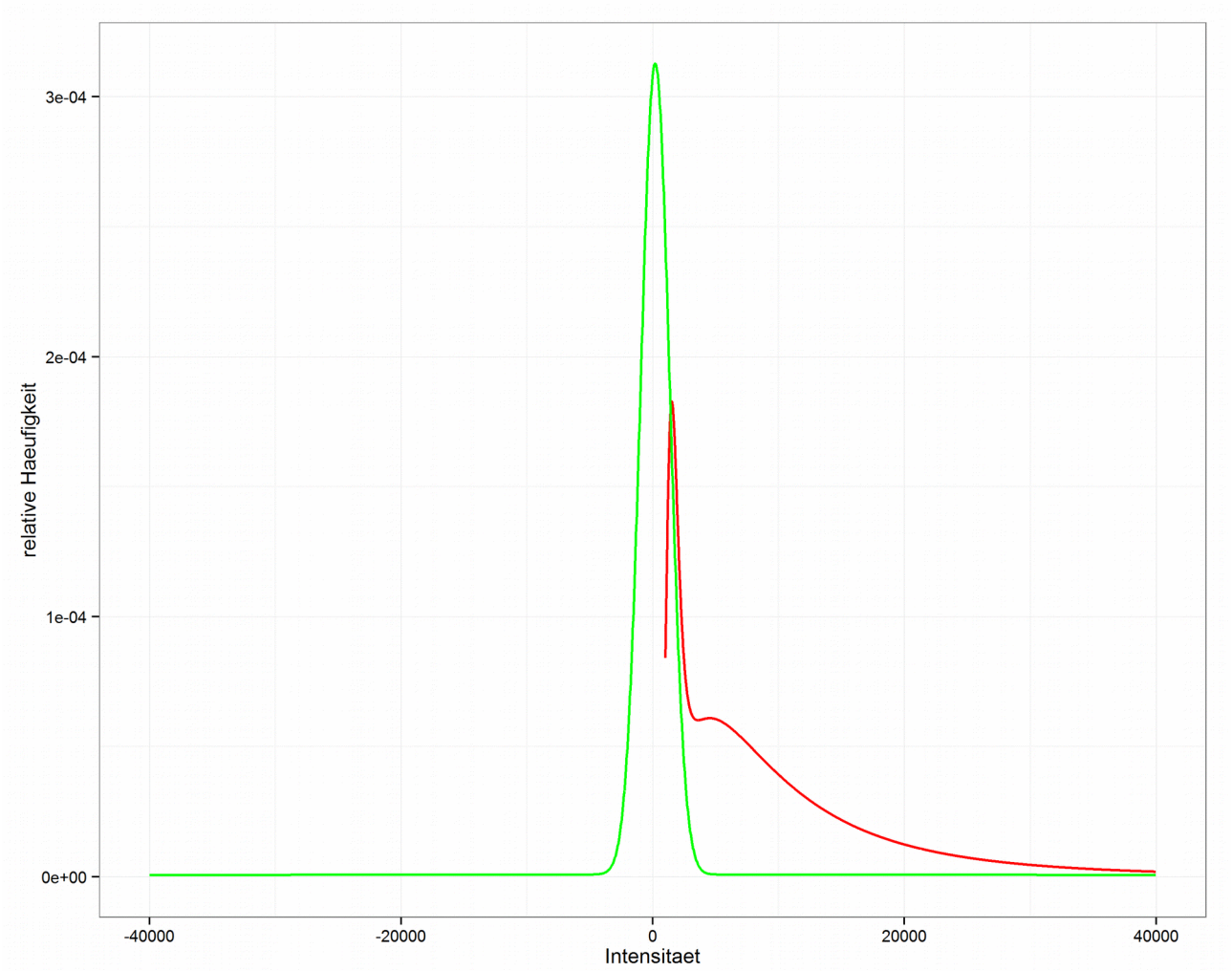

Abb. 45: Wahrscheinlichkeitsdichteverteilungen der optimierten Zielfunktion. Dargestellt sind die Verteilungen LOGNLOGN der Klasse Signal (rot) und der Klasse Rauschen (grün) mit der Zielfunktion NN für die Eigenschaft Intensität. Für die Generierung der Datenbasis für diese Abbildung lag das experimentelle Spektrum des Proteins PfTrx zugrunde (siehe Abschnitt 2.2).

Betrachtet man die rote Kurve aus Abb. 45, erkennt man, dass diese nahe bei der Intensität 1006 abbricht. Dieser Effekt kommt dadurch zustande, da zur Darstellung der Kurve lediglich die Intensitäten der vorhandenen Signale aus der Peakliste (also der Datensatz zur Eingenschaft Intensität) verwendet wurden. Die Kurve würde bei einer künstlichen Erweiterung des Datensatzes durch negative Signal-Intensitäten weiter fortgesetzt werden (siehe Schlüssel usedminmaxvalue="1006;102423" aus der Konfigurationsdatei aus Tabelle 26).

Die Ergebnisse der besten Kombinationen aus der jeweiligen *Hitliste* wurden als Zwischenergebnis abgespeichert. Die verwendeten Kombinationen wurden in die Konfigurationsdatei (SAKombos.ini) gespeichert, welche die besten Kombinationen aus allen erstellten Hitlisten der Eigenschaften enthält. Hieraus wurde zu einer jeden Eigenschaft die benötigten Verteilungstypen der Klassen ausgelesen und anschließend aus der Hauptkonfigurationsdatei (SimulatedAnnealing.ini) mit den optimierten SA-Parametern (μ_1, σ_1, μ_2, σ_2 und λ) ergänzt.

Analysiert man nun die Ergebnisse aller Diskriminierungen aus Tabelle 28 fällt auf, dass z. B. die Ergebnisse der Eigenschaft *multivariate Diskriminierung des Volumens bei den verschiedenen Segmentierungen* durch die Verwendung der theoretischen Verteilungen das gleiche Ergebnis (Mittelwert richtig positiv und richtig negativ: 0,832) in Tabelle 28 aufweisen als bei der Verwendung der **geglätteten Verteilung** aus Tabelle 21 (Mittelwert richtig positiv und richtig negativ: 0,832). Dies ist auch in in Abb. 46 ersichtlich, in der sich die theoretischen Verteilungen (blau und cyan) weitgehend mit den geglätteten Verteilungen (rot und grün) decken.

*Tabelle 28: Ergebnisse der Diskriminierung aller Eigenschaften durch Verwendung der **theoretischen Verteilungen** mit der verwendeten besten Kombination der Verteilungsfunktionen (LOGN und/oder N in der letzten Spalte) analog zur Tabelle 21.*

Eigenschaft	Mittelwert in Prozent aus richtig positiv und richtig negativ	Prozent der wieder-erkannten Signale richtig positiv (%)	Prozent der wieder-erkannten Störsignale richtig negativ (%)	Summe richtig positiv und richtig negativ	Anzahl der wieder-erkannten Signale richtig positiv	Anzahl der wieder-erkannten Störsignaleri chtig negativ	Kombinationen der Zielfunktion Signal/ Rauschen
Verhältnis Intensität zu lokalen Rauschen	95,77	95,19	96,35	48517	6414	42103	LOGN/ N
Intensität	95,23	100	90,47	46271	6738	39533	LOGNLOGN/ NN
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rausch	94,63	93,56	95,71	48127	6304	41823	LOGNN/ LOGN
Volumen	92,81	94,81	90,81	46069	6388	39681	LOGNLOGN/ NN
Verhältnis Intensität zur aufsummierten Linienbreite	91,97	96,84	87,1	44588	6525	38063	NN/ LOGNN
Verhältnis der Peakintensität zur Linienbreite w1	91,6	95,81	87,38	44639	6456	38183	NLOGN/ LOGNN
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w1	91,42	95,12	87,72	44741	6409	38332	LOGNN/ LOGNN
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w2	89,63	93,53	85,73	43766	6302	37464	NLOGN/ LOGNN
Verhältnis der Peakintensität zur Linienbreite w2	89,5	92,8	86,19	43917	6253	37664	LOGNN/ NLOGN
Linienbreite w2	85,05	88,51	81,59	41616	5964	35652	LOGN/ N
Pseudodistanz	85,02	87,16	82,87	42086	5873	36213	LOGNLOG/ LOGNLOGN
Volumengrundfläche	84,04	79,53	88,55	44053	5359	38694	LOGNLOGN/ N
Linienbreite Summe (w1+w2)	83,89	91,85	75,93	39369	6189	33180	N/ NLOGN
Verhältnis der Peakintensität zur Volumengrundfläche	83,79	88,02	79,56	40695	5931	34764	NLOGN/ NN
multivariates diskriminiertes Volumen (Bayes früher)	83,16	79,52	86,79	43285	5358	37927	LOGNLOGN/ NN
Linienbreite w1	82,13	94,57	69,7	36830	6372	30458	NN/ NLOGN
Breite der Volumengrundfläche w2	82,08	76,82	87,34	43344	5176	38168	LOGNLOGN/ NN
Äußere Symmetrie	81,13	77,38	84,88	42304	5214	37090	NLOGN/ N
Kreuz-Rauschen	80,42	82,37	78,47	39840	5550	34290	LOGNLOGN/ LOGN
Breite der Volumengrundfläche w1	76,77	69,95	83,59	41238	4713	36525	LOGNN/ LOGNLOGN
Volumenfehler	71,11	51,91	90,31	42963	3498	39465	NN/ NLOGN
Innere Symmetrie	68,83	75,93	61,74	32095	5116	26979	LOGN/ N
Abstand Schwerpunkt zum Maximum	55	51,63	58,36	28983	3479	25504	NLOGN/ LOGNN

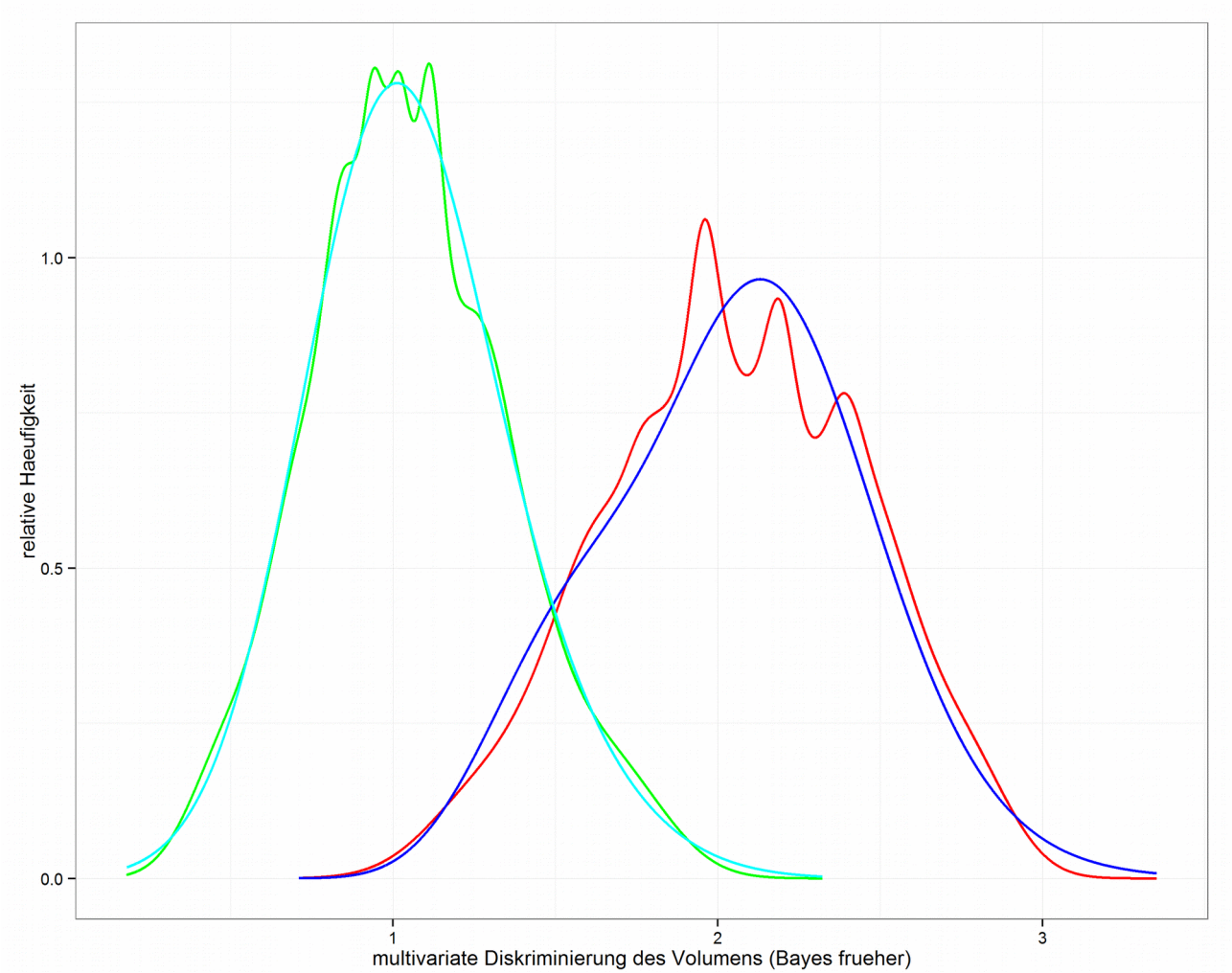

Abb. 46: Vergleich der ***theoretischen Verteilungen*** *mit den* ***geglätteten Verteilungen*** *der Eigenschaft der* ***multivariaten Diskriminierung der Volumen an drei verschiedenen Segmentierungstiefen*** *aus dem* ***ursprünglichen Bayes-Picking****. Es ist leicht zu erkennen, dass die* ***theoretischen Verteilungen*** *stark den geglätteten Verteilungen gleichen. In diesem Fall wurden die Daten logarithmisch skaliert und der adaptive Glättungsfaktor von -3 (also 3 Extrema) ergab im Falle der Signalverteilung (rot) den errechneten Glättungsfaktor von 63. Die geglättete Verteilung des Rauschens (grün) ergab den Glättungsfaktor 175. Für die Optimierung der theoretischen Signalverteilung wurde die Kombination LOGNLOGN (blau) gefunden. Für die Optimierung der Rauschverteilung wurde die Kombination NN (cyan) festgelegt.*

Abschließend war noch zu bewerten, ob und wie sich die Ergebnisse der Klassifizierung durch die Verwendung der *theoretischen Verteilungen* von den Ergebnissen der *geglätteten Verteilungen* unterscheiden. Dazu wurden alle Ergebnisse beider Verteilungstypen in Tabelle 29 gegenübergestellt. Man kann erkennen, dass die wiedererkannten Signale und Störsignale der *theoretischen Verteilungen* den gleichen

Trend aufweisen und kleine Abweichungen zu den *geglätteten Verteilungen* aufweisen und deren Ergebnisse relativ gut abbilden.

*Tabelle 29: Gegenüberstellung der Ergebnisse aus der Klassifizierung durch Verwendung der **geglätteten Verteilungen** und der Verwendung der **theoretischen Verteilungen** der Klassen Signal und Rauschen zur Wiedererkennung der 50436 NMR-Signale aus der Peakliste.*

Eigenschaft	Anzahl der korrekt w iedererkannten Signale und Störsignale bei Verw endung der **geglätteten** Verteilungen	Anzahl der korrekt w iedererkannten Signale und Störsignale bei Verw endung der **theoretischen** Verteilungen	Abw eichung der Anzahl der korrekt w iedererkannten Signale und Störsignale der **theoretischen** Verteilungen zu den **geglätteten** Verteilungen
Verhältnis Intensität zu lokalen Rauschen	48517	48477	40
Gaußsche Peakw ahrscheinlichkeit basierend auf den lokalen Rausch	48127	47939	188
Intensität	46271	46389	-118
Volumen	46069	46068	1
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w 1	44741	44578	163
Verhältnis der Peakintensität zur Linienbreite w 1	44639	44185	454
Verhältnis Intensität zur aufsummierten Linienbreite	44588	44401	187
Volumengrundfläche	44053	44053	0
Verhältnis der Peakintensität zur Linienbreite w 2	43917	43762	155
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w 2	43766	43505	261
Breite der Volumengrundfläche w 2	43344	43344	0
multivariates diskriminiertes Volumen (Bayes früher)	43285	43285	0
Volumenfehler	42963	42966	-3
Äußere Symmetrie	42304	42268	36
Pseudodistanz	42086	42239	-153
Linienbreite w 2	41616	42774	-1158
Breite der Volumengrundfläche w 1	41238	41228	10
Verhältnis der Peakintensität zur Volumengrundfläche	40695	40736	-41
Kreuz-Rauschen	39840	40385	-545
Linienbreite Summe (w 1+w 2)	39369	39369	0
Linienbreite w 1	36830	37507	-677

3.4.3 Die Erweiterung der Klassen *Signal* und *Rauschen* durch die Klasse *Wasser*

Zusätzlich wurde auch getestet, ob sich eine zusätzliche Verbesserung der Diskriminierung durch die Erweiterung der vormals verwendeten Klassen *Signal* und *Rauschen* durch eine weitere Klasse *Wasser* verbessern lässt. Dazu wurde der Bereich des Wasserstreifens aus dem Rauschbereich herausgelöst und in eine eigene Klasse *Wasser* verschoben (siehe Abschnitt 2.6.5.6). Die Gesamtanzahl der NMR-Signale in den Klassenbereichen bleibt somit erhalten. Das Vorgehen wurde analog zur Auswertung der Klassen *Signal* und *Rauschen* ausgeführt und wurde in den folgenden Ergebnissen stets mit einbezogen.

3.4.4 Die Varianten zur Bestimmung der Eigenschaftskombinationen zur Diskriminierung mittels des erweiterten Bayesschen Peak-Pickens bei Verwendung der geglätteten Verteilungen und ihrer optimalen Berechnungsparameter

Um ein optimales Ergebnis der Klassifizierung durch die Bayessche Wahrscheinlichkeit zu erreichen, wurde evaluiert, welche Kombinationsvariante der geglätteten Wahrscheinlichkeitsdichteverteilungen der Eigenschaften (also die Verteilung, welche mit den optimalen Parametern bestimmt wurde) das beste Resultat liefert. Diese Kombination soll später dann auch bei den Berechnungsmethoden der Bayesschen Wahrscheinlichkeit angewendet werden. Zudem soll gezeigt werden, dass die Erhöhung der Anzahl der Eigenschaften gegenüber dem ursprünglichen Bayesschen Peak-Picken mit lediglich drei Eigenschaften (Intensität, diskriminiertes Volumen und äußere Symmetrie) eine Verbesserung darstellt.

Da insgesamt N=23 Eigenschaften getestet werden mussten, ergab dies eine Gesamtanzahl von K=8.388.607 Kombinationsmöglichkeiten:

$$K = 2^N - 1 \tag{36}$$

Da ein manueller Test durch diese hohe Anzahl an Möglichkeiten zu aufwändig war, wurden folgende Varianten von Gruppierungen der Kombination der zu verwendenden Eigenschaften getestet:

Variante 1:

Alle 23 Eigenschaften aus Tabelle 21 wurden für die Klassifizierung verwendet.

Variante 2:

Bei dieser Variante wurden insgesamt 8.388.607 Diskriminierungen durchgeführt, um jede Kombination der Eigenschaften zu erreichen. Die Ergebnisse wurden wiederum in der Form von Hitlisten gespeichert und nach dem Mittelwert der Prozentualen Wiedererkennung von Signal und Störsignal absteigend sortiert. In dieser Liste wurde die Information der jeweils verwendeten Eigenschaften für die Klassifizierung durch die Bayessche Wahrscheinlichkeit mit abgespeichert. An erster Stelle der Liste war dann die Kombination zu finden, welche mit den verwendeten Eigenschaften die höchste Wiedererkennung für Signale und Störsignale erreichte. Für diese *Variante 2* wurden somit nur die Eigenschaften für die Klassifizierung verwendet, welche den ersten Platz der Hitliste einnahm.

Die Tabelle 30 zeigt eine Übersicht der verwendeten Eigenschaften zur Diskriminierung, welche den höchsten Mittelwert der prozentual wiedererkannten Signale und Störsignale erzielten. Die angekreuzten Eigenschaften wurden dann für die Variante 2 verwendet. Dabei wurde bei der Anwendung der beiden Klassen Signal und Rauschen eine mittlere prozentuale Wiedererkennung von 96,28 % erreicht. Im Falle der Verwendung der drei Klassen Signal, Rauschen und Wasser wurden 96,79 % erzielt (siehe Tabelle 33).

*Tabelle 30: Auflistung aller für **Variante 2** verwendeten Eigenschaften für die Klassifizierung durch die Bayessche Wahrscheinlichkeit, welche von allen 8.388.607 Diskriminierungen den höchsten*

Eigenschaft	**2 Klassen:** *Signal* und *Rauschen*	**3 Klassen:** *Signal*, *Rauschen* und *Wasser*
Verhältnis Intensität zu lokalen Rauschen	X	X
Intensität	X	X
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rauschen		
Volumen	X	
Verhältnis Intensität zur aufsummierten Linienbreite		
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w1	X	X
Verhältnis der Peakintensität zur Linienbreite w1	X	X
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w2		
Verhältnis der Peakintensität zur Linienbreite w2	X	X
Pseudodistanz		
Linienbreite w2		X
Volumengrundfläche		
Linienbreite Summe (w1+w2)	X	
Verhältnis der Peakintensität zur Volumengrundfläche		
multivariates diskriminiertes Volumen (Bayes früher)		
Breite der Volumengrundfläche w2		
Linienbreite w1		X
Äußere Symmetrie		
Kreuz-Rauschen		X
Breite der Volumengrundfläche w1		
Volumenfehler	X	X
Innere Symmetrie		X
Abstand Schwerpunkt zum Maximum		

Mittelwert der prozentual wiedererkannten Signale und Störsignale erreichte.

Variante 3:

Alle Eigenschaften, welche **alleine verwendet** je zu einem Mittelwert der prozentualen Wiedererkennung aus Signal und Störsignal von 85 % oder höher führten, wurden zur Berechnung der Diskriminierungen mit Variante 3 festgelegt. Diese 9 Eigenschaften wurden in Tabelle 31 grau hinterlegt.

Tabelle 31: Auflistung aller Eigenschaften, welche von allen Diskriminierungen bei Verwendung nur einer einzelnen Eigenschaft die Schwelle von 85 % der Mittelwerte aus richtig positiv und richtig negativ überschritten (grau hinterlegt).

Eigenschaften	Mittelwert Richtig positiv und Richtig negativ mit den Klassen Signal und Rauschen	Mittelwert Richtig positiv und Richtig negativ mit den Klassen Signal, Rauschen und Wasser
Verhältnis Intensität zu lokalen Rauschen	95,83	95,82
Intensität	94,82	95,33
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rauschen	94,56	94,67
Volumen	92,81	92,81
Verhältnis Intensität zur aufsummierten Linienbreite	91,44	91,51
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w1	91,08	91,12
Verhältnis der Peakintensität zur Linienbreite w1	91,07	91,15
Verhältnis der Peakintensität zur volumenbasierten Linienbreite w2	89,18	89,28
Verhältnis der Peakintensität zur Linienbreite w2	88,84	88,81
Pseudodistanz	84,79	84,82
Linienbreite w2	84,27	84,98
Volumengrundfläche	84,04	84,04
Linienbreite Summe (w1+w2)	83,89	83,89
Verhältnis der Peakintensität zur Volumengrundfläche	83,77	83,77
multivariates diskriminiertes Volumen (Bayes früher)	83,16	83,35
Breite der Volumengrundfläche w2	82,08	82,08
Linienbreite w1	81,72	82,13
Äußere Symmetrie	81,16	81,17
Kreuz-Rauschen	80,33	80,34
Breite der Volumengrundfläche w1	76,75	76,76
Volumenfehler	71,12	71,06
Innere Symmetrie	68,96	68,93
Abstand Schwerpunkt zum Maximum	55,02	55,05

Variante 4:

Nur die drei Eigenschaften, welche der ursprünglichen Berechnung der Bayesschen Wahrscheinlichkeit (Antz et al. 1995) zu Grunde lagen (also Intensität, multivariate diskriminierte Volumen und äußere Symmetrie) wurden verwendet.

Variante 5:

Hier wurden alle Eigenschaften ausgewählt (grau hinterlegt), welche öfter als 400.000 mal in den besten 800.000 Kombinationen (aus den 8.388.607 Möglichkeiten, welche bereits für Variante 2 errechnet wurden) beteiligt waren.

*Tabelle 32: Auflistung aller für **Variante 5** verwendeten Eigenschaften, welche von allen 8.388.607 Diskriminierungen innerhalb der Besten 800.000 Diskriminierungen mehr als 400.000 in den verwendeten Kombinationen auftraten.*

Eigenschaften	Häufigkeit (Klassen Signal und Rauschen)	Häufigkeit (Klassen Signal, Rauschen und Wasser)
Verhältnis Intensität zu lokalen Rauschen	800000	800000
Verhältnis der Peakintensität zur volumenbasierenden Linienbreite w1	562635	406140
Intensität	514803	800000
Volumen	497557	374901
Volumenfehler	494481	405728
Verhältnis der Peakintensität zur volumenbasierenden Linienbreite w2	489134	390033
Verhältnis Intensität zur aufsummierten Linienbreite	455483	397467
Verhältnis der Peakintensität zur Linienbreite w1	454414	419795
Verhältnis der Peakintensität zur Linienbreite w2	450722	418723
Äußere Symmetrie	432183	396235
Kreuz-Rauschen	420667	471017
Abstand Schwerpunkt zum Maximum	391741	398735
Linienbreite Summe (w1+w2)	352967	396767
Breite der Volumengrundfläche w2	346748	387800
Linienbreite w1	325923	401293
Verhältnis der Peakintensität zur Volumengrundfläche	310360	459900
Innere Symmetrie	299763	390303
Pseudodistanz	292670	311677
Breite der Volumengrundfläche w1	267850	400905
Volumengrundfläche	254044	387404
Gaußsche Peakwahrscheinlichkeit basierend auf den lokalen Rauschen	230204	291078
multivariate diskriminiertes Volumen (Bayes früher)	210810	383885
Linienbreite w2	182307	401212

Mit diesen *fünf* verschiedenen *Verwendungsvarianten* der Eigenschaftskombinationen wurden dann die Diskriminierungen durch *sechs Methoden* der Klassifizierung durch die Bayessche Wahrscheinlichkeit untersucht, um zu zeigen, ob die Erhöhung der Eigenschaftszahl zum einem und die Anwendung theoretischer Verteilungen zum Anderen einen Vorteil verschafft.

Dabei wurden die fünf Varianten mit folgenden Berechnungsmethoden getestet:

a) Berechnung ausschließlich über *geglättete Verteilungen* (**ohne Simulated Annealing**) mit den Klassen *Signal* und *Rauschen*.

b) Berechnung über Generierung von *theoretischen Verteilungen* mittels **Simulated Annealing** direkt aus dem zu bestimmenden Spektrum mit den Klassen *Signal* und *Rauschen*.

c) Berechnung der *theoretischen Signalverteilungen* aus einem simulierten Spektrum, bei der, die ebenfalls durch **Simulated Annealing** generierten Verteilungen aus b), nach-optimiert wurden mit den Klassen *Signal* und *Rauschen*.

d) Analog zu a), jedoch mit den Klassen *Signal*, *Rauschen* und *Wasser*

e) Analog zu b), jedoch mit den Klassen *Signal*, *Rauschen* und *Wasser*

f) Analog zu c), jedoch mit den Klassen *Signal*, *Rauschen* und *Wasser*.

*Tabelle 33: Übersicht der Ergebnisse aller **6 Methoden** kombiniert mit **Varianten** 1 bis 5.*

Methode	Variante	Mittelwert in Prozent aus Richtig positiv und Richtig negativ	Prozent der wieder-erkannten **Signale** Richtig positiv	Prozent der wieder-erkannten **Störsignale** Richtig negativ
a	1	95,56	94,75	96,37
	2	96,22	95,53	96,91
	3	95,89	95,15	96,64
	4	88,57	88,82	88,31
	5	96,1	95,53	96,66
b	1	93,72	92,83	94,61
	2	93,92	95,35	92,48
	3	94,72	97,42	92,03
	4	87,2	82,9	91,5
	5	94,4	98,29	90,51
c	1	94,33	95,65	93
	2	94,84	98,26	91,42
	3	94,55	97,88	91,21
	4	81,88	77,92	85,85
	5	94,75	97,92	91,58
d	1	96,35	96,05	96,66
	2	96,78	96,59	96,97
	3	96,52	96,38	96,67
	4	NA	NA	NA
	5	96,76	96,56	96,96
e	1	93,13	90,37	95,89
	2	93,84	94,85	92,84
	3	94,6	98,46	90,75
	4	NA	NA	NA
	5	94,38	97,39	91,37
f	1	92,52	90,28	94,76
	2	92,63	94,14	91,11
	3	94,14	99,07	89,21
	4	NA	NA	NA
	5	92,91	95,79	90,03

Tabelle 33 zeigt die Übersicht aller sechs verwendeten Berechnungsmethoden unter Einbeziehung einer jeden Variation 1 bis 5. Da die ursprüngliche Methode der Berechnung der Bayesschen Wahrscheinlichkeit nicht für mehr als zwei Klassen ausgelegt war, wurde für die Fälle d)-f) (also drei Klassen) die Variante 4 jeweils ausgelassen (NA).

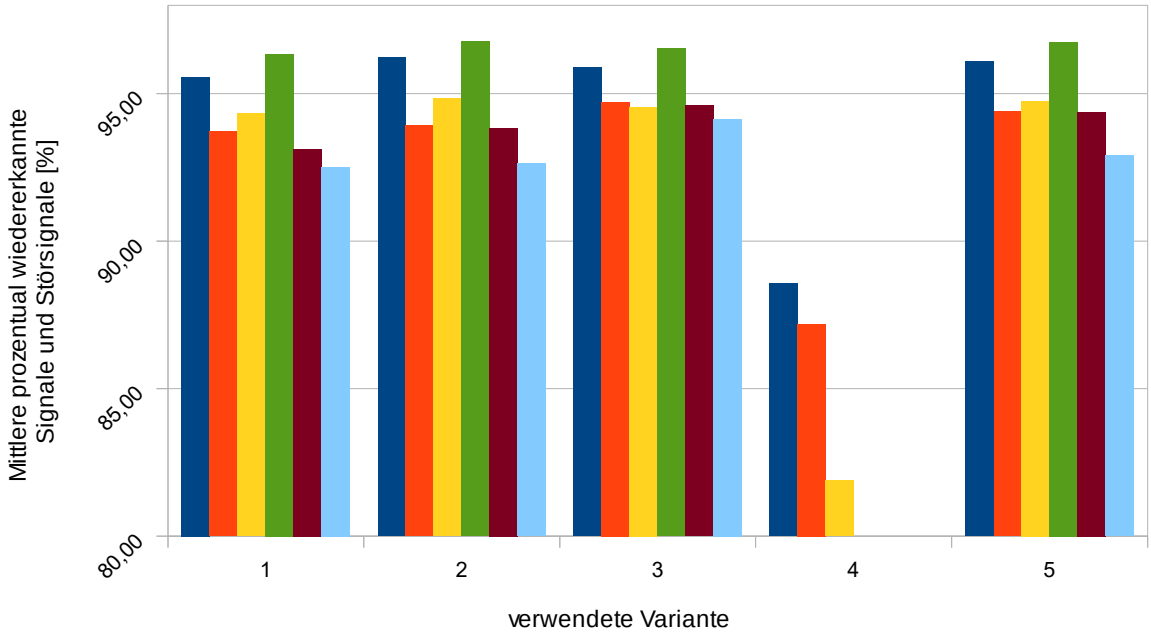

Abb. 47: Vergleich der Ergebnisse aller sechs verwendeten Methoden einer jeden Variante (1-5). ***Methode a)*** *(blau) verwendet keine theoretischen Verteilungen und nur die Klassen Signal und Rauschen.* ***Methode b)*** *(orange) verwendet nur theoretische Verteilungen des eigenen Spektrums und wieder nur die Klassen Signal und Rauschen.* ***Methode c)*** *(gelb) bezieht theoretische Signal-Verteilungen aus einem simulierten Spektrum mit ein und adaptiert diese Verteilungen auf die Verteilungen des experimentellen Spektrums der Klassen Signal und Rauschen. Die* ***Methode d)*** *(grün) ist analog zur Methode a), bezieht jedoch zusätzlich die Klasse Wasser mit ein. Auch die* ***Methode e)*** *(braun und analog zur Methode b)) und die* ***Methode f)*** *(cyan und analog zu c)) verwenden zusätzlich die Klasse Wasser.*

Betrachtet man die Ergebnisse aus Tabelle 33 bzw. aus Abb. 47, können folgende Feststellungen getroffen werden:

- Unabhängig von der Methode (Anzahl der Klassen, Verwendung simulierter Verteilungen usw.) konnte gezeigt werden, dass bereits die Anzahl der Eigenschaften einen Einfluss auf das Ergebnis hat. Denn alle Varianten (1, 2, 3 und 5) lieferten durch die höhere Anzahl an Eigenschaften bessere Resultate als Variante 4 bei drei Eigenschaften. Denn in Variante 4 konnte die Eigenschaft *Intensität* die anderen Eigenschaften mit schwachen Beitrag zur Klassifizierung

(*äußere Symmetrie* und *multivariate Diskriminierung der Volumen*) alleine nicht ausgleichen (siehe Tabelle 21). Das heißt, die Erhöhung der Anzahl der verwendeten Eigenschaften stabilisiert die Ergebnisse und stellt eine Verbesserung gegenüber der Variante 4 dar.

- Die Verwendung von Verteilungen, welche anhand desselben Spektrums generiert wurden (Methode b), an dem sie letztendlich auch getestet wurden, lieferten überraschenderweise ein leicht schlechteres Ergebnis bei den meisten Varianten, als Verteilungen, welche mittels eines simulierten Spektrums erstellt wurden (Methode c).
- Die Methoden a) und d), welche lediglich die *geglätteten Verteilungen* verwenden, lieferten bessere Wiedererkennungen als die Methoden mittels *theoretischen Verteilungen* (unabhängig von der Wahl der Variation 1,2 oder 3). Dies liegt daran, dass aus dem Vorwissen der Zuordnung die bereits bekannten Störsignale aus dem Signalbereich entfernt wurden und alle Signale im Bereich enthalten waren.
- Die Erweiterung der Klassenanzahl von zwei auf drei Klassen wirkte sich auf die Ergebnisse bei Anwendung der Methode d) gegenüber Methode a) positiv aus.

3.4.5 Analyse der Varianten und Berechnungsmethoden durch Reduktion der Signalklasse auf verschiedene Größen durch die gaußsche Peak-Wahrscheinlichkeit

Da in den vorausgegangenen Abschnitten aus der Klasse Signal die Störsignale entfernt wurden und die Zuordnung aller Signale des Proteins PfTrx bereits bekannt war, konnte die Qualität der Ergebnisse des Moduls getestet werden, ob die Erweiterungen einen Erfolg brachten.

Diese Situation zeigt sich jedoch nicht im realen Fall, da kein Wissen dafür existiert, welche Signale und Störsignale zu den jeweiligen Klassenbereichen gehören. Aus diesem Grund wurde der Realfall getestet, die Signalklasse vorab zu bereinigen, ohne Vorwissen wie in Abschnitt 3.4.4 beschrieben einzubringen. D.h. es galt so viele Störsignale wie möglich aus der Signalklasse zu entfernen ohne dabei wichtige Signale zu entfernen und dadurch die Generierung der Verteilungen zu verschlechtern. Dazu wurde der Ansatz der Reduktion mittels der gaußschen Wahrscheinlichkeit analog zu 2.7.2.9 angewendet und

der Datensatz einer jeden Eigenschaft bezüglich deren gaußschen Signalwahrscheinlichkeit prozentual um die Störsignale bereinigt.

Um eine sinnvolle Reduktion der Signalklassen (siehe Abb.9 und 10) einer jeden Eigenschaft zu erreichen, wurden mehrere Reduktionen für die jeweilige Kombination (also einer Variante) durchgeführt.

*Tabelle 34: Die Mittelwerte aus richtig positiv und richtig negativ des zweidimensionalen H-H-NOESY-Spektrums von PfTrx bei der Anwendung der sechs Methoden unter Verwendung der hier exemplarischen Kombinations-**Variante 5** (also alle Eigenschaften, welche in den 800.000 besten Signal-Wiedererkennungs-Wahrscheinlichkeiten über 400.000 mal in jeder Kombination auftritt). Die Reduktionen wurden von 5 % bis 85 % in 5 %-Schritten durchgeführt. Die höchsten Wiedererkennungen wurden grau hinterlegt und der gesamte Verlauf der Reduzierungen in Abb. 48 dargestellt.*

Größe des Signal-Datensatzes relativ zur Ausgangsgröße [%]	Methode a) bei Variante 5	Methode b) bei Variante 5	Methode c) bei Variante 5	Methode d) bei Variante 5	Methode e) bei Variante 5	Methode f) bei Variante 5
5	83,94	92,8	94,04	91,18	94,49	93,59
10	93,07	92,97	94,05	90,89	94,62	93,33
15	93,47	93,03	94,06	94,88	94,48	93,19
20	94,45	93,48	94,06	95,21	94,2	93,19
25	95,39	94,33	94,09	95,35	94,16	93,22
30	94,64	94,33	94,21	94,44	93,55	93,26
35	95,05	94,19	94,45	95,11	93,56	93,42
40	94,58	93,67	94,6	94,28	93,57	93,64
45	93,92	92,42	94,08	93,67	93,51	93,72
50	93,89	92,49	93,61	93,43	92,79	93,62
55	93,82	92,61	93,63	92,51	92,62	93,54
60	93,03	92,57	93,57	91,46	90,87	92,32
65	90,03	92,81	92,98	89,55	82,75	91,24
70	88,43	92,87	92,96	88,6	89,57	90,18
75	87,21	92,82	92,93	87,56	80,63	89,16
80	86,07	92,96	92,95	86,5	81,05	88,01
85	85,36	93,07	92,94	85,71	86,92	86,79

Tabelle 34 zeigt exemplarisch eine Übersicht der Reduktionen der Variante 5 für jede Berechnungsmethode a) bis f). Die Reduktionen wurden von 5 % bis 85 % in 5 %-Schritten für alle Methoden a) bis f) durchgeführt.

Die Reduktionsschritte geben Aufschluss darüber, innerhalb welcher Schritt-Bereiche die Ergebnisse durch Reduktion des Signal-Datensatzes am stabilsten bleiben. D.h. es

werden NMR-Signale sukzessive innerhalb der Signalbereiche entfernt, und danach geprüft, ob die Verteilungen noch gute Ergebnisse liefern.

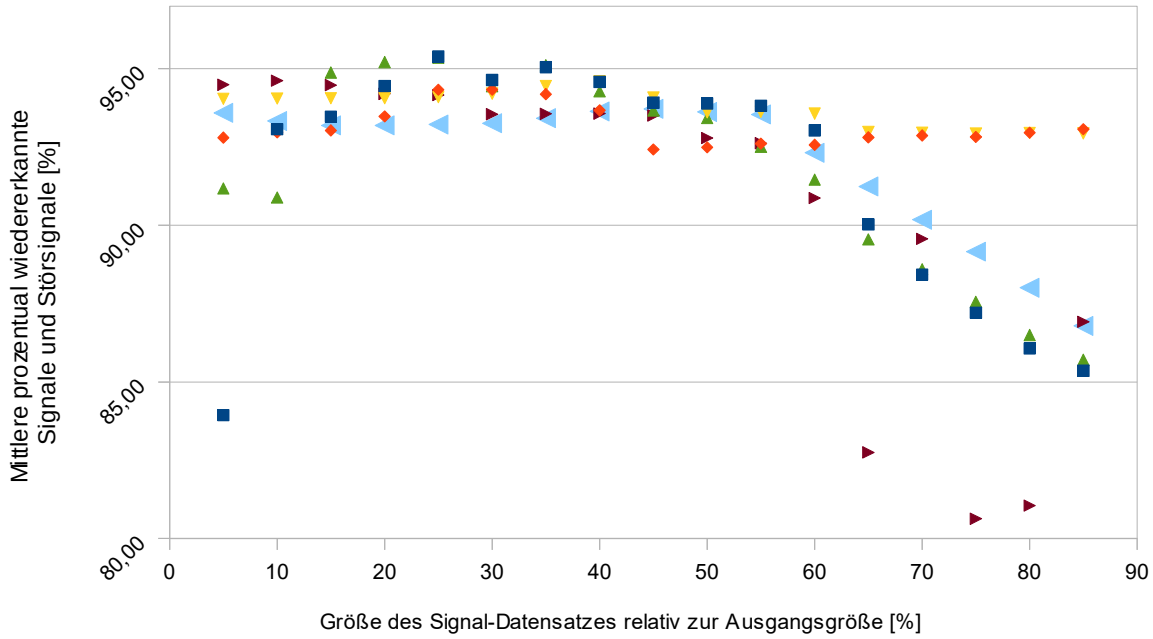

*Abb. 48: Mittelwerte aus richtig positiv und richtig negativ des zweidimensionalen ^{1}H-^{1}H-NOESY-Spektrums des Proteins PfTrx bei der Anwendung der sechs Methoden unter exemplarischer Verwendung der Kombinations-**Variante 5** (also alle Eigenschaften, welche in den 800.000 besten Signal-Wiedererkennungs-Wahrscheinlichkeiten über 400.000 mal in jeder Kombination auftreten). Die Datenpunkte geben die Methoden a(blau), b(orange), c(gelb), d(grün), e(braun) und f(hellblau) wieder. Die Reduktionen der Signalklasse um die Störsignale wurden von 5 % bis 85 % in jeweils 5 %-Schritten durchgeführt.*

Die Daten aus Abb. 48 geben den Verlauf der Mittelwerte der korrekt wiedererkannten Signale und Störsignale einer jeden Reduktion wieder. Dabei ist für die Variante 5 aller Methoden ein stabiler Bereich von 15 % bis 55 % zu erkennen. Die Methoden b) und c), welche jeweils die *theoretischen Verteilungen* für die Klassen *Signal* und *Rauschen* verwenden, weisen die geringste Empfindlichkeit gegenüber einer Reduktion der Signalklasse um Störsignale auf.

Dies wurde für jede Variante der Kombination der Eigenschaften 1 bis 5 durchgeführt und die optimale Reduktionsschwelle in der zusammenfassenden Tabelle 35 analog zur Tabelle 33 aufgeführt.

Tabelle 35: Übersicht der Ergebnisse aller sechs verwendeten Methoden einer jeden Variante 1 bis 5 unter Verwendung der optimalen Reduktion der Signal-Klassen.

Methode	Variante	Reduktion der Klasse Signal (auf %)	Mittelwert in Prozent aus Richtig positiv und Richtig negativ	Prozent der wieder-erkannten **Signale** Richtig positiv	Prozent der wieder-erkannten **Störsignale** Richtig negativ
a	1	35	94,81	94,85	94,77
	2	25	95,7	96,85	94,54
	3	25	95,35	96,08	94,62
	4	30	87,63	83,18	92,07
	5	25	95,39	97,09	93,69
b	1	20	93,97	93,16	94,79
	2	10	94,77	98,47	91,08
	3	5	95	95,35	94,64
	4	35	86,94	81,98	91,89
	5	30	94,33	97,92	90,73
c	1	55	94,61	95,79	93,43
	2	35	94,63	98,38	90,87
	3	35	94,18	97,57	90,8
	4	5	83,05	81,83	84,27
	5	40	94,6	97,45	91,75
d	1	35	95,12	94,39	95,84
	2	25	95,38	99,15	91,61
	3	35	95,41	94,4	96,42
	4	NA	NA	NA	NA
	5	25	95,35	99,21	91,48
e	1	25	93,28	90,65	95,9
	2	5	94,52	96,94	92,11
	3	35	94,43	98,25	90,62
	4	NA	NA	NA	NA
	5	10	94,62	98,12	91,13
f	1	60	93,29	91,23	95,35
	2	5	93,37	96,26	90,48
	3	65	94,03	98,58	89,48
	4	NA	NA	NA	NA
	5	5	93,59	97,64	89,54

Die Gegenüberstellung der Methoden und deren Varianten in Abb. 49 weist insgesamt ein schlechteres Ergebnis zu Abb. 47 auf. Die Methoden a) bis f) spiegeln den realen Fall wieder und bleiben daher natürlich hinter den Ergebnissen aus Kapitel 3.4.4 zurück, bei denen die korrekte Zuordnung der Signale bereits vorher bekannt war und die Klasse *Signal* von allen Störsignalen bereinigt wurde. Jedoch kann noch immer eine Verbesserung der bisherigen Methode der *geglätteten Verteilungen* gegenüber den *theoretischen Verteilungen* erkannt werden. Auffällig ist jedoch, dass auch in diesem Fall

durch die Reduzierung der Signalklasse die Verwendung der ursprünglichen drei Eigenschaften der Variante 4 (also Intensität, diskriminiertes Volumen und äußere Symmetrie) hinter den übrigen Methoden zurück liegt.

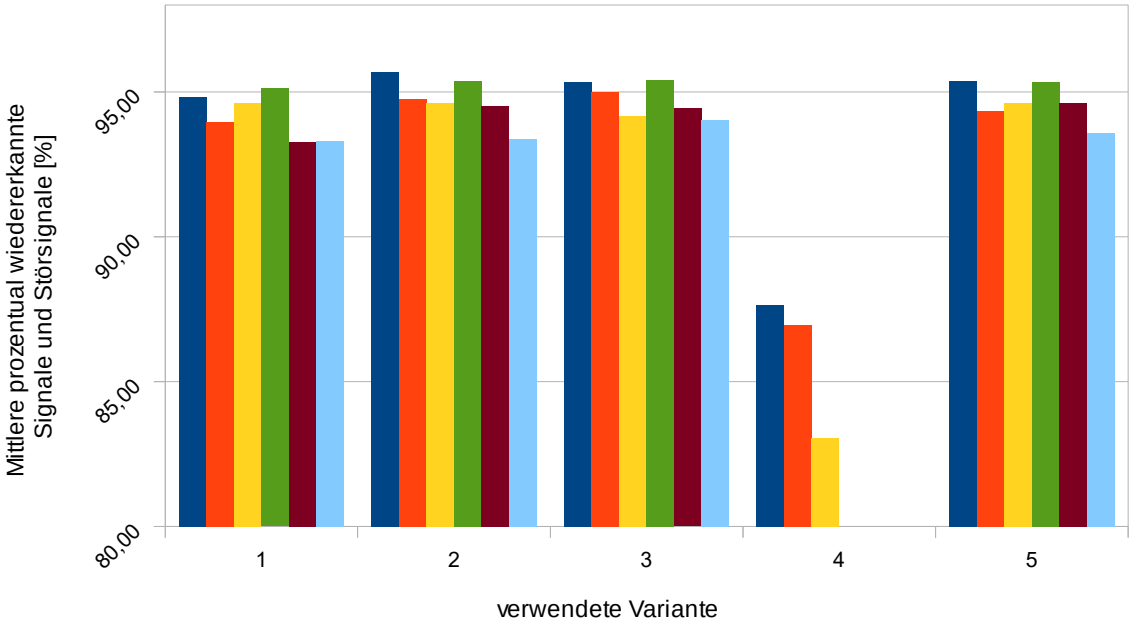

Abb. 49: Vergleich der Ergebnisse aller sechs verwendeten Methoden einer jeden Variante 1 bis 5 nach der Reduktion durch die optimale Reduktionsschwelle der Signalklasse. ***Methode a)*** *(blau) verwendet keine theoretischen Verteilungen und nur die Klassen Signal und Rauschen.* ***Methode b)*** *(orange) verwendet nur theoretische Verteilungen des eigenen Spektrums und wieder nur die Klassen Signal und Rauschen.* ***Methode c)*** *(gelb) bezieht theoretische Signal-Verteilungen aus einem simulierten Spektrum mit ein und adaptiert diese Verteilungen auf die Verteilungen des experimentellen Spektrums der Klassen Signal und Rauschen. Die* ***Methode d)*** *(grün) ist analog zur Methode a), bezieht jedoch zusätzlich die Klasse Wasser mit ein. Auch die* ***Methode e)*** *(braun und analog zur Methode b)) und die* ***Methode f)*** *(cyan und analog zu c)) beziehen zusätzlich die Klasse Wasser mit ein.*

Das beste Ergebnis wurde durch die Methode a) bei Verwendung der Variante 2 erreicht. Die hierzu verwendeten Eigenschaften finden sich in Tabelle 30 wieder. Es wurden dabei lediglich zwei Klassen (Signal und Rauschen) verwendet und der Datensatz der Klasse Signal durch die Reduktionsschwelle um 75 % auf die bereinigte Größe von 25 % reduziert.

4 Diskussion

Zu Beginn dieser Arbeit befand sich das Softwarepaket AUREMOL in einem Entwicklungsstadium, welches die Erweiterungen in dieser Arbeit nicht ohne weiteres erlaubte. So waren nahezu alle Algorithmen derart implementiert, dass eine Erweiterung der Module nicht „straight forward“ möglich war. Daher mussten einige Module neu implementiert werden, um die Weiterentwicklung der Algorithmen und die Performance der Module (wie z. B. das Multithreading oder die Optimierung der Speicherstrukturen) zu ermöglichen.

So wurde zuvor in nahezu allen Modulen der Speicher vom „Stack“ in fest codierter Größe angefordert. Da von diesem Speicherbereich des Rechners keine dynamische Anforderung möglich ist, musste im Redesign auf eine dynamische Anforderung des Speichers aus dem „Heap“ umgestellt werden. Zusätzlich galt es zu gewährleisten, dass nur soviel Speicher angefordert wurde, welcher auch zur Ausführung des Algorithmus erforderlich war. Zudem wurde darauf geachtet, dass so selten wie möglich neuer Speicher vom System angefordert wurde, da sich jede Anforderung eines neuen Speicherbereiches negativ auf die Performance auswirkte. Dazu wurden bereits angeforderte Speicherbereiche wieder verwendet.

Die Anforderung der neuen Module, mittels zentraler Algorithmen Spektren verschiedener Dimensionen verwalten zu können, machte es notwendig, alle Lese- und Schreibzugriffe auf die Rohdaten (Spektren und deren Peaklisten) neu zu entwickeln, bei denen nur eine Schnittstelle für alle Dimensionen nötig war. Dies gelang mit dem rekursiven Ansatz aus 3.1.1 mit entsprechenden Speicheroptimierungen, welche in Module dieser Arbeit sowie in weitere nicht hier aufgeführte Module in AUREMOL implementiert wurden. Zudem ist dieser Ansatz vollständig unabhängig von der AUREMOL-Basis AURELIA. Dies ermöglicht allen künftigen Funktionalitäten, entweder eigenständig in einem Modul zu arbeiten oder in ein anderes Softwarepaket eingebunden zu werden.

4.1 Verbesserte Integration der Signalvolumen

Hinsichtlich der Performance konnte das Modul „Integration“ am meisten durch die Neuimplementierung gewinnen. Der hohe Entwicklungsaufwand reduzierte die Integrationszeit von hochaufgelösten Spektren von mehreren Stunden auf wenige Minuten. Der Grund der zu langen Rechenzeiten der ursprünglichen Version war, dass die Hauptvorteile der Sprache C/C++ nicht genutzt wurden. Dazu zählten vor allem die Verwendung der Zeigerarithmetik und die dynamische Speicherverwaltung. Zudem ist in dieser Arbeit ein spezieller Vektor-Container entwickelt worden, der freien Speicher möglichst selten neu anfordert und damit zur Entlastung des Prozessors führte. Auch der Speicherzugriff wurde ungünstig realisiert. Anstatt der schnelleren und resourcenschonenden Zeigerarithmetik war in AUREMOL oft der langsamere Indexzugriff auf die Inhalte der Vektoren gewählt worden. Auch hatte die Integration in der früheren Version eine unangemessen hohe Prozessorauslastung, um Inhalte von Arrays auf den Wert 0 zu überprüfen. Dies wurde durch mehrere Zeigerarrays optimiert, da diese nur auf die relevanten Werte (ungleich 0) referenzieren. Des Weiteren kann die Integration durch die rekursiven Rohdatenzugriffe nun auf beliebig hohe Dimensionen der Spektren angewendet werden. Dies ist aber nur möglich, wenn das Modul entkoppelt von AUREMOL ausgeführt wird, denn AUREMOL konnte bislang nur maximal vierdimensionale Spektren bearbeiten. Für eine jede Dimension existierte dabei eine eigene Implementierung.

Eine weitere für diese Arbeit essentielle Erweiterung brachte die Einführung der Struktur *Integrations-Hash* im Rahmen der Weiterentwicklung der Integration. Diese zusätzliche Struktur war in der Lage, alle Pixel, die einem Signalvolumen zugehörig waren, separat zu speichern. Dies hatte den Vorteil, dass nach der Integration ein schneller Zugriff auf Volumen verschiedener Segmentierungstiefen möglich war. Dazu musste nicht erneut bei jeder Variation der Segmentierungstiefe das Volumen neu berechnet werden. So war es nun möglich, jedes Signal-Volumen mit einer gewünschten Segmentierungstiefe abzurufen. Diese muss jedoch höher sein, als die Segmentierungstiefe mit der die Struktur *Integrations-Hash* durch die Integration erstellt wurde.

Diese Struktur ermöglichte die Erweiterung der Eigenschaften des Bayesschen Peak-Pickings durch die Erstellung der volumenbasierten Eigenschaften. So konnten bei der

Berechnung der volumenbasierten Eigenschaften verschiedene Segmentierungstiefen zeitsparend variiert werden. Auch die Berechnung des Massenschwerpunkts basiert auf dieser Struktur.

Die Visualisierung der Volumen von ein- und zweidimensionalen Volumen erlaubt es, die Volumen der einzelnen Signale in getrennten Ansichten oder in einer gewünschten zusammengefassten Übersicht qualitativ zu bewerten.

Die Ermittlung des optimalen Integrationsbereiches, in dem der Algorithmus das Volumen jedes NMR-Signals berechnet, stellt gegenüber der früheren Methode einen Vorteil dar, da dieser keinerlei Vorwissen durch manuelle Bereichsdefinition, Linienbreite oder Nachbarschaftsverhältnisse benötigt. Die Schwäche in der ursprünglichen Bestimmung des Integrationsbereiches war, dass sich durch das manuelle Hinzufügen (bzw. Entfernen) von NMR-Signalen in die (bzw. aus der) Peakliste auch der Integrationsbereich ändert und daher der Algorithmus ein anderes Volumen ermittelt als vor dem Hinzufügen (bzw. dem Entfernen). Diese Inkonsistenz wurde durch den neuen Algorithmus beseitigt.

Die Ergebnisse der Integration bei problematischen Spektren mit einer hohen Zerklüftung zeigen, dass die Methode der Glättung des Spektrums während der Integration eine Verbesserung darstellt, da der Wachstumsalgorithmus bei jedem Extremum (verursacht durch das Rauschen oder eine schlechte Aufnahme) stoppen würde. Eine Glättung nimmt zwar eine Veränderung der Rohdaten in Kauf, ist aber in der Lage, das Ergebnis am korrekten Volumen anzunähern (siehe Abb. 15b), während der ursprüngliche Algorithmus an dieser Stelle fehlschlagen würde (siehe Abb. 15a).

Der Glättungsansatz könnte aber noch dahin verbessert werden, dass die Glättung nicht auf das gesamte Spektrum angewendet wird, sondern auf die Integrationsboxen eines jeden zu integrierenden Signals. Dies hätte den Effekt, dass weniger zerklüftete Signalformen auch einen schwächeren Glättungsfilter erhalten würden und dadurch die Integrationsbereiche schärfer abgegrenzt werden.

Wurde ein synthetisches Spektrum mit dem AUREMOL-Modul RELAX generiert, waren alle wichtigen Informationen in der erstellten Peakliste vorhanden. Würde jedoch dieses synthetische Spektrum erneut integriert werden, würden sich die integrierten Volumen von den synthetischen Volumen unterscheiden, da diese analytisch bestimmt wurden. Die Gründe für die Abweichungen sind, dass sich die Positionen der Signale überwiegend

abseits eines Extremum befunden haben und mehrere Signale an derselben Positionen anzutreffen waren. Das Volumen dieser Signale war mit der ursprünglichen Methode nicht integrierbar, da diese nur Signale akzeptierte, deren Position sich an einem Extremum befanden. Zudem wurden Mehrfachzuordnungen von NMR-Signalen derselben Position so integriert, als würde diese Mehrfachbelegung nicht existieren, was zu einem überhöhten Volumen führte. Der Benutzer musste in diesem Fall manuell die Zuordnungen zu einem einzigen Signal zusammenfassen und dies in der Peakliste vermerken. Es wurde gezeigt, dass die neue Methode diese Nachteile nicht mehr aufweist und sowohl Volumen von Mehrfachzuordnungen als auch Signale, deren Positionen nicht an einem Extremum liegen, berechnen kann. Die Positionierungen der Signale „Nicht-Extremum„ und „mehrere Signale an derselben Position“ treten nach einem automatischen Peak-Peaking nicht auf. Jedoch ist es üblich, manuell NMR-Signale nachträglich an eine andere Position zu verschieben, oder eine Signal manuell an ein „Nicht-Extremum“ zu setzen. Egal welche Methode zur Bestimmung der Position eines Signals verwendet wird, erhält man mit der erweiterten Methode der Integration stets einen Volumenwert. Dieser ist, wie gezeigt wurde, je nach Überlagerung, in beinahe allen Fällen näher am korrekten Wert als der der ursprünglichen Integration (Geyer et al. 1995). Zumal diese sogar bei Signalen, deren Position nicht an einem Extremum liegt, die Integration verweigerte (siehe Abb. 20 und 21). Diese Verbesserungen der Volumenberechnung von NMR-Signalen und die Einführung der Struktur *Integrations-Hash* erlaubte weitere Ansätze in dieser Arbeit bezüglich der Bestimmung der Position durch den Schwerpunkt von Signalen und der Berechnung der Bayesschen Wahrscheinlichkeit für ein Signal bzw. Störsignal.

Da durch den *Integrations-Hash* die Signalform am digitalen Raster nach der Integration gespeichert wird, kann zu jeder Segmentierungstiefe die Volumenform abgefragt und zusätzlich visualisiert werden. Dies ermöglicht mit sehr niedrigem Rechenaufwand die Berechnung der Positionen der Schwerpunkte aller NMR-Signale aus der Peakliste durch der im Volumen enthaltenen Voxel (Abb. 23 und 24) (Formel 6 oder 7).

Zur Bestimmung der Volumen gibt es verschiedene Methoden. Die wichtigsten sollen daher kurz diskutiert werden. Eine manuelle Festlegung der Integrations-Bereiche wäre enorm aufwändig, da jedes Signal interaktiv markiert werden müsste. Diese Variante kann daher nicht empfohlen werden und verhindert den Automatisierungsansatz.

Die Multiplikation der Linienbreite auf halber Höhe eines Signals jeder Dimension mit der Intensität des Signals (Fejzo et al. 1990) repräsentiert die einfachste Fit-Methode zur Ermittlung der Volumen von Resonanzsignalen. Da bei dieser Methode die Signalintensität in der Regel die Signalposition darstellt, wird bei einem Signal, dessen Position sich abseits eines Extremums des digitalen Rasters befindet, nicht die Projektion der Intensität am Extremum verwendet, sondern die der Signal-Position. Diese ist in diesem Fall niedriger als die Intensität des Extremums. Ermittelt man dazu nun die Halbwertsbreite des Signals der jeweiligen Dimensionen, werden die Signalformen des ganzen Signals verwendet. Da die Intensität abseits des Extremums zu niedrig wäre, führt dies zu einem zu geringen Volumen. Einen ähnlichen Effekt induzieren überlappende Signale. Je stärker die Anhebung der Intensität an der Signalposition durch eine oder mehrere Überlappungen ist, desto mehr weicht der Wert der Intensität von der korrekten Intensität des Signals ab. Auch wirkt sich die Überlappung negativ auf die Berechnung der Halbwertsbreiten aus. Eine niedrige digitale Auflösung verstärkt diese Effekte zudem. Aufwändigere Fit-Methoden, wie die Anpassung von theoretischen Linienformen (Gauss oder Lorenz) oder benutzerdefinierten Linienformen (Brown und Huestis 1994; Sze et al. 1995) an die durch das Peak-Picking erhaltenen Signale stellen eine bessere Wahl dar. Jedoch unterliegen diese Methoden ebenfalls den bereits im vorhergehenden Punkt aufgeführten Nachteile. Zudem ist die Methode der Angleichung von benutzerdefinierten Signalen als Referenz auf das Spektrum stark davon abhängig, die Referenzsignale korrekt auszuwählen und benötigt daher einen interaktiven Eingriff des Benutzers (Eccles et al. 1991; Denk et al. 1986).

Einen komplexeren Ansatz zum automatischen Peak-Picking und zur Integration von NMR-Signalen bietet die Software AUTOPSY (Koradi et al. 1998). Dabei wird die Bestimmung der Signalvolumen durch Fits von Linienformen und Amplituden umgesetzt, welche durch den Pick-Algorithmus erhalten worden sind. Der Pick-Algorithmus berechnet dabei das Rauschniveau (Summe aus dem globalen Rauschen und des lokalen Rauschens) und wendet an diesem Niveau einen „flood fill“-Algorithmus an. Der Effekt dabei ist, dass NMR-Signale mit Intensitäten oberhalb des „Wassers“ zu isolierten Signalen und zu Gruppen von überlagerten Signalen führen. Signale unterhalb der Wasseroberfläche werden als Nutzsignale für die Auswertung ausgeschlossen, da deren Signalform vollständig durch die Zusammenführung zu verbundenen Regionen aus dem

Spektrum gelöscht werden. Prinzipiell entspricht dies der Segmentierung dieser Arbeit, bei der jedoch die Segmentierung relativ zu jedem Extremum der Signalform angewandt wird. Die isolierten Signale liefern nach Anwendung des „flood fill“-Algorithmus die potentiellen Signale und werden durch Prüfung von Symmetrie- und Linienform-Eigenschaften als Nutzsignal identifiziert und dann die Linienformen für die spätere Berechnung des Volumens durch Überlagerung gespeichert. Bei einer stark verzerrten Baseline, bei hohen thermischen Rauschen, bei niedrigen digitalen Auflösungen oder bei stark ausgeprägten Artefakten kann dies zu einer Verschlechterung der Linienformen und Symmetrieeigenschaften führen. Da bei dreidimensionalen Spektren zwar insgesamt mehr Pixel zu Verfügung stehen, jedoch die digitale Auflösung in den jeweiligen Frequenzdomänen niedriger ist, als bei zweidimensionalen Spektren, wirkt sich dies negativ auf die Bildung der benötigten Linienformen aus, falls ggfs. zu wenig Pixel zur Verfügung stehen, um die Linienform adäquat abzubilden.

Ein großer Unterschied des AUTOPSY-Algorithmus zur Methode in dieser Arbeit liegt darin, dass nicht die Information aus dem „Picken“ für die Integration verwendet wird, sondern umgekehrt die Ergebnisse der Integration einen Teil der Pick-Routine darstellen und die Basis zur Verbesserung der Signalpositionen bilden. Im Gegensatz zu AUTOPSY können für die meisten Spektren die Standardparameter bei der erweiterten Integration dieser Arbeit verwendet werden. Alleine durch die wiederholte Integration nach Veränderung der Startparameter (z. B. automatische Variation der Grösse der Integrationsbereiche oder die automatische Erhöhung der Anzahl der Integrationsschritte für den Wachstumsalgorithmus) konnten die Ergebnisse verbessert werden.

Der ursprüngliche Region-Wachstums-Algorithmus (Geyer et al. 1995) bei Verwendung der Segmentierung (Neidig und Kalbitzer 1990) wurde durch folgende Erweiterungen der Integration realisiert:

- Möglichkeit der Integration von NMR-Signalen, deren Position sich am digitalen Raster nicht an einem Extremum befinden. Diese Signale wurden in der ursprünglichen Integration verworfen.
- Abfrage verschiedener Segmentierungstiefen ohne Neuberechnung des Volumens durch Speicherung der anteiligen Intensitäten am Volumen in der Struktur *Integrations-Hash*.

- Ausnutzung aller verfügbaren logischen oder physikalischen Prozessoren bei optimaler Speichernutzung durch problemspezifisch optimierte dynamische Speichercontainer (Zeigerarithmetik und Wiederverwendung bereits angeforderter Speicherbereiche).
- Visualisierung der Volumen von Signalen in ein- und zweidimensionalen Spektren.
- Verbesserung der Volumen durch die automatische Größenermittlung des Integrationsbereiches für Signale der Peakliste und variable Erhöhung der Integrationsschritte im Vergleich zur ursprünglichen Integration.
- Verbesserung der Integration stark verrauschter Signale durch Glättung des Rohspektrums.
- Einsatzmöglichkeit für n-dimensionale Spektren durch einen bereits in das Softwarepaket AUREMOL eingebundenen rekursiven Ansatz.

4.2 Das Modul Schwerpunktbestimmung zur Verbesserung der Positionsbestimmung von NMR-Signalen

Wie bereits in Abschnitt 3.3.1 anhand des Beispiels aus Abb. 24 dargestellt wurde, lag die Position des Signals 166 durch die Berechnung dessen Schwerpunkts näher an der Position der Simulation als die Position des Signals, welche durch die Lage des Pixels am Extremum (*Maximum-Methode*) bestimmt wurde. Die Methoden zur Berechnung des Schwerpunkts sind an verschiedenen Spektren verschiedener digitalen Auflösungen, jeweils mit und ohne Rauschen und mit unterschiedlichen und gleichen Volumen der beiden Signale getestet worden. Im Falle zweidimensionaler Spektren wurde die Annäherung eines Signals auf ein anderes sowohl horizontal als auch diagonal untersucht.

Um einen aussagekräftigen Verlauf darstellen zu können, sollten 160 Schritte für einen Durchlauf vollzogen werden. Bei jedem Schritt wurde der Schwerpunkt und das zugehörige Maximum bestimmt (siehe Bewegungsgraphen in Abb. 27). Durch die vielen Variationen und der zusätzlichen Untersuchung der Methode zur Berechnung des Volumens im Falle der Überlappung der beiden Signale bis nur noch ein Extremum vorhanden war, wurde der Umfang der Daten so hoch, dass eine sinnvolle Zusammenfassung nötig war. Durch die Trennung und Untersuchung all dieser Variationen gestaltete sich die Auswertung als sehr aufwändig. Es konnte gezeigt werden, dass die Methode der Bestimmung der Position eines Signals durch den Schwerpunkt der Volumenanteile durchwegs bessere Ergebnisse erreichte, als die Festlegung der Position des Signals am Extremum der Signalform bezüglich des digitalen Rasters.

So schneidet die Methode „Bestimmung der Position des Signals durch die Schwerpunkt-Methode“ gegenüber der ursprünglichen Methode (siehe Abschnitt 3.3.7) bei eindimensionalen Spektren etwa um den Faktor 3 und im Falle von zweidimensionalen Spektren um etwa den Faktor 20 besser ab (siehe Tabelle 16). Es konnte gezeigt werden, dass bessere Ergebnisse erzielt werden, falls die anteiligen Intensitäten am Volumen durch die Segmentierungstiefe reduziert wurden. Dabei lagen die optimalen Segmentierungstiefen zwischen 0,0001 und 0,2. Hier kann empfohlen werden, eine möglichst tiefe Segmentierung zu verwenden.

In Abschnitt 3.3.8 verdeutlichte sich, dass nahezu alle Methoden (siehe Tabelle 19) der Bestimmung der Signal-Position durch den Schwerpunkt bessere Ergebnisse lieferten als die Methode durch die Positionsbestimmung durch das Extremum am digitalen Raster, falls die Überlappung der Signale so groß ist, dass ein beeinflussendes anderes Signal kein Extremum mehr aufweist. Dabei war das Modul „getrennte Volumen“ die Methode, welche die besten Ergebnisse erbracht hatte. In diesem Fall waren die Verbesserungen nicht so signifikant wie in Abschnitt 3.3.7. Dabei lagen die optimalen Segmentierungstiefen zwischen 0,2 und 0,5. Daher wird empfohlen, eine möglichst hohe Segmentierung in etwa der halben Linienbreite (also 0,5) zu verwenden.

Da jedoch in einem Spektrum beide Varianten aus 3.3.7 und 3.3.8 vorliegen können und die Verbesserung aus 3.3.8 weniger ins Gewicht fallen, wird empfohlen, stets eine Segmentierung von 0,2 zu verwenden. Dies stellt eine gute Balance zwischen Laufzeitverhalten und Resultat dar.

Da die Ergebnisse den Mittelwert der Abweichungen der Positionen in ppm zur Referenzposition darstellen, fallen die Werte stets sehr klein aus. So wäre die Wahl der Abweichung in Prozent womöglich besser gewesen, da die Werte damit nicht so niedrig skaliert hätten.

Zusammenfassend konnte gezeigt werden, dass die Bestimmung der Position des Signals durch die *Schwerpunktbestimmung mit Abschneidung* an der Segmentierungstiefe der *Maximum-Methode* vorzuziehen ist.

Einen ähnlichen Ansatz verfolgt CAPP (Garrett et al. 2011). Dieser umgeht die aufwändige interaktive Festlegung vieler Nutzsignale und Störsignale als Referenz für einen Vergleich mit den gepickten NMR-Signalen (wie z. B. in STELLA Kleywegt et al. 1990 durch einen Vergleich von der gelernten Datenbank mittels des Cosinuskriteriums) zu umgehen, indem er lediglich eine sogenannte interessante Signalform definiert.

Diese benötigte Signalform bildet die Grundlage für die Erstellung von Höhenlinien in der Form von Ellipsen, welche bei einem optimalen isolierten Signal konzentrisch aufeinander liegen und das Signal optimal abbilden. Um eine NMR-Signal mit den Höhenlinien zu rekonstruieren, durchläuft der Algorithmus vier Schritte:

1. Generierung der realen Höhenlinien (bzw. in allen Schnitten bei höheren Dimensionen) basierend auf allen Intensitäten des digitalen Rasters des Spektrums. Zu jeder Höhenlinie wird dann später eine an diese Kontur passende Ellipse berechnet. Hierzu verlangt diese Methode die Angabe des „level multipliers“ und einer absoluten Schwelle der Intensität für die Berechnung der Konturen durch den Benutzer. Dieser Multiplikator legt in diesem Schritt die Unterteilung der Konturen innerhalb des Konturdiagramms für ein NMR-Signal.
2. Festlegung der Parameter aller benötigen Ellipsen, welche die Höhenlinien am Besten wiedergeben. Dazu werden Ellipsen durch annähernd korrekte Startparameter festgelegt und anschließend durch das Simplex-Verfahren (also die Abweichung der realen Höhenlinie zur Ellipse) optimiert.
3. Suche der Ellipsen, welche die Wellenkämme aufgrund des Rauschens entlang einer jeden Frequenzdomäne aufweist beschreiben.
4. Bestimmung der Nutzsignale als solche und deren Position durch die Anwendung der Ellipsen aus Schritt 2 und Ausschluss durch Vergleich mit den Kämmen durch Rauschen aus Schritt 3. Jedoch dürfen lediglich Ellipsen verwendet werden, welche zuvor manuell festgelegte Grenzwerte nicht über oder unterschreiten dürfen.

Im Falle eines zweidimensionalen Spektrums kann die Rekonstruktion der Signalformen direkt erfolgen und bei höheren Dimensionen (lediglich bis zur vierten Dimension) werden zweidimensionale Schnitte aus den Spektren iterativ abgearbeitet. Das heißt, es werden in jedem zweidimensionalen Schnitt alle Signalformen rekonstruiert und es muss dabei jedes mal auch das Optimierungsverfahren der Ellipsen durchlaufen werden. Dies führt bei Spektren mit einer hohen digitalen Auflösung zu einem sehr hohen Rechenaufwand, welcher bei der Schwerpunktmethode nahezu wegfällt, da diese Methode den *Integrations-Hash* aus der bereits geschwindigkeitsoptimierten Integration direkt verwenden kann. Die Berechnung aller Signalpositionen an deren Schwerpunkten in erfolgt dabei in wenigen Sekunden.

Da CAPP viele vom Benutzer teils spektrumspezifische Parameter für Schritt 1 und 4 benötigt, erschwert dies den Ansatz der Automatisierung. Außerdem kommt es bei starken Überlappungen von NMR-Signalen dazu, dass der Algorithmus in Schritt 1 bei den unteren Konturen die überlagerten NMR-Signale umschließt und dadurch mehrere Signale zu

einem Signal durch gemeinsame Ellipsen zusammenfasst. Der Ansatz in CAPP ist durch die Festlegung der Position des Signals durch das Mittel der Zentren der verwendeten Ellipsen der in dieser Arbeit entwickelten Schwerpunktmethode ähnlich, da beide auf der Analyse der Signalform basieren. Aufgrund der angeführten Nachteile von CAPP wird die Verwendung der Schwerpunktmethode empfohlen.

4.3 Signalidentifizierung durch die Bestimmung der Bayesschen Wahrscheinlichkeit

Um den Zeitaufwand für eine Strukturbestimmung zu optimieren, musste ein automatisierter Ansatz gefunden werden, welcher die Arbeit des Experimentators erleichtert bzw. beschleunigt.

Viele alternative Methoden zur Identifizierung von Nutzsignalen aus NMR-Spektren sind in den letzten 30 Jahren entwickelt worden. Die einfachste Methode stellt die manuelle Bestimmung der Nutzsignale durch den Benutzer dar. Da diese Methode sehr zeitaufwändig ist, sollte eine automatische Bestimmung diese Aufgabe schneller bewältigen.

Eine Methode war, alle Signale über einem durch den Benutzer bestimmten Schwellwert der Intensität (Neidig et al. 1984; Cieslar et al. 1988) als Nutzsignal festzulegen. Diese Methode tendiert aber dazu, Nutzsignale mit einer zu niedrigen Intensität zu verwerfen oder stark angehobene Störsignale durch Wasserstreifen oder eine gestörte Baseline als Nutzsignal zu bewerten.

Aufwändigere Methoden, welche die Trennung der Nutzsignale von den Störsignalen verbessern, sind z. B. Ansätze, welche auf neuronale Netze basieren (Carrara et al. 1993; Corne et al. 1992; Pons und Delsuc 1999). Diese haben jedoch den Nachteil, dass sich die Netzwerkstruktur im Falle eines Backpropagation-Netzes (Anzahl der Eingabeneuronen, Neuronen der verborgenen Schicht oder der Ausgabeneuronen) zwischen den verschiedenen Typen und Dimensionen von Spektren erheblich unterscheiden kann, so dass für jeden Typ eines Spektrums ein Netz bestimmt werden muss. Auch haben Untersuchungen gezeigt (Carrara et al. 1993), dass diese Methoden sehr langsam sind.

Die Erstellung der Verteilungen durch Simulated Annealing oder durch die Generierung der geglätteten Verteilungen in dieser Arbeit entsprechen dem Lernvorhang der neuronalen Netze und liegen innerhalb weniger Minuten vor. Als ein weiterer Nachteil neuronaler Netze erweist sich, dass bei abgeschlossenem Lernvorgang nicht bekannt ist, welche mathematische Funktion die Diskriminierung leistet. Es liegt also lediglich eine

„black box“ vor, welche keine weitere analytische Betrachtung und Bewertung der gefundenen Funktion zur Diskriminierung erlaubt.

Ein anderes Verfahren, welches die Bestimmung von Nutzsignale durch die Überlagerung von Ellipsen erlaubt, wurde im vorangegangenen Abschnitt 4.2 bereits diskutiert (CAPP Garrett et al. 2011). Das am meisten zitierte AUTOPSY (Koradi et al. 1998) wurde ebenfalls in 4.1 betrachtet und wendet nach der Erstellung der Peakliste weitere Filtermethoden durch Signaleigenschaften wie Symmetrie der Signale bezüglich der Spektrumdiagonalen an, damit sich die Qualität der Peakliste verbessert. Sehr ähnlich arbeitet die Methode aus PICKY (Alipanahi et al. 2009). Diese wendet ebenfalls am Anfang einen „flood fill“-Algorithmus an (basierend auf dem lokalen und globalen thermischen Rauschen), um vorab schon Signale zu verwerfen. Leider werden hier auch alle Intensitäten jedes Pixels an dieses Wasserlevel angepasst, um zusammenhängende Bereiche (aus Intensitäten) bilden zu können. Die Rohdaten des Spektrums werden in der Methode dieser Arbeit nicht verändert, sondern nur bei speziellen Berechnungen von Signaleigenschaften variiert. So existiert bei allen volumenbasierten Eigenschaften ein Segmentierungslevel. Dieses Level gilt nicht global im gesamten Spektrums, sondern es gilt nur relativ innerhalb des Integrationsbereiches des zu integrierenden Signals. Damit wird gewährleistet, dass die Volumeninformation weniger durch eine fehlerhafte Baseline verzerrt wird. Um die Klasse *Rauschen* festlegen zu können, sind auch beliebig viele Extrema (entstanden durch thermisches Rauschen) innerhalb des Klassenbereiches erlaubt, da diese ohnehin zur Klassifizierung benötigt werden. Die Angabe eines Thresholds, welcher nur NMR-Signale mit einer Intensität über diesen in der Peakliste erlaubt, ist daher für die Funktionalität der Methode in dieser Arbeit irrelevant und dient nur zur Verkürzung der Laufzeit.

Das Filterkriterium der Symmetrie eines Signals kann in AUTOPSY nur bei symmetrischen Spektren angewendet werden. Die Ergebnisse dieser Arbeit haben aber gezeigt, dass diese Eigenschaft eine eher schwache Diskriminierung erreichte und ist daher für die Methode dieser Arbeit nur optional. Nach der Anwendung dieses „flood fill“-Algorithmus werden in AUTOPSY die erhaltenen Bereiche über dem Wasserlevel in eindeutige und überlappende Signale durch Symmetriekriterien eingeteilt. Alle Bereiche mit eindeutigen Signalen dienen zur Festlegung von Linienformen. Diese werden gruppiert und dann zur Darstellung überlappter Bereiche verwendet.

Der Ansatz in PICKY beruht nicht auf Symmetriekriterien, sondern es wird eine SVD (Singular Value Decomposition) auf die erhaltenen Bereiche angewandt. Da beide Algorithmen die erwähnten schwachen Signale als Störsignale identifizieren und verwerfen, birgt dies die Gefahr eines möglichen Informationsverlustes. Daher war es das Ziel dieser Arbeit, eine Methode anzubieten, welche Störsignale zwar als „schlecht“ bewertet, aber das potentielle Störsignal nicht als Kandidaten für die endgültige Peakliste entfernt. Damit galt es, die Berechnung der Wahrscheinlichkeit, dass es sich bei einem NMR-Signal tatsächlich um ein Nutzsignal handelt, zu verbessern.

Die alleinige Bewertung von NMR-Signalen durch die Gaußsche Wahrscheinlichkeit ist zumeist nicht ausreichend, da viele Rauschpeaks zu hohe Wahrscheinlichkeiten erhalten. Daher wird in dieser Arbeit der Bayes-Ansatz zur Signalidentifizierung (Antz et al. 1995) erweitert, um eine verbesserte Bewertung der NMR-Signale zu erreichen.

Leider bot das ursprüngliche Modul zu Beginn dieser Arbeit keine akzeptable Funktionalität. Der Hauptgrund dafür war, dass die NMR-Signale vor der Klassifizierung durch die Peak-Picking-Routine „adaptives Peak-Picking“ (Trenner 2006) bereits massiv reduziert wurden. Dabei wurden Signale entfernt, welche keine Störsignale darstellten und dies führte dazu, dass die Klasse *Rauschen* kaum mehr signifikante Störsignale enthielt. Das Fehlen von Störsignalen in der Klasse *Rauschen* gestaltete daher eine Klassifizierung als problematisch. Um diesen Fehler zu beheben, wurde die einfache Threshold-basierende Pick-Methode erweitert, so dass mehr NMR-Signale in die Peakliste aufgenommen werden konnten. Dabei wurden diese NMR-Signale zwar mit einer Wahrscheinlichkeit basierend auf dem lokalen Rauschen versehen, jedoch nicht aus der Peakliste entfernt und nur im Viewer ausgeblendet. Die ausgeblendeten NMR-Signale standen aber weiterhin für die Signalidentifizierung zur Verfügung. Ein weiterer Grund für die fehlerhafte Funktionalität des ursprünglichen Moduls war, dass bei der Eigenschaft „dekorreliertes Volumen an drei Segmentierungstiefen“ bei der Bildung der Diagonalisierungsmatrix zur Bestimmung der Eigenwerte ungültige Einträge durch numerische Überläufe entstanden sind und nicht im Vorfeld verhindert wurden. Nach weiteren Korrekturen von Logikfehlern im Modul waren die Ergebnisse trotzdem nicht befriedigend, so dass zusätzlich ältere Versionen zurück bis zum Jahr 2005 getestet wurden. Die Ergebnisse waren mit all diesen Versionen nicht befriedigend und es wurde daher entschieden, das Modul neu zu entwickeln.

4.3.1 Die Erweiterung des Moduls durch Erhöhung der Anzahl und Optimierung der Berechnungsmethoden der Eigenschaften von NMR-Signalen sowie die Variation der Klassenanzahl und des Glättungsfaktors der geglätteten Wahrscheinlichkeitsdichteverteilungen

Da die ursprüngliche Routine des Bayesschen Peak-Pickens (Antz et al. 1995) lediglich drei Eigenschaften (Intensität an der Signalposition, dekorrellierte Signalvolumen und Signalsymmetrie bezüglich der Diagonalen eines homonuklearen Spektrums) zur Erstellung der Verteilungen unterstützte, war die Annahme, dass sich eine Eigenschaft, welche einen schwachen Beitrag zur Diskriminierung beiträgt (durch eine schlechte Beschaffenheit der Spektrum-Rohdaten), negativ auf die Bewertung eines Signals auswirkt. Daher sollte die Erhöhung der Anzahl der Eigenschaften die Qualität der Ergebnisse verbessern, indem mehrere verwendete Eigenschaften einen eventuellen negativen Beitrag auf die Güte der Bewertung der Signale durch eine oder mehrere schlechte Eigenschaften ausgleichen.

Um auch die Qualität der drei ursprünglich verwendeten Eigenschaften zu optimieren, wurde die Verbesserung der Volumen aus Kapitel 3.2 genutzt und ein hoher Aufwand in die Verbesserung der Eigenschaft äußere Symmetrie (Schulte et al. 1997) eines Signals bezüglich der Symmetriediagonalen im Falle von homonuklearen Spektren investiert (siehe äussere Symmetrie unter 2.6.5.1 und Tabelle 24).

Nach Festlegung der zu verwendenden Eigenschaften wurden die Methoden zur Berechnung der zugrundeliegenden Datenbasis für die einzelnen Eigenschaften variiert. Im Anschluss daran erfolgte nach Möglichkeit eine Skalierung der errechneten Datensätze (siehe Kapitel 3.4.1). Aus den jeweiligen Klassen konnte dann zu jeder Eigenschaft eine Wahrscheinlichkeitsdichteverteilung generiert werden. Diese Verteilungen wurden zusätzlich mit verschiedenen Glättungsfiltern erstellt und auf die optimale Größe des Glättungsfilters hin untersucht (siehe Tabelle 23 und 24).

Um den Beitrag der Eigenschaften für eine möglichst gute Diskriminierung bewerten zu können, wurde mit jeder Variation der Verteilungen eine Diskriminierung durchgeführt. Aus dem Ergebnis aus Tabelle 21 bei Verwendung der Klassen *Signal* und *Rauschen* kann man feststellen, dass die in der ursprünglichen Methode des Bayesschen Peak-Pickings die verwendeten Eigenschaften „äusseren Symmetrie“ (bei einem Mittelwert der richtig

positiv und richtig negativ erkannten Signalen von 81,16 %) und „multivariates diskriminiertes Volumen“ (bei einem Mittelwert der richtig positiv und richtig negativ erkannten Signalen von 83,16 %) eine schlechte Diskriminierung aufwies.

Bei der Verwendung einer einzigen Eigenschaft zur Diskriminierung konnten folgende Eigenschaften eine gute Bewertung der Signale und Störsignale erreichen (siehe vollständige Tabellen 21 und 22):

- **Verhältnis Signal-Intensität zu lokalen Rauschen**:

  Der Mittelwert der richtig erkannten Signale war 95,83 % (Klassen *Signal* und *Rauschen*) und 95,82 % (Klassen *Signal*, *Rauschen* und *Wasser*)

- **Signal-Intensität**

  Der Mittelwert der richtig erkannten Signale war 95,82 % (Klassen *Signal* und *Rauschen*) und 95,33 % (Klassen *Signal*, *Rauschen* und *Wasser*)

- **Gaußsche Signalwahrscheinlichkeit basierend auf dem lokalen Rauschen**

  Der Mittelwert der richtig erkannten Signale war 94,56 % (Klassen *Signal* und *Rauschen*) und 94,67 % (Klassen *Signal*, *Rauschen* und *Wasser*)

- **Signal-Volumen**

  Der Mittelwert der richtig erkannten Signale war 92,81 % (Klassen *Signal* und *Rauschen*) und 92,81 % (Klassen *Signal*, *Rauschen* und *Wasser*)

Bereits hier zeigt sich, dass die Verwendung von drei Klassen anstatt von zwei Klassen keine signifikanten Veränderungen bewirkt.

In Tabelle 23 ist ersichtlich, dass die Verwendung der adaptiven Glättung der Wahrscheinlichkeitsdichtefunktionen am besten diskriminieren (1 Maximum bis 4 Maxima). Im Gegensatz zur ursprünglichen Version des Bayesschen Peak-Pickens, bei der alle Werte als Absolutwerte genommen wurden (Antz et al. 1995; Schulte et al. 1997), stellte sich heraus, dass für die optimale Datenbasis keine Absolutwerte genommen werden werden dürfen, um eine bessere Diskriminierung zu erreichen. Die logarithmische Reskalierung der Daten zur Erstellung der Verteilungen brachte lt. Tabelle 23 ebenfalls eine Verbesserung der Klassifizierung.

4.3.2 Die Erzeugung der theoretischen Verteilungen basierend auf dem optimalen Parametersatz der geglätteten Verteilungen

Ein weiterer Ansatz dieser Arbeit war, zu testen, ob die Verteilungen der Eigenschaften pro Klasse durch eine theoretische Verteilung wiedergegeben werden kann. Es galt, je Eigenschaft und Klasse eine Funktion zu finden, welche die Verteilungen der Wahrscheinlichkeiten einer jeden Eigenschaft aus den verschiedenen Klassen optimal wiedergibt. Zudem sollten diese Verteilungen auch auf sehr wenige NMR-Signale in einem Spektrum anwendbar sein (beschränkte Statistik), welche keine Generierung von Verteilungen durch ihren geringen Umfang zulassen. Dazu wurden Wahrscheinlichkeitsverteilungen aus Abschnitt 2.7.1 definiert und sowohl einzeln als auch miteinander kombiniert getestet, um die beste Diskriminierung zu erhalten.

Damit die freien Parameter der Verteilungsfunktionen (24 bis 29) für jede Eigenschaft bezüglich jeder Klasse bestimmt werden konnten, wurde die Maximum-Likelihood-Schätzung (R.A. Fisher and the making of maximum likelihood 1912-1922 1997) für die Parameter μ1, μ2, σ1, σ2 und λ verwendet, indem die Maximum-Likelihood-Funktion (30) durch Simulated Annealing (Kirkpatrick et al. 1983) maximiert wurde. Aufgrund des hohen Stichprobenumfangs wurde dieses Verfahren der Momentenmethode (Hazewinkel 2002) vorgezogen, da die Berechnung durch die numerische Iteration einen sehr hohen Resourcenbedarf hat.

Während der Optimierungsläufe mussten für die Eingrenzung der Nachbarschaftssuche erlaubter Zustände durch Parameter aus den geglätteten Verteilungen extrahiert werden (siehe 2.7.2.2). Die Evaluation des Abbruchkriteriums (siehe 2.7.2.6) und die Adaption des Metropoliskriteriums (siehe 2.7.2.3) stellte dabei sicher, dass der Optimierungslauf sowohl zeitnah endet als auch konsistente Ergebnisse liegert.

Danach wurde für die beiden Fälle der Verwendung von zwei Klassen bzw. drei Klassen für jede einzelne Eigenschaft (siehe Tabelle 25 als Beispiel für die Eigenschaft Intensität bei der Verwendung von zwei Klassen) eines Signals die Kombination aus den Verteilungstypen (24 bis 29) ermittelt, welche für jede der 23 Eigenschaften am besten diskriminiert und in eine Konfigurationsdatei gespeichert. Diese Datei kann anschließend für eine Diskriminierung von NMR-Signalen eines anderes Spektrums des gleichen Typs angewendet werden (siehe 3.4.2). Die Energielandschaft der Eigenschaften zu jeder

Klasse wurde durch die Rohdaten aufgebaut, welche mittels der optimalen Parameter aus Kapitel 3.4.1 aus dem Spektrum und der Peakliste bestimmt wurden. Die theoretischen Verteilungsfunktionen trafen dabei optimal die Verteilungen aus den geglätteten Verteilungen (siehe z. B. Abb. 46). Der Mittelwert der richtig positiv und richtig negativ wiedererkannten Signale und Störsignale bewegte sich dabei im Falle der Diskriminierung mittels zwei Klassen von der besten Diskriminierung von 95,77 % der Eigenschaft „Verhältnis Intensität zu lokalen Rauschen“ durch die Verteilungen LOGN (Klasse *Signal*) und N (Klasse *Rauschen*) bis hin zur schlechtesten Diskriminierung von 55 % (NLOGN/LOGNN) der Eigenschaft „Abstand Schwerpunkt zum Maximum“. Die Ergebnisse der Diskriminierung aller Eigenschaften durch Verwendung der theoretischen Verteilungen finden sich für den Fall der Verwendung der Klassen *Signal* und *Rauschen* in Tabelle 28 wieder. An dieser Stelle sei erwähnt, dass die Verwendung der Verteilungen NLOGN und LOGNN nahezu identisch war. Dies lag daran, dass die Optimierung lediglich das Gewicht der beiden Funktionen (also λ_1 und λ_2) vertauschte. Damit wies in beiden Fällen dieselbe Funktion annähernd gleiches Gewicht auf.

Im Falle der Verwendung von den drei Klassen *Signal*, *Rauschen* und *Wasser* (siehe Abschnitt 3.4.3) war der Ablauf analog und der Trend nahezu identisch und wurde daher lediglich in die Endauswertung mit einbezogen.

Zu Beginn der Erstellung der Optimierung konnte oftmals die Zielfunktion nicht gefunden werden, da durch die Exponentialfunktionen der Funktionen ein numerischer Überlauf entstand. Dieser Überlauf wurde durch Absicherungen vermieden, was jedoch eine negative Auswirkung aus das Laufzeitverhalten der Optimierung hatte, obwohl die Berechnungen optimal auf die CPU-Kerne verteilt wurden. Daher wurde das Metropoliskriterium dahingehen adaptiert, dass der Optimierungslauf für alle 23 Eigenschaften innerhalb weniger Minuten (je nach der Verfügbarkeit von Prozessorkernen) die Parameter der Funktionen bestimmen konnte.

Die generierte Datenmenge war so enorm, dass eine manuelle Auswertung dieser Daten nahezu unmöglich war. Daher musste die Auswertung in AUREMOL integriert werden. Der Aufbau der nötigen Zwischenergebnisse gestaltete sich als schwierig, da bei der Implementierung stets darauf geachtet werden musste, die Übersicht über die temporären

Dateien der Zwischenergebnisse und das nachfolgende Einlesen in den Arbeitsspeicher nicht zu verlieren.

4.3.3 Die beste Kombination der Signal-Eigenschaften um die optimale Diskriminierung zu erreichen

In Kapitel 3.4.4 wurde abschließend ermittelt, welchen Einfluss die Kombination bestimmter Eigenschaftsgruppen (Varianten 1 bis 5) auf die Diskriminierung hatte.

Varianten:

1. Alle 23 Eigenschaften wurden verwendet.
2. Aus den 8.388.607 Kombinationsmöglichkeiten aller Eigenschaften wurde die höchste Wiedererkennung ermittelt und die dabei verwendeten Eigenschaften stellten somit diese Variante dar (siehe Tabelle 30).
3. Alle Eigenschaften, welche bei der Verwendung nur dieser Eigenschaft alleine eine Wiedererkennung von über 85 % erzielten, wurden für diese Variante herangezogen. Diese neun Eigenschaften wurden in Tabelle 31 (grau markiert) aufgeführt und zeigen, dass die Verwendung einer zusätzlichen Klasse (Wasser) keine signifikante Änderung der Wiedererkennung aufweist (Abb. 47 dritter Säulenblock).
4. Nur die drei Eigenschaften, welche der ursprünglichen Berechnung der Bayesschen Wahrscheinlichkeit (Antz et al. 1995) zu Grunde lagen (also Intensität, multivariate diskriminierte Volumen und äußere Symmetrie), wurden für diese Variante verwendet.
5. Bei dieser Variante wurden all die Eigenschaften ausgewählt, welche öfter als 400.000 mal in den besten 800.000 Kombinationen der beteiligten Eigenschaften (aus den 8.388.607 Möglichkeiten aus Variante 2) beteiligt waren (siehe Tabelle 32).

Der Grund für die Festlegung der Varianten 2, 3 und 5 war festzustellen, ob eine Filterung der Eigenschaften nach deren Qualität das Ergebnis zur Verwendung aller 23 Eigenschaften einen Vorteil bringt.

Da durch die Einführung der theoretischen Verteilungen die Möglichkeit bestand, die Anwendung der Verteilungstypen (*theoretisch* und *geglättet*) zu variieren, wurde auch dies untersucht.

Daraus ergab sich in der weiterentwickelten Routine ein Modul des Bayesschen Peak-Pickens mit fünf verschiedenen Varianten der Kombination der Eigenschaften. Folgende sechs Methoden dienen zur Bestimmung der Bayesschen Wahrscheinlichkeiten:

a) Erstellung von **geglätteten Verteilungen** aus den Rohdaten der Eigenschaften aller zugehörigen Klassen mit verschiedenen Glättungsfiltern **aus dem experimentellen Spektrum**. Diese geglätteten Verteilungen stellen die Basis für die Diskriminierung dar.

b) Erstellung von **theoretischen Verteilungen** durch Simulated Annealing aus den Rohdaten der Eigenschaften aller zugehörigen Klassen **aus dem experimentellen Spektrum**. Ausschließlich diese theoretischen Verteilungen wurden dann zur Diskriminierung verwendet. Dabei dienten die geglätteten Verteilungen lediglich dazu, die erlaubten Nachbarschaften der Zustände in der Energielandschaft während des Optimierungslaufs zu bestimmen.

c) Erstellung von **theoretischen Verteilungen** durch Simulated Annealing aus den Rohdaten der Eigenschaften. Hier wurden die Verteilungen der Klasse *Signal* aus den Rohdaten des synthetischen Spektrums und die Verteilungen der übrigen Klassen aus den Rohdaten des experimentellen Spektrums gewonnen.

d) Methode a), jedoch mit der zusätzlichen Klasse *Wasser.*

e) Methode b), jedoch mit der zusätzlichen Klasse *Wasser.*

f) Methode c), jedoch mit der zusätzlichen Klasse *Wasser.*

Die Ergebnisse sind in Tabelle 33 und in Abb. 47 aufgeführt und lassen folgende Schlussfolgerungen zu.

- Es bildete Variante 4 (ursprüngliche Version des Moduls) bezüglich der übrigen Varianten unabhängig von der verwendeten Methode das Schlusslicht. Dies zeigt, dass der schwache Beitrag die Eigenschaften „äußere Symmetrie“ und „multivariate

Diskriminierung der Volumen“ zur Diskriminierung nicht alleine durch die gut diskriminierende Eigenschaft Intensität ausgeglichen werden konnte.

- Methode b) lieferte ein leicht schlechteres Ergebnis bei den meisten Varianten, als Verteilungen, welche mittels eines simulierten Spektrums durch Methode c) erstellt wurden. Dies bekräftigt die Vermutung, dass aus dem simulierten Spektrum mehr Informationen gewonnen werden konnte, als aus dem experimentellen Spektrum.
- Die Methoden a) und d), welche lediglich die geglätteten Verteilungen verwenden, lieferten bessere Wiedererkennungen als die Methoden mittels theoretischer Verteilungen (unabhängig von der Wahl der Variation 1,2 oder 3). Dies liegt daran, dass die bereits bekannten Störsignale aus dem Signalbereich komplett entnommen wurden. Da die geglätteten Verteilungen teilweise viele Extrema zugelassen haben, tritt ein gewisser „auswendig lernen“-Effekt ein (siehe Spalte Glättungsfaktor aus Tabelle 23). Da die Maximalanzahl der Extrema bei den theoretischen Verteilungen maximal 2 beträgt und dadurch einen glatten Verlauf aufweisen, wird dieser Effekt verhindert und sollte einen Vorteil bei der Verwendung anderer Spektren bringen (vor allem bei Spektren mit sehr wenig Signalen).
- Die Erweiterung der Klassenanzahl von zwei auf drei Klassen verbesserte lediglich Methode a) und d) in allen Varianten. Methode b) und c) blieben durch die zusätzliche Klasse nahezu unverändert. Im Falle der Methoden b) und c) verschlechterte die Klasse Wasser die Wiedererkennung sogar.

Da bisher die Klasse *Signal* durch das Vorwissen einer abgeschlossenen Zuordnung von den Störsignalen befreit wurde, stellt dies nicht den realen Fall dar. Daher wurde weiter untersucht, wie die Klasse *Signal,* ohne Vorwissen über die Zugehörigkeit der NMR-Signale zu einer Klasse zu haben, reduziert werden kann.

4.3.4 Reduzierung der Signalklasse durch die Entfernung der Störsignale aus der Signalklasse vor der Klassifizierung durch die Information des lokalen Rauschens

Im ursprünglichen Modul wurde die Klasse der Signale durch eine Korrektur der Verteilung (bei niedrigen Intensitäten) durch die Peakdichte innerhalb der Signalklasse festgelegt und bei hohen Intensitäten nur die ausreichend hohen Intensitäten in der Klasse belassen. Da

das neue Modul auch auf andere Spektrumtypen anwendbar ist, kam diese Variante der Korrektur der Signalklasse nicht mehr in Frage. Daher wurde der Ansatz einer Bereinigung der Signalklasse durch die Information des lokalen Rauschens getestet. Somit wurden die Verteilungen nicht durch die Dichte der NMR-Signale korrigiert (siehe in 2.7.2.9 Abb. 9 vor der Reduzierung und Abb. 10 nach der Reduzierung), sondern es wurde stets die Anzahl der Peaks in der Signalklasse reduziert.

Dazu wurden die Wahrscheinlichkeitsdichteverteilung der Eigenschaft „gaußsche Signalwahrscheinlichkeit basierend auf dem lokalen Rausch“ für die Klasse *Signal* erstellt und dann wurde schrittweise (5 %-Schritte) das Konfidenzlevel verändert und Peaks über diesem Level aus der Signalklasse entfernt. In Abb. 48 (bei Verwendung der Variante 5) zeigt sich, dass die Wiedererkennung bei der Anwendung die beiden Klassifizierungs-Methoden b) und c) (theoretische Verteilungen mit zwei Klassen) relativ stabil bleibt.

In Tabelle 35 sind die Ergebnisse der Klassifizierung mit der optimalen relativen Reduzierungsgröße analog zum vorigen Kapitel aufgelistet. Vergleicht man Abb. 49 (realer Fall; kein Vorwissen der Signale bezüglich der Klassenzugehörigkeit) und Abb. 47 (Wissen über Klassenzugehörigkeit vorhanden), erreichen nahezu alle Methoden in allen Varianten, außer der Variante 4, ähnliche Ergebnisse und stellen daher eine sehr gute Methode zur Klassifizierung von Signalen und Störsignalen dar.

Das beste Ergebnis wurde durch die Methode a) (nur geglättete Verteilungen) bei Verwendung der Variante 2 erreicht. Das heißt, es wurden zwei Klassen (Signal und Rauschen) verwendet und die Verwendung der folgenden Eigenschaften sollen als Empfehlung dienen:

- *Das Verhältnis der Signalintensität zum lokalen Rauschen*
- *Die Intensität des Signals*
- *Das Volumen eines Signals*
- *Das Verhältnis der Signalintensität zur volumenbasierten Linienbreite*
- *Das Verhältnis der Signalintensität zur Linienbreite*
- *Die Summe der Linienbreiten*
- *Volumenfehler*

Für den Parameter zur Reduktion der Klasse Signal wird die Reduktionsschwelle um 75 % auf die bereinigte Größe von 25 % empfohlen. Die Anwendung von theoretischen Verteilungen auf die Signalklasse direkt auf das zu bewertende Spektrum lieferte stets schlechtere Ergebnisse als diese empfohlene Methode. Dabei war es irrelevant, welche Eigenschaften zur Diskriminierung verwendet worden waren.

Liegt jedoch ein Spektrum vor, welches nur wenige Nutzsignale enthält, stehen zu wenig Signale aus dem Signalbereich zur Generierung der geglätteten Verteilungen zu Verfügung. Methode a) schließt sich für diesen Fall aus. Daher wird empfohlen, für solche Spektren Methode c) zu wählen, welche die theoretischen Signal-Verteilungen aus einem synthetischen Spektrum bestimmt.

4.4 Bewertung

Die entwickelten Module lassen viele verschiedene Parameter zu, die es dem Experten ermöglicht, eine Feinabstimmung der Algorithmen auf das Spektrum zu erlauben. Jedoch ist es für die Automatisierung der Strukturbestimmung nötig, möglichst viele Standard-Parameter festzulegen. Die empfohlenen Einstellungen der Module in dieser Diskussion sollten beachtet werden, um ein interaktives Eingreifen durch den Benutzer zu minimieren und gleichzeitig optimale Ergebnisse zu erreichen. Der durchgängig rekursive Ansatz macht die Module darüber hinaus robust und zukunftssicher.

Da die Module *Integration* und *Schwerpunktbestimmung* alle Spektren mit beliebiger Dimension und ohne Vorwissen über das Spektrum verarbeiten können, ermöglicht dies die einfache Einbindung in ein automatisches Verfahren. Die automatische Anpassung der Startparameter der Integration zur Umsetzung besserer Ergebnisse hat sich als sehr effektiv gezeigt. Trotz der Erhöhung des Rechenaufwandes, während der Integration der NMR-Signale, konnte durch die Optimierung des Algorithmus eine hohe Performance des Moduls gewährleistet werden.

Nach Analyse der Ergebnisse durch die Schwerpunktbestimmung zur Verbesserung der Positionsbestimmung von NMR-Signalen wird empfohlen, diese der ursprünglichen Methode zur Festlegung der Positionen der Signale (meist am Extremum der Signalform) vorzuziehen. Die Positionen am Schwerpunkt lagen nach Variation des Grades der Überlagerung, der Signalform und der Bewegungsrichtung bei verschiedenen digitalen

Auflösungen der Spektren überwiegend näher an der theoretischen Position als die der ursprünglichen Positionen.

Durch die Möglichkeit des schnellen und einfachen Auslesens von Volumeninformationen aus dem *Integrations-Hash*, konnte die Anzahl der volumenbasierten Eigenschaften von NMR-Signalen leicht erhöht werden. Die Annahme, dass die Diskriminierung durch die Erhöhung der Eigenschaften verbessert werden kann, konnte in dieser Arbeit gezeigt werden. Obwohl es sich bei dem getesteten Spektrum um ein ^{1}H-^{1}H-NOESY-Spektrum des Proteins *PfTrx* mit vielen Nutzsignalen (6738) und Überlagerungen handelte, konnte dennoch eine hohe Wiedererkennung von Nutzsignalen und Störsignalen erreicht werden. Besonders gute Resultate erzielte dabei die Methode, welche die verbesserten geglätteten Verteilungen verwendete. Diese Methode erreichte bei Anwendung der vorher genannten Eigenschaften eine Wiedererkennung von 95,7 % (Mittelwert in Prozent aus richtig positiv und Richtig negativ wiedererkannten Signalen) mit der Reduktion der Klasse *Signal* auf 25 %. Die Erweiterung der Anzahl der Klassen und die Klasse *Wasser* blieb jedoch hinter der Annahme einer Verbesserung der Diskriminierung zurück. Einzig die Wahl einer Schwelle für die Wahrscheinlichkeit, welche festlegt, ab wann ein NMR-Signal ein Nutzsignal darstellt, muss noch durch den Benutzer definiert werden. Diese Schwelle könnte z. B. durch eine am Anfang festgelegte Zahl der zu erwarteten Nutzsignale angenähert werden, wenn von der Annahme ausgegangen wird, dass ca. 95 % der Signale wiedererkannt werden.

Es konnte in dieser Arbeit gezeigt werden, dass AUREMOL durch die Neuerungen und Erweiterungen in dieser Arbeit die Signalidentifizierung zur Strukturbestimmung von Proteinen verbessern konnte.

5 Literaturverzeichnis

Alipanahi, Babak; Gao, Xin; Karakoc, Emre; Donaldson, Logan; Li, Ming (2009): PICKY: a novel SVD-based NMR spectra peak picking method. In: *Bioinformatics (Oxford, England)* 25 (12), S. 75. DOI: 10.1093/bioinformatics/btp225.

Antz, C.; Neidig, K. P.; Kalbitzer, H. R. (1995): A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. In: *Journal of Biomolecular NMR* 5 (3), S. 287–296. DOI: 10.1007/BF00211755.

Bartels, Christian; Güntert, Peter; Billeter, Martin; Wüthrich, Kurt (1997): GARANT-a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. In: *J. Comput. Chem.* 18 (1), S. 139–149. DOI: 10.1002/(SICI)1096-987X(19970115)18:1<139::AID-JCC13>3.0.CO;2-H.

Brown, James W.; Huestis, Wray H. (1994): Quantification of two-dimensional NOE spectra via a combined linear and nonlinear least-squares fit. In: *Journal of Biomolecular NMR* 4 (5), S. 645–652. DOI: 10.1007/BF00404275.

Brunner, Konrad (2006): Modellierung, Strukturverbesserung und sequentielle Zuordnung als vollautomatische Module für die automatisierte Proteinstrukturbestimmung im Softwareprojekt AUREMOL.

Carrara, Enrico A.; Pagliari, Franco; Nicolini, Claudio (1993): Neural networks for the peak-picking of nuclear magnetic resonance spectra. In: *Neural Networks* 6 (7), S. 1023–1032. DOI: 10.1016/S0893-6080(09)80012-9.

Cavanagh, John (2007): Protein NMR spectroscopy. Principles and practice. 2. ed. Amsterdam, Boston: Academic Press. Online verfügbar unter http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10167118.

Cieslar, Christian; Marius Clore, G.; Gronenborn, Angela M. (1988): Computer-aided sequential assignment of protein 1H NMR spectra. In: *Journal of Magnetic Resonance (1969)* 80 (1), S. 119–127. DOI: 10.1016/0022-2364(88)90063-7.

Claridge, Timothy D. W. (2009): High-resolution NMR techniques in organic chemistry. 2nd ed. Amsterdam, London: Elsevier (Tetrahedron organic chemistry series, v. 27). Online verfügbar unter http://www.sciencedirect.com/science/bookseries/14601567/27.

Corne, Simon A.; Johnson, A.Peter; Fisher, Julie (1992): An artificial neural network for classifying cross peaks in two-dimensional NMR spectra. In: *Journal of Magnetic Resonance (1969)* 100 (2), S. 256–266. DOI: 10.1016/0022-2364(92)90260-E.

Denk, Winfried; Baumann, Rudolf; Wagner, Gerhard (1986): Quantitative evaluation of cross-peak intensities by projection of two-dimensional NOE spectra on a linear space spanned by a set of reference resonance lines. In: *Journal of Magnetic Resonance (1969)* 67 (2), S. 386–390. DOI: 10.1016/0022-2364(86)90449-X.

Eccles, C.; Guntert, P.; Billeter, M.; Wuthrich, K. (1991): Efficient analysis of protein 2D NMR spectra using the software package EASY. In: *Journal of Biomolecular NMR* 1 (2), S. 111–130.

Fahrmeir, Ludwig; Brachinger, Wolfgang (Hg.) (1996): Multivariate statistische Verfahren. 2., überarb. Aufl. Berlin: de Gruyter.

Fejzo, Jasna; ZOLNAI, ZSOLT; MACURA, SLOBODAN; Markley, John L. (1990): Quantitative evaluation of two-dimensional cross-relaxation NMR spectra of proteins. Interproton distances in Turkey ovomucoid third domain. In: *Journal of Magnetic Resonance (1969)* 88 (1), S. 93–110. DOI: 10.1016/0022-2364(90)90111-L.

Garrett, Daniel S.; Gronenborn, Angela M.; Clore, G. Marius (2011): A short recollection on the paper entitled "A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams" by D.S. Garrett, R. Powers, A.M. Gronenborn, and G.M. Clore [J. Magn. Reson. 95 (1991) 214-220]. In: *J. Magn. Reson.* 213 (2), S. 364–365. DOI: 10.1016/j.jmr.2011.08.009.

Geyer, M.; Neidig, K. P.; Kalbitzer, H. R. (1995): Automated Peak Integration in Multidimensional NMR Spectra by an Optimized Iterative Segmentation Procedure. In: *Journal of Magnetic Resonance, Series B* 109 (1), S. 31–38. DOI: 10.1006/jmrb.1995.1143.

Görler, A.; Kalbitzer, H. R. (1997): Relax, a flexible program for the back calculation of NOESY spectra based on complete-relaxation-matrix formalism. In: *J. Magn. Reson.* 124 (1), S. 177–188. DOI: 10.1006/jmre.1996.1033.

Hausser, Karl H.; Kalbitzer, Hans R. (1989): NMR für Mediziner und Biologen. Strukturbestimmung, Bildgebung, In-vivo-Spektroskopie. Berlin: Springer.

Hazewinkel, Michiel (2002): Encyclopaedia of mathematics. Berlin, New York: Springer-Verlag.

Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. (1983): Optimization by simulated annealing. In: *Science* 220 (4598), S. 671–680. DOI: 10.1126/science.220.4598.671.

Kleywegt, Gerard J.; Boelens, Rolf; Kaptein, Robert (1990): A versatile approach toward the partially automatic recognition of cross peaks in 2D 1H NMR spectra. In: *Journal of Magnetic Resonance (1969)* 88 (3), S. 601–608. DOI: 10.1016/0022-2364(90)90291-G.

Koradi, R.; Billeter, M.; Engeli, M.; Güntert, P.; Wüthrich, K. (1998): Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. In: *Journal of magnetic resonance (San Diego, Calif. : 1997)* 135 (2), S. 288–297. DOI: 10.1006/jmre.1998.1570.

Koutrouvelis, Ioannis A.; Meintanis, Simos (2002): Estimating the parameters of Poisson-exponential models. In: *Aust NZ J Stat* 44 (2), S. 233–245. DOI: 10.1111/1467-842X.00225.

Leutner, Michael; Gschwind, Ruth M.; Liermann, Jens; Schwarz, Christian; Gemmecker, Gerd; Kessler, Horst (1998): Automated backbone assignment of labeled proteins using the threshold accepting algorithm. In: *Journal of Biomolecular NMR* 11 (1), S. 31–43. DOI: 10.1023/A:1008298226961.

Levitt, Malcolm H. (2009): Spin dynamics. Basics of nuclear magnetic resonance. 2. ed., repr. with corr. Chichester: Wiley.

LIMPERT, ECKHARD; STAHEL, WERNER A.; ABBT, MARKUS (2001): Log-normal Distributions across the Sciences. Keys and Clues. In: *BioScience* 51 (5), S. 341. DOI: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.

Markley, John L.; Bax, Ad; Arata, Yoji; Hilbers, C. W.; Kaptein, Robert; Sykes, Brian D. et al. (1998): Recommendations for the presentation of NMR structures of proteins and nucleic acids (IUPAC Recommendations 1998). In: *Pure and Applied Chemistry* 70 (1). DOI: 10.1351/pac199870010117.

Munte, Claudia Elisabeth; Becker, Katja; Schirmer, Rolf Heiner; Kalbitzer, Hans Robert (2009): NMR assignments of oxidised thioredoxin from Plasmodium falciparum. In: *Biomolecular NMR assignments* 3 (2), S. 159–161. DOI: 10.1007/s12104-009-9163-7.

Neidig, K. P.; Bodenmueller, H.; Kalbitzer, H. R. (1984): Computer aided evaluation of two-dimensional NMR spectra of proteins. In: *Biochemical and Biophysical Research Communications* 125 (3), S. 1143–1150.

Neidig, K. P.; Geyer, M.; Görler, A.; Antz, C.; Saffrich, R.; Beneicke, W.; Kalbitzer, H. R. (1995): AURELIA, a program for computer-aided analysis of multidimensional NMR spectra. In: *Journal of Biomolecular NMR* 6 (3), S. 255–270. DOI: 10.1007/BF00197807.

Neidig, Klaus-Peter; Kalbitzer, Hans Robert (1990): Improved representation of two-dimensional NMR spectra by local rescaling. In: *Journal of Magnetic Resonance (1969)* 88 (1), S. 155–160. DOI: 10.1016/0022-2364(90)90119-T.

Neuhaus, David; Williamson, Michael P. (1989): The nuclear Overhauser effect in structural and conformational analysis. Weinheim, New York, New York, Cambridge: VCH.

Olson, JohnB.; Markley, JohnL. (1994): Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances. A demonstration of the connectivity tracing assignment tools (CONTRAST) software package. In: *Journal of Biomolecular NMR* 4 (3). DOI: 10.1007/BF00179348.

Pons, J. L.; Delsuc, M. A. (1999): RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. In: *Journal of Biomolecular NMR* 15 (1), S. 15–26.

Pons, J. L.; Malliavin, T. E.; Delsuc, M. A. (1996): Gifa V. 4: A complete package for NMR data set processing. In: *Journal of Biomolecular NMR* 8 (4), S. 445–452. DOI: 10.1007/BF00228146.

R.A. Fisher and the making of maximum likelihood 1912-1922 (1997): The Institute of Mathematical Statistics.

Ried, Andreas; Gronwald, Wolfram; Trenner, Jochen M.; Brunner, Konrad; Neidig, Klaus-Peter; Kalbitzer, Hans Robert (2004): Improved simulation of NOESY spectra by RELAX-JT2 including effects of J-coupling, transverse relaxation and chemical shift anisotropy. In: *Journal of Biomolecular NMR* 30 (2), S. 121–131. DOI: 10.1023/B:JNMR.0000048945.88968.af.

Schulte; Gorler; Antz; Neidig; Kalbitzer (1997): Use of global symmetries in automated signal class recognition by a bayesian method. In: *J. Magn. Reson.* 129 (2), S. 165–172. DOI: 10.1006/jmre.1997.1241.

Shen, Hengyi; Poulsen, Flemming M. (1990): Toward automated determination of buildup rates of nuclear overhauser effects in proteins, using symmetry projection operators. In: *Journal of Magnetic Resonance (1969)* 89 (3), S. 585–594. DOI: 10.1016/0022-2364(90)90343-8.

Sze, K. H.; Barsukov, I. L.; Roberts, G.C.K. (1995): Quantitative Evaluation of Cross-Peak Volumes in Multidimensional Spectra by Nonlinear-Least-Squares Curve Fitting. In: *Journal of Magnetic Resonance, Series A* 113 (2), S. 185–195. DOI: 10.1006/jmra.1995.1079.

Trenner, Jochen Markus (2006): Accurate proton-proton distance calculation and error estimation from NMR data for automated protein structure determination in AUREMOL. Univ., Diss--Regensburg, 2006. Online verfügbar unter http://www.opus-bayern.de/uni-regensburg/volltexte/2007/649/pdf/Dissertation_JM_Trenner.PDF.

Vuister, G. W.; Bax, A. (1993): Measurement of Two- and Three-Bond Proton to Methyl-Carbon J Couplings in Proteins Uniformly Enriched with 13C. In: *Journal of Magnetic Resonance, Series B* 102 (2), S. 228–231. DOI: 10.1006/jmrb.1993.1089.

Wolfram Gronwald; Hans Robert Kalbitzer (2004): Automated structure determination of proteins by NMR spectroscopy. In: *Progress in Nuclear Magnetic Resonance Spectroscopy* 44 (1–2), S. 33–96. DOI: 10.1016/j.pnmrs.2003.12.002.

Zimmerman, Diane; Kulikowski, Casimir; Wang, Lingze; Lyons, Barbara; Montelione, GaetanoT. (1994): Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods

from artificial intelligence. In: *Journal of Biomolecular NMR* 4 (2). DOI: 10.1007/BF00175251.

6 Zusammenfassung

Ziel der vorliegenden Arbeit war, die automatische Auswertung multidimensionaler NMR-Spektren von Proteinen zur Strukturbestimmung durch das Softwareprojekt AUREMOL zu verbessern.

Im Rahmen dieser Arbeit wurden Basisfunktionen durch effizientere, zentrale und rekursive Schnittstellen ersetzt. Dies ermöglicht eine einfache Einbindung der Funktionalitäten in zukünftige Module. Alle Module, welche in dieser Arbeit erstellt bzw. erweitert wurden, sind auch durch ihren rekursiven Ansatz in der Lage, Spektren beliebiger Dimension verarbeiten zu können.

Der Schwerpunkt der Arbeit lag in der Automatisierung und Optimierung der Signalerkennung und der Extraktion wichtiger Signaleigenschaften, da sie die Qualität aller Folgeroutinen bei der automatischen NMR-Strukturbestimmung wesentlich beeinflussen.

Das Modul Integration wurde in mehreren Punkten verbessert. So wurde eine automatische Anpassung der Integrationsschritte und der Größe des Integrationsbereichs für den Wachstumsalgorithmus während des Integrationsprozesses realisiert. Eine weitere wichtige Erweiterung bietet nun auch die Möglichkeit, Signale zu integrieren, deren Position sich nicht an einem Extremum der Signalform befindet. Durch weitere Optimierungen und eine Parallelisierung konnte die Performance des Algorithmus signifikant erhöht werden. Zudem ist es dem Modul nun durch den rekursiven Ansatz möglich, n-dimensionale Spektren zu integrieren. Um dem Benutzer die Möglichkeit zu bieten, die Ergebnisse der Integration noch einmal stichprobenartig verifizieren zu können, wurde eine Visualisierung der Volumen für ein- und zweidimensionale Spektren in AUREMOL umgesetzt. Da alle anteiligen Intensitäten am Volumen zu den jeweiligen Signalen in eine zusätzliche Struktur gespeichert werden, ist es möglich, schnelle Abfragen von Volumeninformationen anderer Module durchzuführen.

Durch die Festlegung der Positionen am Schwerpunkt der Volumenform eines Signals, konnte an synthetischen Spektren gezeigt werden, dass die Schwerpunktposition näher an der theoretisch bestimmten Position liegt, als die Position, welche durch das Extremum der Signalform definiert wird.

Im letzten Abschnitt dieser Arbeit wurde der Ansatz zur Bestimmung der Bayesschen Wahrscheinlichkeit, dass ein NMR-Signal aus der Peakliste ein Nutzsignal ist, verbessert. Die Güte der Diskriminierung der NMR-Signale in Nutzsignale und Störsignale konnte vor allem durch die Einführung neuer Signal-Eigenschaften verbessert werden. Ebenso konnte gezeigt werden, dass aus einem synthetischen Spektrum theoretische Verteilungen durch Simulated Annealing generiert werden können und dadurch eine ausreichende Diskriminierung gewährleisten. Dies ist vor allem dann nötig, falls zu wenig Nutzsignale in der Klasse Signal vorzufinden sind, denn dies erlaubt keine Erstellung von dem Modul benötigten geglätteten Verteilungen der verwendeten Eigenschaften. Die geglätteten Verteilungen wurden ebenfalls durch einen dynamischen Glättungsfilter erweitert, welcher erlaubt, Wahrscheinlichkeitsdichteverteilungen mit einer definierbaren Anzahl von Extrema zu generieren. Generell ist der erweiterte Ansatz der Berechnung der Bayesschen Wahrscheinlichkeit für Signale auf nicht zweidimensionale Spektren höherer Dimension anwendbar, falls keine Symmetrieeigenschaften verwendet werden.

7 Danksagung

An dieser Stelle möchte ich mich vor allem bei meiner Familie und den Kollegen bedanken, die mir die Kraft gaben, diese Arbeit gelingen zu lassen.

Besonders möchte ich mich bei meinem Betreuer Herrn Prof. Dr. Dr. Hans Robert Kalbitzer für die interessanten Themengebiete, die Betreuung dieser Arbeit und die fesselnden Diskussionen und Ideen bedanken, welche die Erstellung dieser Arbeit ermöglichten.

Ebenfalls besonders bedanken möchte ich mich bei Markus Beck Erlach für das Korrekturlesen und meinen Kollegen Jörg Köhler, Tobias Harsch, Michael Spörner, Dörte Repenning-Rochelt, Claudia Munte und allen anderen Kollegen für das hervorragende Arbeitsklima, das mich stets motivieren konnte, diese Arbeit fertig zu stellen.

Außerordentlich bedanken möchte ich mich bei meiner Lebensgefährtin Marion Veitl und meinen beiden Kindern Eric und Sarah Donaubauer, welche während dieser langen Zeit noch Geduld aufbrachten und mir darüber hinaus noch genug Kraft gaben, diese Arbeit doch noch fertig zu stellen.

8 Anhang

8.1 Die Bewegungsgraphen aus der horizontalen Bewegung des Signals „Peak 2“ und „Peak 1“ bei zweidimensionalen Spektren mit verschiedenen digitalen Auflösungen bei der Schwerpunktbestimmung

Es soll die horizontale Bewegung von „Peak 2“ mit verschiedenen digitalen Auflösungen der Frequenzdomänen dargestellt werden. Dazu werden die Bewegungsgraphen mit Spektren der digitalen Auflösung von 512x2048 Pixeln und mit Spektren mit den Auflösungen 2048x512 Pixeln dargestellt.

Dabei erfährt das sich bewegende Signal „Peak 2“ sowohl die Auswirkung der geringen Auflösung als auch der hohen Auflösung:

- Das Signal „Peak 2“ erfährt eine horizontale Bewegung entlang der direkten Dimension mit der **geringeren** 512 Pixel-Auflösung (Abb. 50).

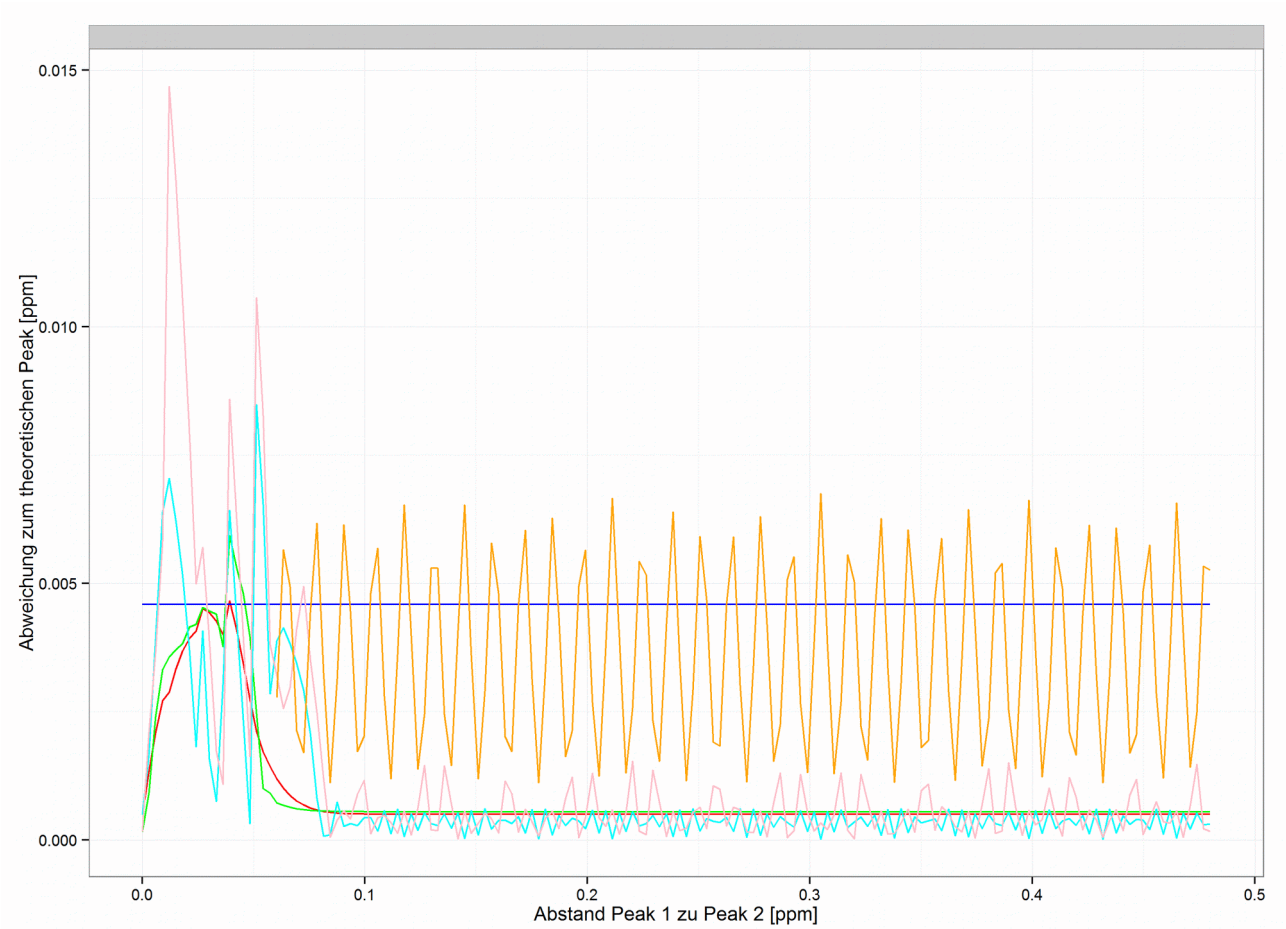

Abb. 50: Bewegung horizontal entlang der Frequenzachse mit der niedrigeren Auflösung von 512 Pixel. Die Spektren haben die digitalen Auflösung von 2048x512 Pixel ohne Rauschen analog zur Abb. 29. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.

- Das Signal „Peak 2“ erfährt wieder eine horizontale Bewegung entlang der direkten Dimension, jedoch hier mit der **höheren** Auflösung von 2048 Pixel (Abb. 51).

Das Wellenmuster ist im Falle der Maximum-Methode in Abb. 50 stark oszillierend, da die Repositionierung der Referenzposition während der Iteration in Richtung der 512 Pixel länger unterdrückt wird, d.h. die Position des Referenzsignals verweilt länger auf einem digitalen Pixel. Die beiden Methoden der Schwerpunktbildung sind im Vergleich zur „Maximum-Methode“ wesentlich unempfindlicher.

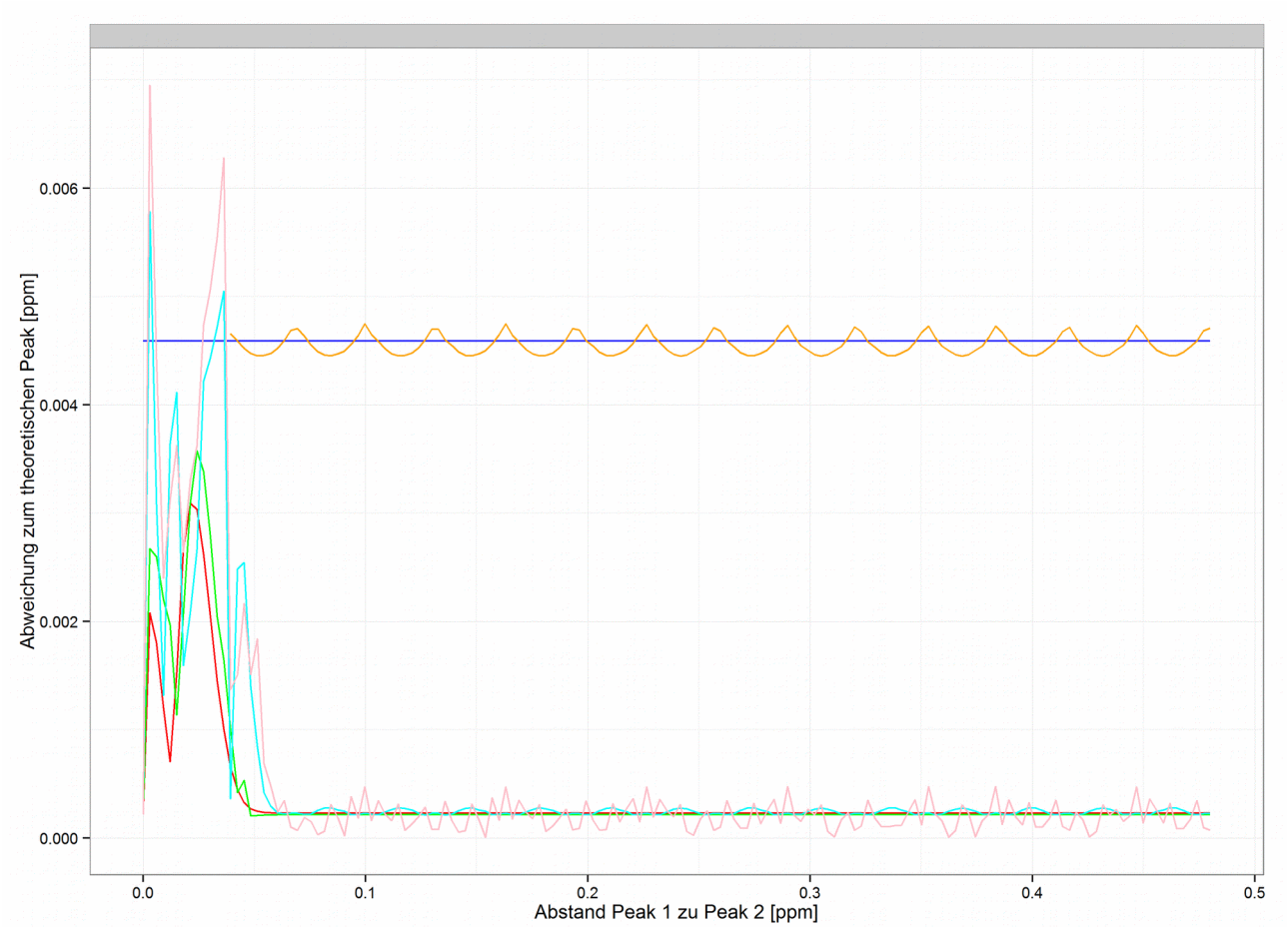

Abb. 51: Bewegung horizontal entlang der höheren Auflösung von 2048 Pixel von Spektren mit einer Auflösung von 512x2048 Pixeln ohne Rauschen analog zur Abb. 29. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.

Das Wellenmuster ist hier im Falle der Maximum-Methode in Abb. 51 weniger stark oszillierend als in Abb. 50, da die Repositionierung der Referenzposition während der Iteration in Richtung der 2048 Pixel rascher erfolgt. Das heißt die Position des Referenzsignals verweilt nun kürzer auf einem digitalen Pixel, da in diese Richtung für den gleichen Weg mehr Pixel vorhanden sind. Die beiden Methoden der Schwerpunktbildung sind dazu im Vergleich wieder unempfindlicher.

8.2 Gemittelte Bewegungsgraphen durch das Modul „Standardvolumen“ für eindimensionale Spektren bei der Schwerpunktbestimmung

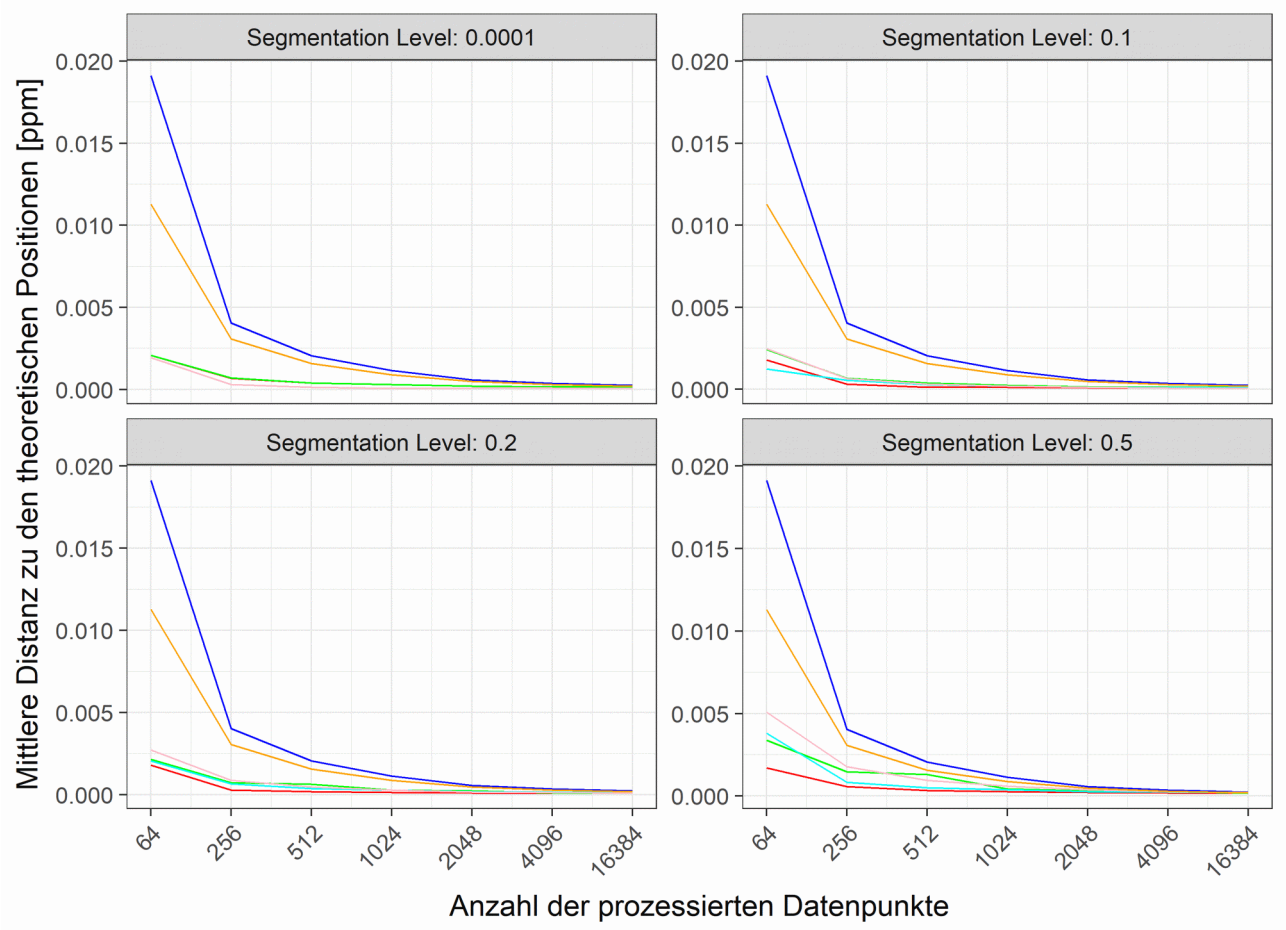

*Abb. 52: Eindimensionale Spektren **ohne Rauschen** analog zu Abb. 32. Die Signale „Peak 1“ und „Peak 2“ haben **gleiches Volumen**. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die „Schwerpunktbildung mit Abschneidung“ am Segmentierungslevel bei 0,1 (rechts oben) mit 0,0004 ppm erreicht. Die mittlere Abweichung der „Maximum-Methode“ beträgt 0,0032 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

Das Ergebnis aus Abb. 52 unterscheidet sich nur dahin gehend zu dem Experiment mit unterschiedlichen Volumen (Abb. 32), dass sich die beiden Signale näher aufeinander zubewegen können, bevor deren Extrema ununterscheidbar werden. Denn beide Signale

dominieren gleich stark bis kurz vor deren Überlappung. Damit erhöht sich lediglich die Anzahl der Werte (Iterationen), die einen Beitrag zum Modul „Standardvolumen“ liefern.

Ein weiteres zu testende Szenario war, den Spektren einen Rauschanteil hinzuzufügen. Dieser Rauschanteil hat Auswirkungen auf die Bildung der Volumen hinsichtlich der Segmentierungstiefe und der resultierenden Signalform. Diese wurde wiederum mit unterschiedlichen und gleichen Volumen der beiden Signale getestet.

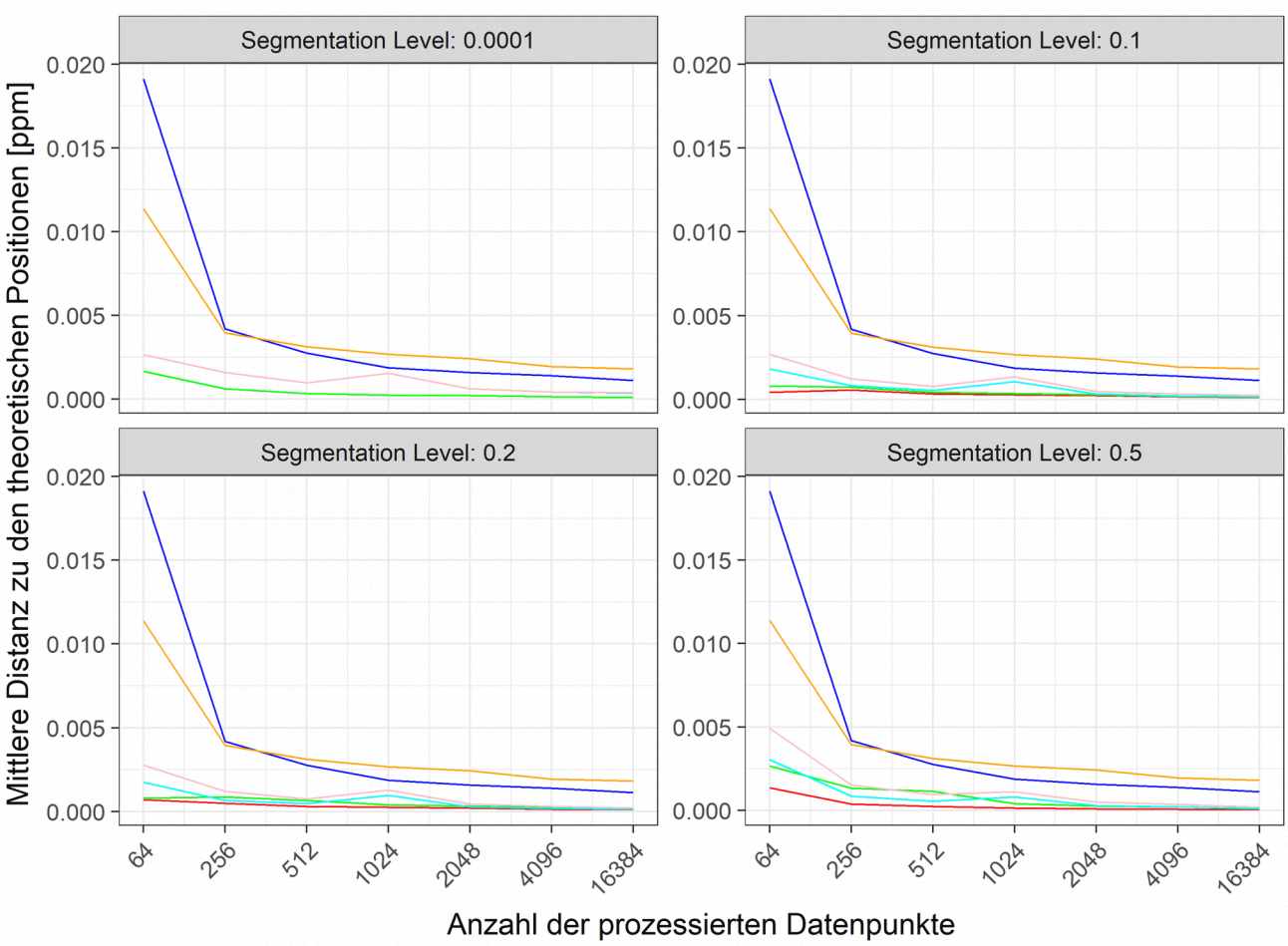

*Abb. 53: **Eindimensionale** Spektren **mit Rauschen** analog zu Abb. 32. Die Signale „Peak 1“ und „Peak 2“ haben **unterschiedliches Volumen**. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die „Schwerpunktbildung mit Abschneiden“ am Segmentierungslevel bei 0,2 (links unten) mit 0,0005 ppm erreicht. Die mittlere Abweichung der „Maximum-Methode“ beträgt 0,004 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*

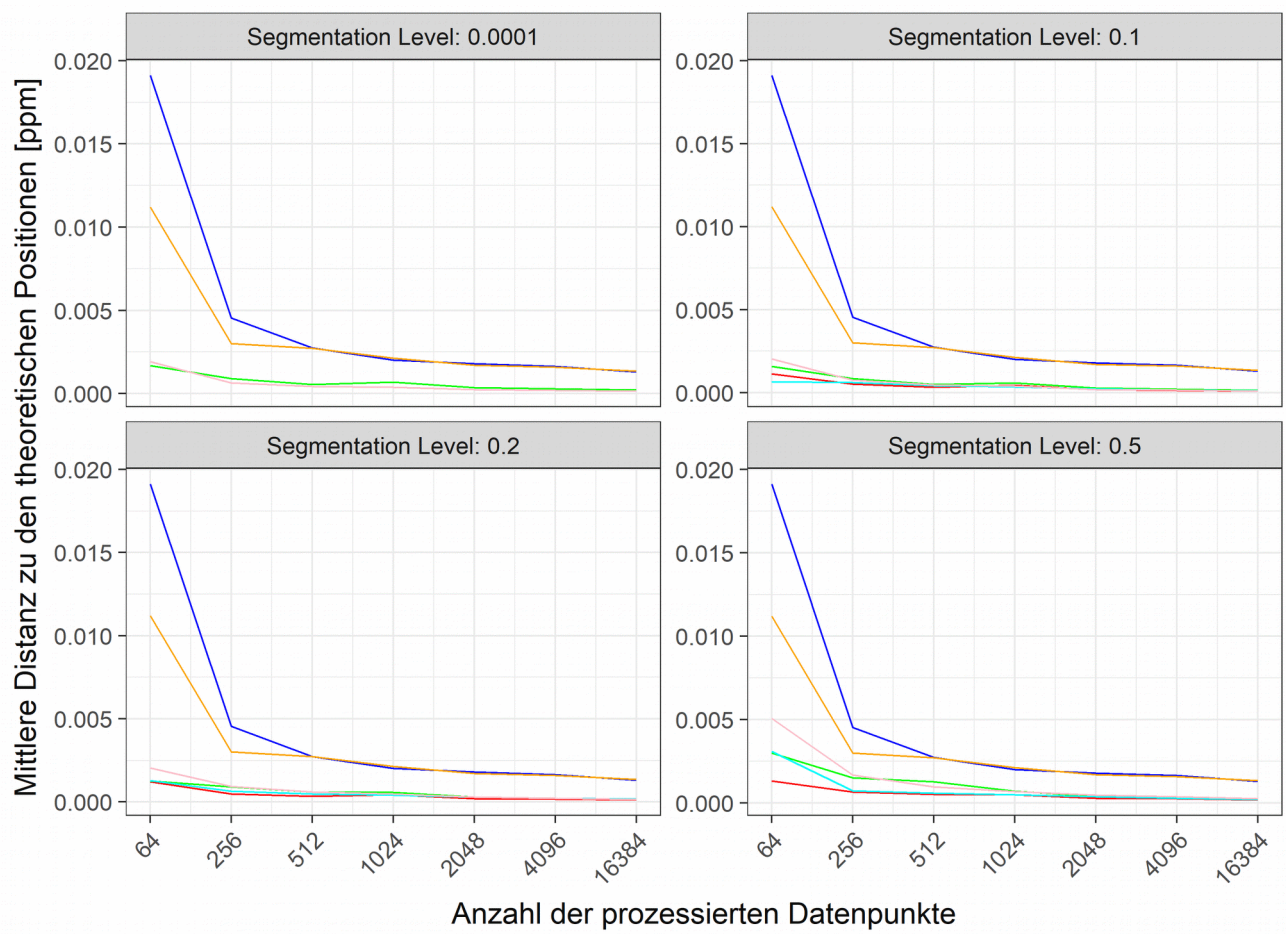

*Abb. 54: **Eindimensionale** Spektren **mit Rauschen** analog zu Abb. 32. Peak 1 und Peak 2 besitzen jedoch **gleich hohe Volumen**. Der kleinste mittlere Abstand zur theoretischen Position wurde durch die Methode Abschneiden am Segmentierungslevel bei 0,1 (rechts oben) mit 0,0004 ppm erreicht. Die mittlere Abweichung der Maximum-Methode beträgt 0,0041 ppm. In dieser Abbildung stellen die Kurven folgende Methoden dar: (blau) Maximum-Methode „Peak 1“, (orange) Maximum-Methode bewegender „Peak 2“, (grün) Schwerpunktmethode ohne Abschneidung „Peak 1“, (rosa) Schwerpunktmethode ohne Abschneidung bewegender „Peak 2“, (rot) Schwerpunktmethode mit Abschneidung „Peak 1“, (cyan) Schwerpunktmethode mit Abschneidung bewegender „Peak 2“.*