

Entwicklung eines Servers zur Analyse, Bewertung und Verfeinerung von Proteinstrukturen

Dissertation

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

der

Naturwissenschaftlichen Fakultät I – Biowissenschaften –

der Martin-Luther-Universität
Halle-Wittenberg,

vorgelegt

von Herrn Christoph Knorr

geb. am 20.11.1987 in Magdeburg

Gutachter: PD Dr. Iris Thondorf
Prof. Dr. Mike Schutkowski
Prof. Dr. Jens Wöhnert

Datum der Verteidigung: 8.7.2020

Science cannot be stopped. Man will gather knowledge no matter what the consequences – and we cannot predict what they will be.

- Linus Pauling, Vorlesung an der Yale Universität im Oktober 1947

Inhaltsverzeichnis

| | |
|--|-------------|
| Inhaltsverzeichnis | I |
| Abbildungsverzeichnis | III |
| Tabellenverzeichnis | V |
| Abkürzungsverzeichnis | VIII |
| Symbolverzeichnis | IX |
| 1. Einleitung und Zielstellung | 1 |
| 1.1 Historie, Bestimmung und Validierung von Proteinstrukturen | 1 |
| 1.2 Zielstellung | 3 |
| 2. Strukturbiochemische Parameter | 5 |
| 2.1 Analytische Untersuchungen von Proteinstrukturen | 5 |
| 2.1.1 Allgemeine Strukturdaten | 5 |
| 2.1.2 Eigenschaften von Aminosäuren | 10 |
| 2.2 Intramolekulare Interaktionen | 20 |
| 2.2.1 Disulfidbindungen | 20 |
| 2.2.2 Salzbrücken | 21 |
| 2.2.3 Kation- π -Wechselwirkungen | 25 |
| 2.2.4 π - π -Wechselwirkungen | 35 |
| 2.2.5 Hydrophobe Wechselwirkungen | 42 |
| 2.2.6 Wasserstoffbrücken des Protein-Rückgrates | 45 |
| 2.3 Sekundärstrukturen | 49 |
| 2.3.1 Bestehende Zuweisungsalgorithmen | 49 |
| 2.3.2 Vergleich verschiedener Zuweisungsalgorithmen | 53 |
| 2.3.3 Helices in Proteinstrukturen | 58 |
| 2.3.4 β -Stränge | 65 |
| 2.3.5 Verfeinerung von <i>tight Turns</i> | 68 |

| | | |
|-----------|--|------------|
| 2.4 | Validierung von Proteinstrukturen | 79 |
| 2.4.1 | Überprüfung von einfachen geometrischen Parametern | 79 |
| 2.4.2 | Analyse von Torsionswinkeln | 82 |
| 2.4.3 | Aminosäure-Volumina und sterische Probleme | 91 |
| 2.4.4 | Energiepotentiale | 96 |
| 2.4.5 | Überprüfung von Sekundärstrukturmerkmalen | 99 |
| 3. | Informatische Implementierung | 102 |
| 3.1 | Die Funktionsserver VAProTS und VAProIN | 102 |
| 3.2 | Datenserver VADoc, VASat und VAPara | 104 |
| 4. | Zusammenfassung | 106 |
| 5. | Ausblick | 108 |
| 6. | Material und Methoden | 110 |
| 6.1 | Datensätze | 110 |
| 6.2 | Suche nach hydrophoben Interaktionen mit Hilfe von <i>k-d</i> Bäumen | 110 |
| 6.3 | Zuweisung von Sekundärstrukturen | 111 |
| 6.4 | Statistische Auswertungen | 112 |
| 6.5 | Sekundärstrukturzuweisung mittels externer Programme | 113 |
| 6.6 | Informatische Module | 114 |
| 7. | Literaturverzeichnis | 115 |
| | Eidesstattliche Erklärung | 135 |
| | Danksagung | 136 |
| | Anhang A1 - Tabellen | 137 |
| | Anhang A2 - Abbildungen | 143 |
| | Anhang A3 - Elektronischer Anhang | 152 |
| | Lebenslauf | 153 |

Abbildungsverzeichnis

| | | |
|----------|--|----|
| Abb. 1: | Mittlerer und maximaler Gyrationradius bezogen auf die Kettenlänge (XRD) | 9 |
| Abb. 2: | α -Zeta-Winkel des Datensatzes XRD im Vergleich zu Daten von Morris <i>et al.</i> | 11 |
| Abb. 3: | Vorkommen von <i>cis</i> -Peptidbindungen in Abhängigkeit von der Auflösung von Kristallstrukturen [105] | 14 |
| Abb. 4: | Anteil Nicht-Xaa-Pro <i>cis</i> -Peptidbindungen in Abhängigkeit vom Veröffentlichungsjahr der Kristallstrukturen [105] | 15 |
| Abb. 5: | Verteilung der χ_1 -Winkel innerhalb XRD | 17 |
| Abb. 6: | Torsionswinkel einer Disulfid-Bindung | 21 |
| Abb. 7: | Mögliche Salzbrücken des Aspartats | 22 |
| Abb. 8: | Sequenzseparationen von Salzbrücken in XRD | 24 |
| Abb. 9: | Schematische Darstellung von Kation- π -Wechselwirkungen | 25 |
| Abb. 10: | Verschiedene Orientierungen von Arginin im Bezug zu Aromaten [123] | 27 |
| Abb. 11: | Verteilung der Theta-Winkel | 31 |
| Abb. 12: | Verteilung der γ -Winkel bei Kation- π -Wechselwirkungen von Arginin mit verschiedenen aromatischen Aminosäuren | 33 |
| Abb. 13: | Vergleich der gefundenen π - π -Wechselwirkungen in Abhängigkeit vom Abstand der Masseschwerpunkte | 36 |
| Abb. 14: | Mögliche Konformationen von π - π -Wechselwirkungen | 37 |
| Abb. 15: | Ebenenwinkel verschiedener π - π -Interaktionsnetzwerke bei verschiedenen Zentroid-Distanzen [128] | 38 |
| Abb. 16: | Ebenenwinkel in Abhängigkeit vom Abstand der Masseschwerpunkte (XRD) | 39 |
| Abb. 17: | Distanzkriterien zur Bestimmung von Wasserstoffbrücken | 46 |
| Abb. 18: | Abstände und Winkel von Wasserstoffbrücken des Datensatzes XRD | 46 |
| Abb. 19: | Sequenzseparationen aller gefundenen Wasserstoffbrücken des Datensatzes XRD | 48 |
| Abb. 20: | PyMOL-Darstellungen der neu-definierten <i>Turn</i> -Typen | 71 |
| Abb. 21: | Beispiele der Phi/Psi-Winkel der Aminosäuren in den <i>Turns</i> | 72 |
| Abb. 22: | Phi/Psi-Winkel von Asparagin und Lysin im <i>Turn</i> EE3 | 74 |
| Abb. 23: | Wasserstoffbrücken von Aspartat in HH2- <i>Turns</i> | 77 |

Abbildungsverzeichnis

| | | |
|------------|--|-----|
| Abb. 24: | Omega-Winkel im Datensatz XRD | 83 |
| Abb. 25: | Klassischer Ramachandran- <i>Plot</i> | 85 |
| Abb. 26: | Vergleich der PhiPsi- <i>Plots</i> von Lovell <i>et al.</i> [220] und den in dieser Arbeit durchgeführten Berechnungen | 86 |
| Abb. 27: | Phi/Psi-Diagramme der Sonderfälle Glycin, Prolin und Prä-Prolin | 88 |
| Abb. 28: | Seitenketten-Torsionswinkel-Diagramme für Arginin | 89 |
| Abb. 29: | Chi1/Chi2-Diagramme von Arginin und Leucin erstellt mittels ProCheck | 90 |
| Abb. 30: | Schema zur Anpassung kovalenter Radien bei der Volumenberechnung | 93 |
| Abb. 31: | Schematische Darstellung der Volumen-Berechnung überlappender Kugeln | 94 |
| Abb. 32: | Geometrische Kriterien zur Validierung von Helices | 100 |
| Abb. 33: | Mögliche Validierungskriterien von β -Strängen | 101 |
| Abb. 34: | Interaktive Interaktionskarten des Servers VAProIN | 104 |
| Abb. 35: | Graphische Repräsentation des Bewertungsserver-Netzwerkes | 109 |
| | | |
| Abb. A. 1: | Vergrößerung der Verteilungen der Zeta-C α -Winkel im Datensatz XRD | 143 |
| Abb. A. 2: | Verteilung der Zeta-C α -Winkel von Prolin im Datensatz XRD | 144 |
| Abb. A. 3: | Verteilung der berechneten Chi3-Torsionswinkel von möglichen Disulfidbindungen | 145 |
| Abb. A. 4: | θ -Winkel von Kation- π -Wechselwirkungen mit Tyrosin | 146 |
| Abb. A. 5: | Elektrostatische Oberflächenpotentiale aromatischer Systeme [125] | 147 |
| Abb. A. 6: | Niedrig-Energieübergänge von π - π -Wechselwirkungen des Benzen-Dimers in der Gasphase | 148 |
| Abb. A. 7: | Schematische Darstellung verschiedener C-O-N-Winkel von Wasserstoffbrücken des Proteinrückgrates | 149 |
| Abb. A. 8: | Pentapeptidmuster in β -Strängen [198] | 150 |
| Abb. A.9: | Phi/Psi- <i>Plots</i> für Glutamat, Histidin, Isoleucin und Valin | 151 |

Tabellenverzeichnis

| | | |
|-------------|--|----|
| Tabelle 1: | Vergleich der Aminosäurezusammensetzung im SwissProt-Sequenzdatensatz sowie im Proteinstrukturdatensatz XRay5000 | 6 |
| Tabelle 2: | Zusammenfassung berechneter Gyrationradien der genutzten Datensätze | 7 |
| Tabelle 3: | Zeta α -Winkel der L- und D-Konfigurationen der kanonischen Aminosäuren | 12 |
| Tabelle 4: | Statistische Verteilungen von D-Aminosäuren in den genutzten Datensätzen | 13 |
| Tabelle 5: | <i>cis</i> -Peptidbindungen innerhalb der genutzten Datensätze | 14 |
| Tabelle 6: | Protonierungszustände von Aspartat und Glutamat | 16 |
| Tabelle 7: | Mittelwerte und prozentuale Anteile der drei Chi1-Konformationen | 18 |
| Tabelle 8: | Prozentuale Verteilungen der Chi1-Konformationen einzelner Aminosäuren | 18 |
| Tabelle 9: | Prolin-Konformationen im Datensatz XRD | 19 |
| Tabelle 10: | Chi3-Torsionswinkel von möglichen Disulfidbindungen | 20 |
| Tabelle 11: | Salzbrücken und daran beteiligte Aminosäuren im Datensatz XRD | 23 |
| Tabelle 12: | Prozentuale Verteilung der Sequenzseparationen von Salzbrücken | 24 |
| Tabelle 13: | Zur Bestimmung von Masseschwerpunkten genutzte Atome | 26 |
| Tabelle 14: | Verhältnisse der verschiedenen θ -Winkel von Burley und Petsko und der aktuellen Arbeit | 26 |
| Tabelle 15: | An Kation- π -Wechselwirkungen beteiligte Aminosäuren | 29 |
| Tabelle 16: | Statistische Auswertung aller möglichen Kation- π -Interaktionen | 30 |
| Tabelle 17: | Verschiedene Konfidenz-Intervalle der θ -Winkel des Datensatzes XRD | 32 |
| Tabelle 18: | Prozentuale Verteilungen der Orientierungen der Guanidinium-Gruppe zum aromatischem Ringsystem | 34 |
| Tabelle 19: | Abstandsabhängigkeit der Konformation der Arginin-Interaktion mit aromatischen Systemen | 35 |
| Tabelle 20: | Idealwerte der Winkel bei π - π -Wechselwirkungen | 37 |
| Tabelle 21: | Vorkommen von aromatischen Aminosäuren in π - π -Wechselwirkungen | 41 |
| Tabelle 22: | Statistische Auswertung möglicher Kombinationen von aromatischen Interaktionen | 42 |
| Tabelle 23: | Genutzte Atome zur Bestimmung von hydrophoben Interaktionen | 43 |

Tabellenverzeichnis

| | | |
|-------------|--|----|
| Tabelle 24: | Statistische Auswertung von Kombinationen hydrophober Wechselwirkungen der Datensätze XRD, MBD und SMD | 44 |
| Tabelle 25: | Hydrophobe <i>Cluster</i> im Datensatz XRD | 45 |
| Tabelle 26: | Von P-SEA genutzte Kriterien zur Zuweisung von Sekundärstrukturen [151] | 52 |
| Tabelle 27: | Sekundärstrukturgehalt des Datensatzes XRD bestimmt mit verschiedenen Programmen | 55 |
| Tabelle 28: | Vergleich verschiedener bekannter Zuweisungsprogramme mit Angaben der Autoren und dem hier genutzten neuen Mechanismus | 55 |
| Tabelle 29: | Prozentuale Übereinstimmungen zwischen verschiedenen Zuweisungsprogrammen in verschiedenen älteren Studien | 56 |
| Tabelle 30: | Übereinstimmung von Zuweisungen einzelner Sekundärstrukturmerkmale | 57 |
| Tabelle 31: | Vorkommen verschiedener Helix-Arten innerhalb der Struktur-Datensätze | 58 |
| Tabelle 32: | Geometrische Parameter von α - und 3_{10} -Helices | 59 |
| Tabelle 33: | Statistische Auswertung von Helix-Präferenzen der einzelnen Aminosäuren für den Datensatz XRD | 60 |
| Tabelle 34: | Salzbrücken zwischen verschiedenen Sekundärstrukturen | 61 |
| Tabelle 35: | Positionsspezifische, statistische Präferenzen von Aminosäuren in Helices | 63 |
| Tabelle 36: | Statische Auswertung von Aminosäure-Präferenzen in β -Strängen | 65 |
| Tabelle 37: | Statistische Auswertung von positionsspezifischen Strang-Präferenzen von β -Strängen mit mindestens neun Aminosäuren für den Datensatz XRD | 67 |
| Tabelle 38: | π - π -Wechselwirkungen zwischen verschiedenen Sekundärstrukturen | 68 |
| Tabelle 39: | Zusammenfassung der hier definierten <i>Turn</i> -Typen | 70 |
| Tabelle 40: | Statistische Auswertung der <i>XX2-Turns</i> | 76 |
| Tabelle 41: | Positionsspezifische Aminosäure-Präferenzen in <i>XX3</i> - und <i>EE4-Turns</i> | 78 |
| Tabelle 42: | Zusammenfassung der berechneten maximalen RMSDs der Aminosäureseitenketten | 79 |
| Tabelle 43: | Mittlere RMS-Distanzen von planaren Gruppen | 80 |
| Tabelle 44: | Berechnete Zeta-Winkel zur Überprüfung der Chiralität von Aminosäuren | 81 |
| Tabelle 45: | Peptidbindungskonformationen der verschiedenen Datensätze | 84 |
| Tabelle 46: | Kovalente und van-der-Waals Radien ausgesuchter Elemente | 91 |
| Tabelle 47: | Berechnete kovalente Volumina der Aminosäure | 92 |
| Tabelle 48: | Vergleich von normierten Aminosäure-Volumina verschiedener Studien | 95 |

Tabellenverzeichnis

| | | |
|--------------|---|-----|
| Tabelle 49: | Standard-Parameter der entwickelten Energiepotentiale | 97 |
| Tabelle 50: | Geometrische Parameter zur Validierung von Sekundärstrukturmerkmalen | 99 |
| Tabelle 51: | Suchkriterien in der PDB zur Erstellung der Datensätze | 110 |
| Tabelle 52: | Wertebereiche der verschiedenen Helixkriterien | 111 |
| Tabelle 53: | Sonderfälle bei der Zuweisung von β -Strängen | 112 |
| Tabelle 54: | Externe Sekundärstrukturzuweisungsprogramme | 114 |
| Tabelle 55: | Im Zuge der Arbeit genutzte Python und C++ Module | 114 |
| | | |
| Tabelle A.1: | Vorkommen aller kanonischen Aminosäuren in den Datensätzen | 137 |
| Tabelle A.2: | Wasserstoffbrücken- <i>Cluster</i> im Datensatz XRD | 138 |
| Tabelle A.3: | Aminosäurezusammensetzungen des Datensatzes XRD im Vergleich zu Datensätzen verschiedener älterer Studien | 139 |
| Tabelle A.4: | Statistische Präferenzen von β -Strängen mit weniger als neun Aminosäuren | 140 |
| Tabelle A.5: | Seitenketten-Torsionswinkel, die zur Erstellung von Diagrammen genutzt wurden | 141 |
| Tabelle A.6: | Mögliche Server des zu entwickelnden Netzwerks | 142 |

Abkürzungsverzeichnis

| | |
|---------|--|
| g- | <i>gauche</i> ⁻ |
| g+ | <i>gauche</i> ⁺ |
| t | <i>trans</i> |
| Akz. | Wasserstoffbrücken-Akzeptor |
| CASP | <i>critical Assessment of protein structure prediction</i> |
| Cryo-EM | Cryo-Elektronenmikroskopie |
| Don. | Wasserstoffbrücken-Donor |
| MBD | <i>ModBase</i> -Datensatz |
| NMR | <i>Nuclear Magnetic Resonance</i> |
| RMSD | <i>root mean square distance</i> |
| SAXS | <i>Small angle Xray-Scattering</i> |
| SMD | <i>SwissModel</i> -Datensatz |
| SSE | Sekundärstrukturelement |
| TRD | Trainingsdatensatz |
| TED | Test-Datensatz |
| VADoc | <i>Validation and Analysis Documentation</i> |
| VAPara | <i>Validation and Analysis Parameters</i> |
| VAProTS | <i>Validation and Analysis of Protein Tertiary Structures</i> |
| VAProIN | <i>Validation and Analysis of Protein Interaction Networks</i> |
| VASat | <i>Validation and Analysis Statistics</i> |
| XRD | XRay5000-Datensatz |

Symbolverzeichnis

| | |
|-----------|---------------------|
| # | Anzahl |
| \bar{x} | Mittelwert |
| σ | Standardabweichung |
| % | prozentualer Anteil |
| R_g | Gyrationsradius |

1. Einleitung und Zielstellung

1.1 Historie, Bestimmung und Validierung von Proteinstrukturen

Ausgehend von den ersten Röntgen- und elektronenmikroskopischen Analysen von Tropomyosin in den späten Vierzigern zeigte sich, dass Proteine keine langen ungeordneten Ketten von Aminosäuren sind, sondern klar strukturierte Bereiche besitzen [1]. Mit der Postulierung von α -Helices und β -Strängen im Jahr 1951 durch Pauling, kam der große Durchbruch auf diesem Gebiet [2–4].

Mit Hilfe dieser Erkenntnisse wurden verschiedene Strukturebenen, namentlich Primär-, Sekundär-, Tertiär- und Quartärstruktur, definiert. Die Primärstruktur entspricht dabei der Aminosäuresequenz, während unter Sekundärstrukturen geordnete Bereiche wie α -Helices, 3_{10} -Helices oder β -Stränge zusammengefasst werden. Die Tertiärstruktur beschreibt dann die Gesamtheit der dreidimensionalen Faltung eines Proteins und die Quartärstruktur die Bildung von Proteinkomplexen [5, 6].

Durch die Entwicklung der Röntgenkristallstrukturanalyse wurde es möglich, Proteinstrukturen experimentell zu bestimmen. Bis zum heutigen Tag findet diese Variante der Strukturbestimmung von Proteinen weitläufige Anwendung in den verschiedensten Bereichen [7, 8]. Weitere oft verwendete Methoden sind die *Nuclear Magnetic Resonance* Spektroskopie (NMR-Spektroskopie) [9–11] und die Cryo-Elektronenmikroskopie (Cryo-EM) [12, 13] sowie Kombinationstechniken [14, 15]. Um die Menge an aufklärbaren Proteinstrukturen zu vergrößern, wurden zusätzliche Techniken, z.B. Elektronen-Kristallographie [16], Neutronen-Diffraktion [17] oder Fiber-Diffraktion [18], entwickelt, mit deren Hilfe etwaige Schwächen der anderen Methoden ausgeglichen werden sollten.

Aufgeklärte Proteinstrukturen, und dabei vor allem der Zusammenhang zwischen der Struktur sowie der Funktion eines Proteins, spielen in den meisten biologischen, medizinischen und pharmazeutischen Forschungsbereichen eine wichtige Rolle [19]. Sie können essentielle Erkenntnisse bei der Bekämpfung von Krebs [20–23] oder bei der Entwicklung von Medikamenten zur Behandlung von Ebola [24–26], Malaria [27, 28] oder Hepatitis C [29, 30] liefern.

Die Aufklärung von Proteinstrukturen ist zwar eine sehr zeitaufwändige und kostspielige Prozedur, aber dennoch unverzichtbar [31]. Leider kann die Anzahl der in der PDB

Einleitung

hinterlegten Strukturen nicht mit der Menge an bekannten Proteinsequenzen, die vor allem durch die ständige Weiterentwicklung von Sequenzierungsverfahren [32–34] sehr schnell wächst, Schritt halten [35, 36]. Des Weiteren kann nur eine begrenzte Anzahl an Proteinstrukturen tatsächlich experimentell aufgeklärt werden, so dass andere Verfahren in Betracht gezogen werden müssen [37, 38].

Zu diesen Alternativen zählt die *in silico* Generierung von Proteinstrukturmodellen [38]. Dabei finden Methoden wie die Homologie-Modellierung, das *Threading* oder die *ab initio* Erstellung von Strukturen Anwendung [39–42]. Die Erstellung der Modelle beruht dabei entweder auf der Suche nach homologen Proteinen deren Struktur bereits bekannt ist [43, 44], der Erkenntnis, dass Proteine mit gleicher Funktion ähnliche Faltungen annehmen [45–47] oder der Suche nach der energetisch günstigsten Faltung [48]. Durch die ständige Weiterentwicklung und Automatisierung dieser Verfahren stellen sie gute Alternativen zur klassischen experimentellen Strukturaufklärung dar [49–52].

Proteinstrukturen, die *in silico* generiert wurden, können für eine große Anzahl an Untersuchungen genutzt werden. So können sie zur Aufklärung von Protein-Protein- oder Protein-Ligand-Interaktionen beitragen [53–56]. Des Weiteren können sie zur Identifizierung von Bindestellen [57, 58] oder zur struktur-basierten Entwicklung von Medikamenten [59–63] genutzt werden. Aufgrund der vielfältigen Anwendungsmöglichkeiten von *in silico* erstellten Proteinstrukturen können automatisierte Datenbanken wie das *Swiss Model Repository* [64] oder die *ModBase* [65] Alternativen zur PDB darstellen.

Bevor diese Proteinstrukturmodelle jedoch für Analysen genutzt werden können, muss eine Validierung erfolgen. Dazu wurden zum einen die CASP (*Critical Assessment of Protein Structure Prediction*) Experimente ins Leben gerufen [66], deren Bewertungskriterien zwar sehr komplex sind [67–69], sich jedoch stark auf den Vergleich mit Kristallstrukturen konzentrieren [70, 71]. Auf der anderen Seite wurden verschiedene Programme und/oder Server wie ProCheck [72] oder WhatCheck [73] zur Analyse von geometrischen Problemen entwickelt. Weitere Server sind ANOLEA [74] oder ProSA [75] zur energetischen Validierung sowie komplexere Programme wie Qmean [76] oder MolProbity [77].

Obwohl die Anzahl an Bewertungsprogrammen in den letzten Jahren immer mehr zunahm [78–83], stellt die Modellvalidierung noch immer ein sehr komplexes Verfahren dar. Die Hauptprobleme sind dabei die Fixierung der Programme auf nur wenige Aspekte einer Proteinstruktur, z.B. nur geometrische Analysen bei ProCheck, die zum Teil fehlende oder

unzulängliche Dokumentation sowie veraltete Standard-Parameter. Außerdem korrelieren die Ergebnisse der verschiedenen Server und Programme nur in den seltensten Fällen miteinander, wodurch eine umfassende Aussage über die Qualität einer Proteinstruktur erschwert wird. Dementsprechend muss eine Möglichkeit gefunden werden, diese Probleme anzugehen und eine umfassende Bewertung unter Nutzung eines Programms oder Servers zu generieren.

1.2 Zielstellung

Das Ziel der vorliegenden Arbeit war die Entwicklung eines Servers zur Analyse, Bewertung und Verfeinerung von Proteinstrukturen. Dazu sollte eine Vielzahl von Analyse- und Validierungsfunktionen bereit gestellt werden, um eine umfassende quantitative und qualitative Bewertung aus einer Hand zu ermöglichen.

Um dies zu bewerkstelligen, mussten zuerst verschiedene Berechnungsalgorithmen entwickelt werden, mit deren Hilfe eine umfassende Analyse von Proteinstrukturen durchgeführt werden kann. Im Anschluss daran sollte eine neue Methode zur Zuweisung von Sekundärstrukturen etabliert werden, um Problemen bei schon bekannten Programmen zu begegnen. Insbesondere Regionen zwischen α -Helices und β -Strängen sollten dabei gesondert untersucht werden, um etwaige Regelmäßigkeiten zu identifizieren. In einem letzten Schritt mussten dann Standard-Parameter für die geometrische, sterische und energetische Bewertung von Proteinstrukturen ermittelt werden. Dabei sollte ein besonderes Augenmerk auf die Aminosäure-Seitenketten, Sekundärstrukturmerkmale und Torsionswinkel gelegt werden. Unter Anwendung und Erweiterung bereits bekannter Validierungssysteme sollten alle Aspekte, die einen Einfluss auf die Qualität von Proteinstrukturen haben, bewertet werden können. Zusätzlich sollten mehrere neue Energiepotentiale sowie Bewertungsmechanismen für Sekundärstrukturelemente etabliert werden.

Die bestimmten Parameter, der neu entwickelte Zuweisungsmechanismus sowie die etablierten Berechnungsalgorithmen sollten dann in Form von Funktionsservern einer breiten Öffentlichkeit zur Verfügung gestellt werden. Des Weiteren sollten Datenserver eine umfangreiche Dokumentation inklusive der Standard-Parameter und statistische

Einleitung

Auswertungen einzelner Datensätze bereitstellen, um Nutzern einen Überblick über die zugrunde liegenden Funktionen zu geben.

2. Strukturbiologische Parameter

Die Validierung von Proteinstrukturen basiert in den meisten Fällen auf einem Vergleich von Standardparametern mit denen des zu untersuchenden Modells. Diese Standardparameter werden dabei in den meisten Fällen direkt aus experimentell bestimmten Proteinstrukturen gewonnen. Im Folgenden soll die Auswahl, Berechnung und Interpretation solcher Daten erläutert werden. Die Reihenfolge der einzelnen Abschnitte ist dabei abhängig von der Komplexität der Implementierung, beginnend mit sehr einfachen Analyse-Modulen bis hin zu sehr komplexen Validierungsprogrammen. Sämtliche Untersuchungen wurden mit Hilfe der in Abschnitt 6.1 beschriebenen Datensätze durchgeführt. Dabei diente der Strukturdatensatz XRay5000 (XRD) zur Bestimmung der Standard-Parameter, während die Modell-Datensätze aus *ModBase* (MDB) und dem *SwissModel Repository* (SMD) zur Überprüfung einiger geometrischer Parameter herangezogen wurden.

2.1 Analytische Untersuchungen von Proteinstrukturen

Um eine *in silico* generierte Proteinstruktur ausführlich validieren zu können, ist es wichtig zu allererst eine tiefgreifende Analyse durchzuführen. Dabei wird nach *cis* Peptidbindungen, D-Aminosäuren und anderen einfachen Parametern gesucht, mit deren Hilfe die Qualität des vorliegenden Modells eingeschätzt werden kann. Durch die Auswertung solcher Parameter kann schnell bestimmt werden, ob die generierte Proteinstruktur im Einklang mit der Literatur steht.

2.1.1 Allgemeine Strukturdaten

Um einen Einblick in die grundsätzliche Qualität von Modellen zu bekommen, sollten in einem ersten Schritt allgemeine Strukturparameter analysiert werden. Diese ermöglichen es, etwaige Besonderheiten der jeweiligen Struktur zu identifizieren und mit Literaturangaben zu vergleichen.

Aminosäurezusammensetzung

Als erstes kann die Aminosäurezusammensetzung eines Modells mit denen von bekannten Proteinstrukturen und -sequenzen verglichen werden. Die dafür nötigen Daten entstammen zum einen der UniProt/SwissProt [84] und zum anderen einem selbsterstellten Datensatz

Strukturbiologische Parameter

hochaufgelöster Proteinstrukturen aus der PDB (XRD, siehe Material und Methoden). Anhand der Daten in Tabelle 1 wird deutlich, dass die grundsätzlichen Verteilungen im Bezug auf einige Aminosäuren deutliche Unterschiede aufweisen. Auffällig ist, dass es bei aromatischen Aminosäuren (Trp und Tyr) zu größeren prozentualen Abweichungen kommt. Möglicherweise sorgt der stabilisierende Effekt von aromatischen Interaktionen [85] für eine bessere Kristallisierbarkeit von Proteinstrukturen und führt somit auch zu einer Erhöhung des Anteils an aromatischen Aminosäuren innerhalb der PDB [35]. Des Weiteren fällt auf, dass Methionin in Proteinsequenzen deutlich häufiger vorkommt als in gelösten Proteinstrukturen. Diese Diskrepanz könnte auf die Funktion des Methionins als Translationsinitiator [86] und die co-translationale Abspaltung des N-terminalen Methionins [87] zurückzuführen sein. Außerdem könnte ein direkter Zusammenhang zwischen den unterschiedlichen Anteilen von Aspartat und Methionin aus der post-translationalen Umwandlung von Methionin zu Aspartat bestehen [88]. Ein weiterer Grund könnte der häufige Austausch von Methionin zu Selenomethionin bei der Röntgenkristallstrukturaufklärung sein [19].

Tabelle 1: Vergleich der Aminosäurezusammensetzung im SwissProt-Sequenzdatensatz sowie im Proteinstrukturdatensatz XRay5000

| Aminosäure | SwissProt [89] | XRD | Aminosäure | SwissProt [89] | XRD |
|------------|----------------|------|--------------|----------------|------|
| Alanin | 8,2% | 8,0% | Leucin | 9,6% | 9,2% |
| Arginin | 5,5% | 5,1% | Lysin | 5,8% | 6,0% |
| Asparagin | 4,0% | 4,5% | Methionin | 2,4% | 1,9% |
| Aspartat | 5,4% | 6,0% | Phenylalanin | 3,8% | 4,1% |
| Cystein | 1,3% | 1,3% | Prolin | 4,7% | 4,7% |
| Glutamin | 6,7% | 6,6% | Serin | 6,6% | 6,1% |
| Glutamat | 3,9% | 3,9% | Threonin | 5,3% | 5,5% |
| Glycin | 7,0% | 7,2% | Tryptophan | 1,0% | 1,5% |
| Histidin | 2,2% | 2,3% | Tyrosin | 2,9% | 3,8% |
| Isoleucin | 5,9% | 5,6% | Valin | 6,8% | 6,9% |

Gyrationsradien von Proteinstrukturen

Der Gyrationsradius ist definiert als RMSD (*root mean square distance*) aller Atome vom Zentroid der Proteinstruktur [90] und kann als mittlerer (Mittelwert über alle drei Raumrichtungen) oder maximaler Radius (größte Ausdehnung in eine der drei Raumrichtungen) angegeben werden.

Strukturbiologische Parameter

In Tabelle 2 sind die mit einem vereinfachten System berechneten Gyrationradien für alle in dieser Arbeit genutzten Datensätze zusammengefasst. Dabei wurden die drei Raumkoordinaten (x, y und z) über alle Atome der Proteinstruktur gemittelt und im Anschluss die Differenz zwischen dem maximalen Wert einer Raumkoordinate und seinem Mittelwert berechnet. Diese Mittelwerte entsprechen dann den Gyrationradien in den drei Raumrichtungen und je nach Angabe kann dann der Mittelwert der drei Radien oder der Maximalwert angegeben werden. Beim Vergleich von mittlerem und maximalem Gyrationradius sind nur geringe Differenzen bei den Mittel- sowie Minimalwerten zu erkennen. Die Maximalwerte weisen beim deutliche Unterschiede (145,97 Å für mittleren zu 278,78 Å für maximalen Gyrationradius) auf. Hauptgründe für diese Diskrepanz liegen in Strukturen die sehr langgestreckte, z.B. Teile des Exocyst (PDB: 2B1E [91]), oder nicht komplett globuläre Konformationen annehmen, z.B. Transaminase von *Pseudomonas* st. ACC (PDB: 4UHO [92]).

Tabelle 2: Zusammenfassung berechneter Gyrationradien der genutzten Datensätze

| Art des Radius | Wert | XRD |
|--------------------------|------------|----------|
| Mittlerer Gyrationradius | Mittelwert | 24,77 Å |
| | Maximum | 145,97 Å |
| | Minimum | 5,33 Å |
| Maximaler Gyrationradius | Mittelwert | 29,78 Å |
| | Maximum | 278,78 Å |
| | Minimum | 7,23 Å |

Abhängigkeit von Kettenlänge und Gyrationradius

Bei einer Auftragung von Gyrationradien gegen die dazugehörigen Kettenlängen (Abb. 1) wird ein nicht-linearer Zusammenhang dieser beiden Parameter erkennbar. Das Verhältnis von Kettenlänge und Gyrationradius entspricht dabei dem von Flory postulierten Potenzgesetz [93]:

$$R_g = R_0 N^{\nu} \quad (1)$$

Der Gyrationradius R_G einer Proteinstruktur ist dabei abhängig von dem Vorfaktor R_0 , der Aminosäureanzahl N der Struktur und dem Exponenten ν (Formel (1)). Die genauen Werte von R_0 und ν wurden in vielen Arbeiten diskutiert. So postulierte Flory einen Wert von 0,6 für ν [93], der durch SAXS (*Small angle Xray-Scattering*) Messungen von Kohn *et al.* an denaturierten Proteinen [94] und smFRET (*single-molecule Förster resonance energy*

Strukturbiologische Parameter

transfer) Messungen von Hofmann *et al.* bestätigt werden konnte [95]. Gong *et al.* fanden im Gegensatz dazu anhand von Monte-Carlo Simulationen mit simuliertem *Annealing* Werte von 2,83 Å für R_0 und 0,34 (Kristallstrukturen vor der Simulation) bzw. 0,40 (nach der Simulation) für v [96]. Eine statistische Analyse von rund 37000 Proteinstrukturen von Hong und Lei bestätigte den Wert von 0,40 für v [97]. Sie konnten dabei zeigen, dass diese Werte nahezu unabhängig von der Faltungsklasse der Proteine (All- α , All- β , α/β oder unstrukturierte Proteine) sind [97]. Ausführliche statistische Untersuchungen von Tanner an Protein-Komplexen verdeutlichten, dass Monomere und Oligomere grundsätzlich denselben Regeln im Bezug auf den Gyrationradius folgen und ein Wert von 0,40 für v unabhängig vom Oligomer-Zustand genutzt werden kann [98]. Sowohl Hong und Lei [97] als auch Tanner [98] nutzten dabei den von Wilkins *et al.* mittels NMR-Messungen bestimmten Wert von 2,0 Å für R_0 [99].

Strukturbiologische Parameter

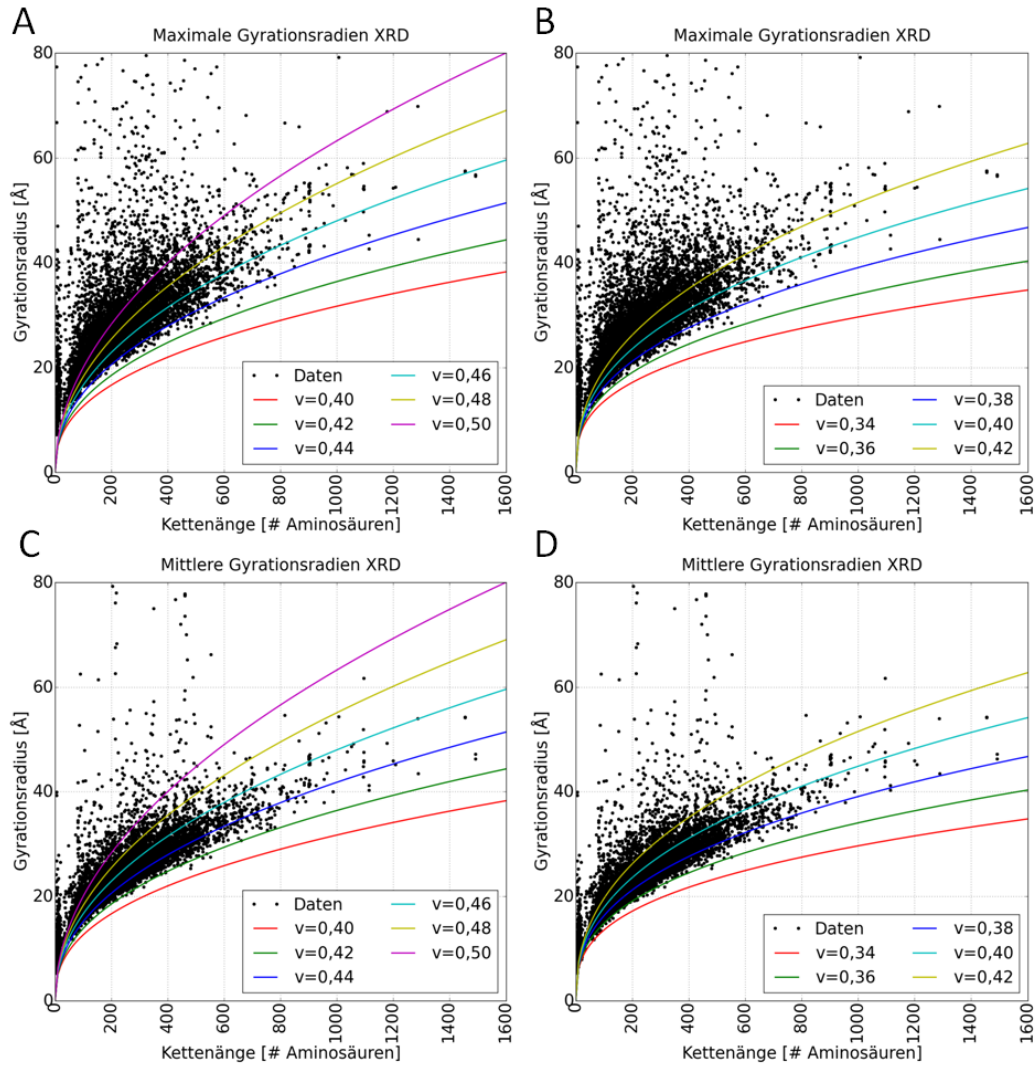


Abb. 1: Mittlerer und maximaler Gyrationradius bezogen auf die Kettenlänge (XRD)

In den Diagrammen sind die berechneten mittleren (C und D) und maximalen (A und B) Gyrationradien gegen die dazugehörigen Struktur­längen für den Datensatz XRD aufgetragen. Auf der linken Seite wurde ein R_0 von 2,0 Å und auf der rechten Seite ein R_0 von 2,83 Å für den nicht-linearen *Fit* verwendet. Die schwarzen Punkte sind die berechneten Gyrationradien und die farbigen Linien zeigen verschieden nicht-lineare *Fits* mit verschiedenen Werten für v .

Die berechneten Radien für den genutzten Datensatz XRD (Abb. 1) zeigen ein etwas anderes Verhalten als bisher postulierte Daten. So konnten für die maximalen Gyrationradien Ideal-Parameter von 2,0 Å für R_0 und 0,46 für v bestimmt werden (Abb. 1 A), während für die mittleren Gyrationradien die Werte von Gong *et al.* [96] ($R_0 = 2,83$ Å und $v = 0,40$, Abb. 1 D) bestätigt werden konnten.

Alles in allem konnte gezeigt werden, dass eine grundsätzliche Unterscheidung von mittlerem und maximalem Gyrationradius zwingend erforderlich ist. Des Weiteren konnte

das Potenzgesetz von Flory [93] sowohl für den mittleren als auch für den maximalen Gyrationradius bestätigt werden. Der Vergleich mit älteren Studien konnte für die mittleren Gyrationradien die Parameter, die von Gong *et al.* mit Hilfe von Monte-Carlo Simulationen bestimmt wurden [96] reproduzieren. Die Ergebnisse der statistischen Analysen von Hong und Lei [97] sowie Tanner [98] ($R_0 = 2,0 \text{ \AA}$ und $\nu = 0,40$) konnten allerdings nicht nachgewiesen werden. Da sowohl Hong und Lei [97] als auch Tanner [98] in ihren Arbeiten Datensätze mit sehr hohen Sequenzidentitäten (90% bei Tanner bzw. Nutzung aller PDB-Strukturen bei Hong und Lei) nutzten, könnten die sehr ähnlichen Proteinstrukturen, im Gegensatz zu den Analysen von Gong *et al.* [96] und der vorliegenden Arbeit, zu einem statistischen Fehler geführt haben.

2.1.2 Eigenschaften von Aminosäuren

Neben den allgemeinen Eigenschaften einer Proteinstruktur spielen auch die individuellen Parameter der einzelnen Aminosäuren eine wichtige Rolle im Bezug auf die Qualität des Modells. Dabei sind vor allem die Konfiguration (D- oder L-Aminosäure), der Omega-Winkel (*cis*- oder *trans*-Peptidbindung), Protonierungsgrad von Aspartat und Glutamat sowie (wenn vorhanden) der Chi1-Winkel der einzelnen Aminosäuren von besonderem Interesse.

D-Aminosäuren

Um die Konfiguration einer Aminosäure zu bestimmen, kann der pseudo-Torsionswinkel Zeta [100] genutzt werden. Morris *et al.* definierten so einen Wert von $33,81^\circ \pm 4,14^\circ$ als Standard [100]. Der Vergleich der berechneten Zeta-Winkel entlang der Bindungen $\text{C}\alpha\text{-N-C-C}\beta$ des Datensatzes XRD (unter Ausschluss von Prolin) mit den Ergebnissen von Morris *et al.* (Abb. 2) zeigt, dass die neueren Werte eine deutliche engere Verteilung annehmen. Auch bei einem genaueren Blick auf die Verteilung der neu berechneten Werte (Abb. A. 1) konnte dieses Verhalten bestätigt werden.

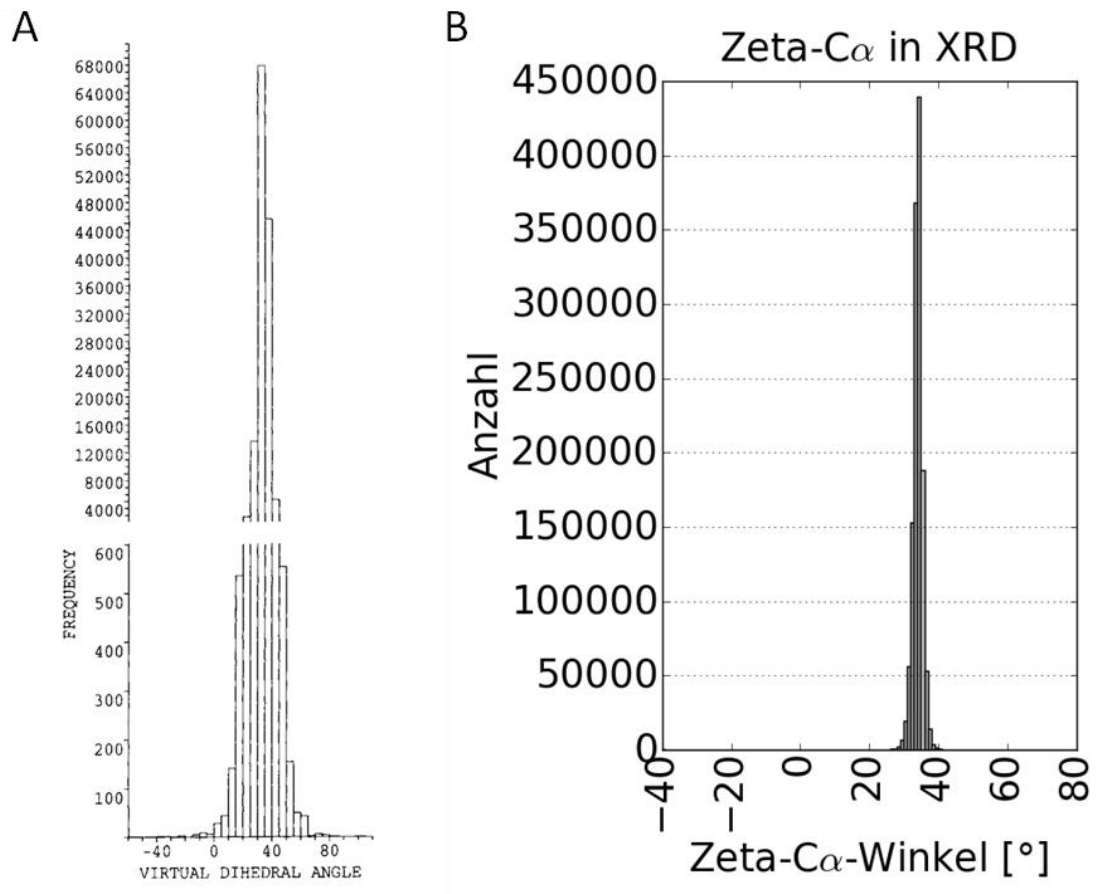


Abb. 2: **C α -Zeta-Winkel des Datensatzes XRD im Vergleich zu Daten von Morris *et al.***

Die Diagramme zeigen die berechneten C α -Zeta-Winkel von Morris *et al.* (A) [100], sowie des Datensatzes XRD (B). Die neuen Daten zeigen eine deutlich engere Verteilung als die der älteren Studie.

Es zeigte sich, dass für fast alle L-Aminosäuren dieser Winkel standardmäßig bei rund $34^\circ \pm 1,43^\circ$ (Tabelle 3) bzw. für Prolin bei $38,4^\circ \pm 1,12^\circ$ liegt (Abb. A. 2). Der leichte Unterschied zwischen diesen Werten basiert auf der Ringgeometrie des Prolins, die dafür sorgt, dass das C β -Atom etwas anders zu den Substituenten rund um das C α -Atom steht. In Anbetracht der Tatsache, dass Morris *et al.* einen allgemeingültigen Zeta-Winkel angeben [100], der mit seiner Standardabweichung sogar in den Bereich des Prolins reicht, wurde von der Verwendung dieser grundsätzlichen Definition Abstand genommen. Stattdessen wurde zumindest Prolin bei der Entwicklung des Servers separat betrachtet.

Strukturbiologische Parameter

Tabelle 3: Zeta α -Winkel der L- und D-Konfigurationen der kanonischen Aminosäuren

Die Zahlen der vierten Spalte geben die Anzahl an gefundenen Proteinstrukturen sowie in Klammern die Anzahl der gefundenen Moleküle der jeweiligen D-Aminosäure an. Ein X in der letzten Spalte bedeutet, es wurden zu wenige Strukturen gefunden um eine Auswertung der Daten zu ermöglichen. Alle Zeta-Winkel wurden entlang der Achse α -N-C- β berechnet.

| Aminosäure | Zeta α [°] | D-Form (PDB ID) | # Strukturen | Zeta α [°] |
|------------|-------------------|-----------------|--------------|-------------------|
| ALA | 34,0±0,9 | DAL | 69 (184) | -34,0±3,7 |
| ARG | 34,1±1,4 | DAR | 10 (33) | -33,1±5,6 |
| ASN | 33,9±1,3 | DSG | 5 | X |
| ASP | 34,0±1,5 | DAS | 11 (24) | -34,5±2,3 |
| CYS | 34,1±1,5 | DCY | 10 (24) | -34,8±1,6 |
| GLN | 34,1±1,2 | DGN | 8 | X |
| GLU | 34,0±1,3 | DGL | 31 (65) | -34,4±1,3 |
| HIS | 34,0±1,5 | DHI | 10 (17) | -34,8±1,3 |
| ILE | 33,8±1,4 | DIL | 5 | X |
| LEU | 34,2±1,3 | DLE | 20 (194) | -35,1±2,7 |
| LYS | 34,0±1,3 | DLY | 14 (45) | -33,9±3,6 |
| MET | 34,2±1,2 | MED | 5 | X |
| PHE | 34,0±1,5 | DPN | 14 (26) | -34,4±3,2 |
| PRO | 38,4±1,1 | DPR | 21 (144) | -39,0±1,7 |
| SER | 34,3±1,4 | DSN | 21 (69) | -34,2±1,6 |
| THR | 34,1±1,5 | DTH | 9 (19) | -35,2±1,7 |
| TRP | 34,1±1,6 | DTR | 14 (41) | -34,6±1,6 |
| TYR | 34,1±1,5 | DTY | 14 (44) | -34,6±2,2 |
| VAL | 33,7±1,4 | DVA | 20 (109) | -33,8±2,9 |

Berechnungen des α -Zeta-Winkels für D-Aminosäuren zeigten, dass sich die Beträge der Winkel in den meisten Fällen nur wenig von denen der L-Aminosäuren unterscheiden (Tabelle 3). Bei D-Cystein, -Histidin, -Leucin und -Threonin wurden jedoch Werte ermittelt die deutlicher von denen der L-Konfigurationen abweichen. Aufgrund der relativ geringen Datenmenge sowie den großen Standardabweichungen, die sich für die D-Aminosäuren ergeben wurde von einer separaten Betrachtung der vier Aminosäuren bei der Entwicklung des Validierungsprogramms jedoch abgesehen.

Schreiner *et al.* konnten bei Moleküldynamik-Simulationen mit NAMD zeigen, dass D-Aminosäuren innerhalb von Helices zu einem Knick innerhalb der Sekundärstruktur führen können [101]. Sie konnten in verschiedenen Datensätzen von Strukturen aus der PDB knapp 1% Kristallstrukturen mit fehlerhaften D-Aminosäuren¹ finden [101]. Eine statistische

¹ Dabei handelt es sich um modellierte D-Aminosäuren die eigentlich keine sind.

Auswertung der hier genutzten Datensätze führte für die Kristallstrukturen und die Modelle der *ModBase* zu deutlich geringeren Anteilen. Im *SwissModel*-Datensatz konnte jedoch ein deutlich erhöhter Anteil fehlerhafter D-Aminosäuren nachgewiesen werden, was zu einigen Problemen bei der Nutzung dieser Modelle im Zuge von weiteren wissenschaftlichen Untersuchungen führen könnte.

Tabelle 4: Statistische Verteilungen von D-Aminosäuren in den genutzten Datensätzen

| | XRD | MBD | SMD |
|-----------------------------------|---------|---------|---------|
| # Aminosäuren | 1442992 | 668196 | 387485 |
| # D-Aminosäuren | 19 | 2 | 321 |
| % D-Aminosäuren | 0,0013% | 0,0003% | 0,0828% |
| # Strukturen | 5545 | 3196 | 1245 |
| # Proteinketten mit D-Aminosäuren | 17 | 2 | 243 |
| % Proteinketten mit D-Aminosäuren | 0,31% | 0,06% | 19,52% |

***cis*-Peptidbindungen**

Untersuchungen von Stewart *et al.* (154 Strukturen) [102], MacArthur und Thornton (903 Proline) [103] sowie Jabs *et al.* (571 Strukturen) [104] verschieden großer Datensätze von Strukturen aus der PDB konnten einen Anteil von 5% bis 7% Xaa-Pro bzw. 0,03% bis 0,05% Nicht-Xaa-Pro *cis*-Peptidbindungen innerhalb von Proteinstrukturen nachweisen. Eine aktuelle Studie von Croll [105] konnte jedoch zeigen, dass der Anteil von fehlerhaften Nicht-Xaa-Pro *cis*-Peptidbindungen² mit sinkender Auflösung der Kristallstrukturen signifikant ansteigt (Abb. 3). Er diskutierte dabei, dass diese Fehler unter anderem auf die automatisierte Nutzung von Programmen wie MolProbity [106] zurückzuführen sein könnten. Aus diesem Grund versuchen Arbeitsgruppen wie das Richardson-Lab Hilfestellungen zu geben, um eine Überinterpretation von Nicht-Xaa-Pro *cis*-Peptidbindungen zu verhindern [107].

² Gemeint sind *trans*-Peptidbindungen die nur leicht verdreht sind und deshalb als *cis*-Peptidbindungen identifiziert wurden.

Strukturbiologische Parameter

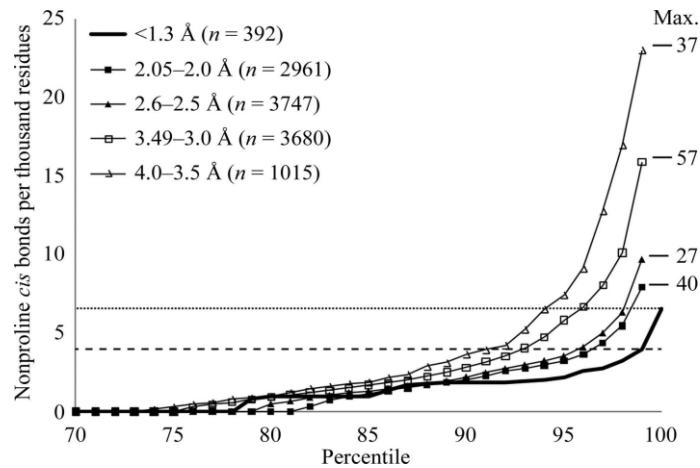


Abb. 3: Vorkommen von *cis*-Peptidbindungen in Abhängigkeit von der Auflösung von Kristallstrukturen [105]

Die Abbildung ist der Studie von Croll [105] entnommen und zeigt wie der Anteil an nicht-Xaa-Pro *cis*-Peptidbindungen mit sinkender Auflösung der Kristallstrukturen ansteigt.

Durch die Berechnung des Rückgrat-Torsionswinkels Omega ($C\alpha-C-N-C\alpha$) können *cis*- und *trans*-Peptidbindungen voneinander unterschieden werden [100]. Für die Datensätze XRD, MBD und SMD wurden alle Peptidbindungen hinsichtlich ihrer Konformation untersucht und etwaige *cis*-Peptidbindungen statistisch ausgewertet (Tabelle 5). Dabei konnte analog zur statistischen Untersuchung der PDB durch Schreiner *et al.* [101] ein Anteil von ungefähr 45% Strukturen mit *cis*-Peptidbindungen gefunden werden. Bei den Modell-Datensätzen war der Anteil mit 28,8% bei *ModBase* geringer und mit 59,3% bei *SwissModel* höher als bei den Kristallstrukturen. Das spricht für eine verminderte Qualität der automatisch generierten Modelle von *SwissModel*.

Tabelle 5: *cis*-Peptidbindungen innerhalb der genutzten Datensätze

| Wert | XRD | MBD | SMD |
|---|---------|--------|--------|
| # Peptidbindungen | 1472256 | 665000 | 386240 |
| # Xaa-Pro Peptidbindungen | 69196 | 35880 | 20726 |
| % Xaa-Pro an Peptidbindungen | 4,7% | 5,4% | 5,4% |
| # <i>Cis</i> -Peptidbindungen | 4246 | 1096 | 1519 |
| % <i>Cis</i> -Peptidbindungen an Peptidbindungen | 0,29% | 0,16% | 0,39% |
| # Xaa-Pro <i>cis</i> -Peptidbindungen | 3359 | 530 | 500 |
| % Xaa-Pro <i>cis</i> an <i>cis</i> -Peptidbindungen | 79,1% | 48,4% | 32,9% |
| % Xaa-Pro <i>cis</i> an Xaa-Pro-Peptidbindungen | 4,85% | 1,48% | 2,41% |
| % nicht Xaa-Pro <i>cis</i> an <i>cis</i> -Peptidbindungen | 20,9% | 52,6% | 77,1% |
| % nicht Xaa-Pro <i>cis</i> an Peptidbindungen | 0,06% | 0,09% | 0,26% |
| # Strukturen mit <i>cis</i> -Peptidbindung | 2538 | 794 | 738 |
| % Strukturen mit <i>cis</i> -Peptidbindung | 45,8% | 28,8% | 59,3% |

Die Daten in Tabelle 5 zeigen, dass ungefähr aller 5% Xaa-Pro-Peptidbindungen als *cis*-Peptidbindungen innerhalb der Datensätze vorliegen, und entsprechen den Ergebnissen von verschiedenen vorhergehenden Untersuchungen. Der Anteil an nicht Xaa-Pro *cis*-Peptidbindungen ist mit 0,06% nah an den postulierten Werten von 0,03-0,05% für tatsächliche *cis*-Peptidbindungen und spiegelt so nicht die von Croll [105] beschriebene Überinterpretation wider (Abb. 4). Bei den Modell-Datensätzen sind die Anteile zum Teil deutlich größer. So konnten bei den Modellen 0,09% (MBD) und 0,26% (SMD) nicht Xaa-Pro *cis*-Peptidbindungen gefunden werden. Während die Ergebnisse der *ModBase* noch in einem annehmbaren Rahmen liegen, sind die Ergebnisse von *SwissModel* sehr problematisch. Diese Modelle sollten vor etwaiger Nutzung noch einmal detailliert überprüft werden.

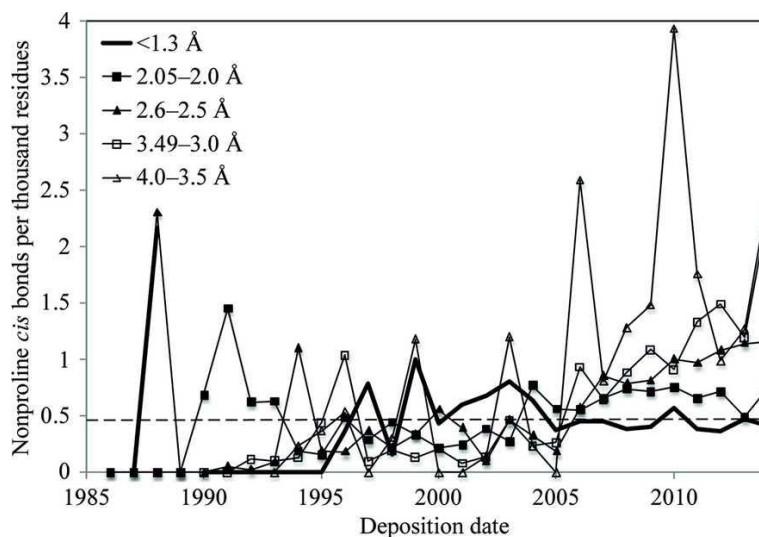


Abb. 4: Anteil Nicht-Xaa-Pro *cis*-Peptidbindungen in Abhängigkeit vom Veröffentlichungsjahr der Kristallstrukturen [105]

Die Abbildung wurde der Studie von Croll [105] entnommen und zeigt, wie der Anteil von Nicht-Xaa-Pro *cis*-Peptidbindungen in Proteinstrukturen im Laufe der letzten Jahre immer mehr ansteigt.

Protonierung von Aspartat und Glutamat

Einige kanonische Aminosäuren besitzen Seitenketten, die in wässriger Lösung einem Säure-Base-Gleichgewicht unterliegen, d.h. sie sind protonierbar. Dieses Verhalten ist sehr gut verstanden (pKa-Werte), jedoch kommen im Inneren von Proteinen andere Eigenschaften zum Tragen. Die Säure-Base-Eigenschaften können sich deutlich von denen in wässriger Lösung unterscheiden und sind von der individuellen Umgebung der Aminosäure abhängig. So können sie an der Ausbildung von Wasserstoffbrücken sowie an der Katalyse beteiligt sein. Daher ist es wichtig, den Protonierungsgrad von titrierbaren Aminosäuren zu

bestimmen. Diesem Umstand Rechnung tragend wurde ein System entwickelt, mit dessen Hilfe der Protonierungszustand von Asp und Glu direkt anhand von Bindungslängen bestimmt werden kann. Dabei wird der Umstand genutzt, dass eine Doppelbindung deutlich kürzer ist als eine Einfachbindung. Dementsprechend gilt analog zu schon früher bestimmten Werten, dass ein Sauerstoffatom von Asp oder Glu protoniert ist, wenn die Länge der dazugehörigen C-O-Einfachbindung $\geq 1,3 \text{ \AA}$ und die benachbarte C-O-Doppelbindung $\leq 1,22 \text{ \AA}$ ist. Auf eine geometrische Analyse von Histidin, Arginin oder Lysin wurde verzichtet, da die hier genutzten Strukturen eine zu geringe Auflösung zeigten um zwischen den einzelnen Protonierungszuständen zu differenzieren [108, 109].

Tabelle 6: Protonierungszustände von Aspartat und Glutamat

| Aminosäure | XRD | MBD | SMD |
|-------------------------|-------|-------|-------|
| # Aspartate | 88015 | 40478 | 24768 |
| # Protonierte Aspartate | 90 | 0 | 1 |
| % Protonierte Aspartate | 0,10% | 0,00% | 0,01% |
| # Glutamate | 96118 | 32609 | 20458 |
| # Protonierte Glutamate | 84 | 0 | 5 |
| % Protonierte Glutamate | 0,09% | 0,00% | 0,02% |

Tabelle 6 zeigt die Anzahl protonierter Asp und Glu zusammen mit ihren prozentualen Anteilen am Gesamtvorkommen innerhalb der genutzten Datensätze. Um den genauen Protonierungsgrad von Aspartat oder Glutamat zu bestimmen, sind Röntgenkristallstrukturen mit sehr hohen Auflösungen nötig [108, 109]. Dementsprechend sind die geringen Anzahlen an protonierten Asp und Glu (Tabelle 6) nicht überraschend.

Chi1-Winkel

Chi1-Winkel können, berechnet entlang der Bindungen N-C α -C β -C γ , grundsätzlich eine von drei Konformationen, namentlich *gauche*⁻ (g⁻, $\bar{x} = -60^\circ$, Auswahlbereich: -120° bis 0°), *gauche*⁺ (g⁺, $\bar{x} = 60^\circ$, Auswahlbereich: 0° bis 120°) sowie *trans* (t, $\bar{x} = \pm 180^\circ$, Auswahlbereich: $\pm 120^\circ$ bis $\pm 180^\circ$), annehmen. Diese stimmen mit den energetisch idealen Isomeren von zwei kovalent-gebundenen, sp³-hybridisierten Kohlenstoff-Atomen überein [110]. Eine statistische Untersuchung von 19 Kristallstrukturen von Janin *et al.* fand eine allgemeine prozentuale Verteilung von 54% *gauche*⁻, 11% *gauche*⁺ und 35% *trans* [111]. Sie konnten dabei zusätzlich nachweisen, dass diese prozentuale Verteilung stark von der jeweiligen Aminosäure abhängig ist [111]. Diese Abhängigkeit konnte durch eine weitere statische

Strukturbiologische Parameter

Analyse von Bhat *et al.* (23 Strukturen) bestätigt werden [112]. McGregor *et al.* konnten in einer Analyse von 61 PDB-Strukturen zusätzlich zur Aminosäure-Abhängigkeit auch eine Sekundärstruktur-Abhängigkeit der Verteilung der Chi1-Winkel feststellen [113].

Prolin stellt eine Ausnahme von diesem Verhalten dar, da seine Seitenketten-Torsionswinkel aufgrund seiner Topologie deutlich von denen anderer Aminosäuren abweichen. Bereits Ramachandran konnte mit Hilfe der Seitenketten-Torsionswinkel von Prolin zwei unterschiedliche Konformationen berechnen [114]. Eine neuere Studie von Ho *et al.* konnte sogar drei verschiedene Prolin-Konformationen anhand der Chi1- und Chi2-Torsionswinkel in Proteinstrukturen identifizieren [115].

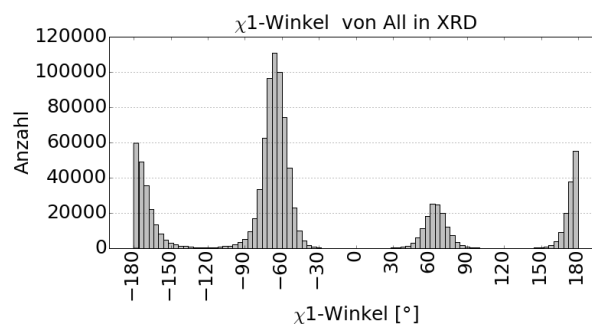


Abb. 5: Verteilung der Chi1-Winkel innerhalb XRD

Das Histogramm zeigt die Verteilung der berechneten Chi1-Winkel für den Datensatz XRD. Es sind drei Maxima zu erkennen, die den Konformationen *gauche*⁻ (rund -65°), *gauche*⁺ (rund 65°) sowie *trans* (±180°) zugeordnet werden können. Die meisten Aminosäuren nehmen eine *gauche*⁻ Konformation an, gefolgt von *trans* und *gauche*⁺ Konformationen.

Eine Analyse der allgemeinen Verteilungen der Chi1-Winkel (berechnet als Winkel entlang der Achse N-C α -C β -C γ), unter Ausschluss von Prolin, bestätigte drei Maxima in den Bereichen -65°, 65° sowie 180° (Abb. 5). Dieses Verhalten sowie die berechneten Mittelwerte (siehe auch Tabelle 7) entspricht nahezu den postulierten Daten, allerdings ist zu beachten, dass in der Anfangszeit die Nomenklatur von *gauche*⁻/*gauche*⁺ nicht einheitlich war. Die berechneten prozentualen Verteilungen sind relativ ähnlich zu den allgemeinen Daten von Janin *et al.* [111] sowie den allgemeinen Ergebnissen von McGregor *et al.* [113]. Die von Willard *et al.* genutzten prozentualen Verteilungen in ihrem VADAR-Webserver [116] zeigen jedoch deutliche Unterschiede bei *gauche*⁺ und *trans*.

Strukturbiologische Parameter

Tabelle 7: Mittelwerte und prozentuale Anteile der drei Chi1-Konformationen

Die angegebenen Prozentzahlen von Morris *et al.* [100] sind der Ausgabe des VADAR-Webservers [116] entnommen.

| | Janin [111] | McGregor [113] | Morris <i>et al.</i> [100] | XRD |
|--------------------------------------|-------------|----------------|----------------------------|----------------|
| \bar{x} <i>gauche</i> ⁻ | -70° ± 21° | | -66,7° ± 15,0° | -65,7° ± 10,2° |
| \bar{x} <i>gauche</i> ⁺ | 61° ± 25° | | 64,1° ± 15,7° | 64,3° ± 10,6° |
| \bar{x} <i>trans</i> | 190° ± 24° | | 183,6° ± 16,8° | 182,1° ± 11,2° |
| % <i>gauche</i> ⁻ | 54% | 49% | 55,00% | 53,01% |
| % <i>gauche</i> ⁺ | 11% | 13% | 20,00% | 12,43% |
| % <i>trans</i> | 35% | 33% | 25,00% | 34,56% |

Die Überprüfung der Verteilungen der Chi1-Konformationen der einzelnen Aminosäuren konnte die von Janin *et al.* [111] beschriebenen Verhältnisse zu großen Teilen bestätigen (Tabelle 8). Größere Abweichungen finden sich lediglich bei Serin und Threonin. Serin nimmt in den hier beschriebenen Analysen deutlich häufiger eine *gauche*⁺ Konformation an als von Janin *et al.* beschrieben [111]. Die Unterschiede beim Threonin resultieren aus Unterschieden in der Berechnung der Winkel. Während Janin *et al.* den Winkel entlang der Bindungen N-C α -C β -O γ_1 berechneten [111], wurde in der vorliegenden Arbeit die Bindungen N-C α -C β -C γ_2 genutzt.

Tabelle 8: Prozentuale Verteilungen der Chi1-Konformationen einzelner Aminosäuren

| Aminosäure | Janin [111] | | | XRD | | |
|--------------|-------------|-----|-----|-------|-------|-------|
| | g- | g+ | t | g- | g+ | t |
| Arginin | | | | 57,5% | 8,5% | 34,0% |
| Asparagin | 51% | 15% | 34% | 56,0% | 14,0% | 30,0% |
| Aspartat | 51% | 15% | 34% | 50,6% | 17,5% | 31,9% |
| Cystein | 57% | 16% | 27% | 51,6% | 20,5% | 27,9% |
| Glutamin | 44% | | | 62,0% | 7,6% | 30,4% |
| Glutamat | | | | 57,9% | 8,6% | 33,5% |
| Histidin | 45% | 10% | 45% | 55,3% | 12,3% | 32,4% |
| Isoleucin | | | | 77,0% | 13,3% | 9,7% |
| Leucin | 60% | 4% | 36% | 66,8% | 1,2% | 32,0% |
| Lysin | | | | 57,7% | 7,8% | 34,5% |
| Methionin | | | | 63,6% | 6,6% | 29,8% |
| Phenylalanin | 57% | 13% | 31% | 54,6% | 11,2% | 34,2% |
| Serin | 34% | 24% | 38% | 28,5% | 47,6% | 12,9% |
| Threonin | 48% | 39% | 13% | 49,5% | 7,8% | 42,7% |
| Tryptophan | 57% | 13% | 31% | 50,0% | 15,0% | 35,0% |
| Tyrosin | 57% | 13% | 31% | 53,8% | 11,7% | 34,5% |
| Valin | 66% | 21% | 13% | | | |

Strukturbiologische Parameter

Mit Hilfe der Seitenketten-Torsionswinkel kann man die unterschiedlichen Konformationen von Prolin bestimmen [115]. Ausgehend von den Daten von Ho *et al.* [115] wurden die drei Konformationen *planar* ($-10^\circ < \text{Chi1} < 10^\circ$), *Up* ($\text{Chi1} \leq -10^\circ$) und *Down* ($\text{Chi1} \geq 10^\circ$) definiert. Die Daten in Tabelle 9 zeigen, dass die Proline in den hier genutzten Datensätzen ähnliche Konformationen aufweisen wie sie von Ho *et al.* [115] beschrieben wurden. Auch die prozentualen Verteilungen sind sehr ähnlich zu den früheren Ergebnissen.

Tabelle 9: Prolin-Konformationen im Datensatz XRD

| Konformation | Ho [115] | XRD | MBD | SMD |
|---------------|----------|-------|-------|-------|
| # Planar | 216 | 124 | 135 | 31 |
| % Planar | 5,0% | 4,6% | 6,9% | 2,9% |
| # <i>Up</i> | 2160 | 1406 | 1099 | 527 |
| % <i>Up</i> | 50,4% | 51,7% | 55,9% | 49,4% |
| # <i>Down</i> | 1913 | 1187 | 733 | 508 |
| % <i>Down</i> | 44,6% | 43,7% | 37,2% | 47,7% |

2.2 Intramolekulare Interaktionen

Ein wichtiger Aspekt im Hinblick auf die Stabilität von Proteinstrukturen sind intramolekulare Wechselwirkungen. Dementsprechend wurden im Rahmen der vorliegenden Arbeit Disulfidbindungen, Salzbrücken, Wasserstoffbrücken des Proteinrückgrates, aromatische π - π -, Kation- π - sowie hydrophobe Wechselwirkungen analysiert.

2.2.1 Disulfidbindungen

Die Identifizierung von Disulfidbindungen erfolgt in zwei Schritten. Zuerst wird der Abstand zweier S γ -Atome bestimmt. Liegt dieser Abstand in einem Bereich von 1,85 Å bis 2,25 Å [100] wird im Anschluss der Chi3-Torsionswinkel der S γ -S γ -Bindung (Abb. 6) berechnet [100]. Die Auswertung aller möglichen Disulfidbindungen ergab, dass der berechnete Winkel entweder zwischen -102,8° und -69,2° oder zwischen 74,7° und 112,2° ($\bar{x} \pm 1,5 \cdot \sigma$ der neu berechneten Werte)³ liegen muss, damit von einer Disulfid-Bindung gesprochen werden kann (Tabelle 10 und Abb. A. 3).

Tabelle 10: Chi3-Torsionswinkel von möglichen Disulfidbindungen

| | Morris [100] | XRD |
|---|--------------|-------------|
| # Rechtsgängig | X | 956 |
| \bar{x} Chi3 Rechtsgängig | 96,8±10,1° | 93,4±12,5° |
| # Linksgängig | X | 1137 |
| \bar{x} Chi3 Linksgängig | -85,8±8,6° | -86,0±11,2° |
| % Strukturen mit Disulfidbindungen | 21,3% | 15,3% |

Die Daten in Tabelle 10 zeigen eine leichte Präferenz für die Ausbildung von linksgängigen Disulfidbindungen im Vergleich zur Ausbildung von rechtsgängigen Disulfidbindungen. Die berechneten Mittelwerte liegen sehr nah an den Ergebnissen von Morris *et al.* [100], weisen jedoch leicht erhöhte Standardabweichungen auf. Insgesamt konnten in rund 15% aller untersuchten Proteinstrukturen Disulfidbindungen gefunden werden. Das ist ein geringerer Anteil als er von Morris *et al.* [100] nachgewiesen wurde.

³ Die Werte wurden Abb. A. 3 entnommen und so gewählt, dass mindestens 95% der untersuchten Bindungen dieses Kriterium erfüllen.

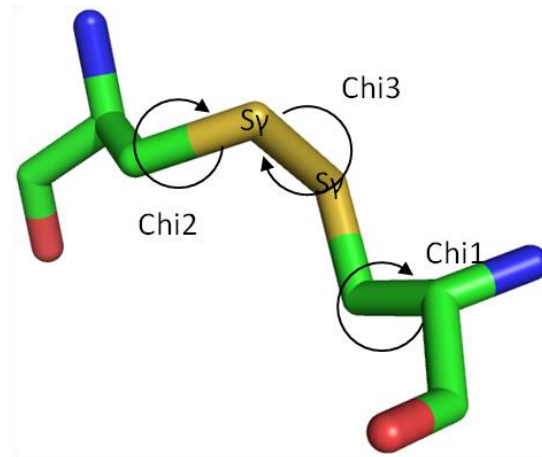


Abb. 6: Torsionswinkel einer Disulfid-Bindung

Die Abbildung zeigt die *Stick*-Repräsentation einer Disulfid-Bindung zwischen zwei Cysteinen. Die Pfeile geben exemplarisch an, für welche Bindung die entsprechenden Seitenketten-Torsionswinkel Chi1, Chi2 und Chi3 berechnet werden. Für eine Disulfid-Bindung ist dabei vor allem letzterer von entscheidender Bedeutung.

2.2.2 Salzbrücken

Salzbrücken können zwischen den negativ geladenen Aminosäuren Aspartat und Glutamat sowie den positiv geladenen Aminosäuren Arginin, Histidin und Lysin gebildet werden. Barlow und Thornton konnten bei Untersuchungen von Kristallstrukturen zwei Abstandsmaxima in den Bereichen 4,0 Å und 6,0 Å für Salzbrücken feststellen [117]. Bei genauerer Betrachtung ihrer Daten konnten sie zeigen, dass wenn sich zwei gegensätzlich geladene Atome auf 4,0 Å annähern, z.B. NH1 von Arg und OD1 von Asp, die restlichen N-O-Abstände, z.B. NH2 von Arg und OD2 von Asp, in einem Bereich von 6,0 Å liegen [117]. Dementsprechend definierten sie einen Ideal-Abstand von maximal 4,0 Å für Salzbrücken [117]. Des Weiteren konnten sie nachweisen, dass Salzbrücken eher an der Ausbildung von Tertiärstrukturen beteiligt sind als an der von Sekundärstrukturen, da mehr 66% der von ihnen gefundenen Salzbrücken eine Sequenzseparation von mehr als sieben Aminosäuren aufwiesen [117]. Weitere statistische Analysen von Salzbrücken wurden zum einen von Lee *et al.* [118] und zum anderen von Sarakatsannis und Duan [119] durchgeführt. Während erstere ebenfalls mit einem simplen Atomabstand zur Definition von Salzbrücken arbeiteten [118], nutzten letztere den Abstand zwischen den Ladungsschwerpunkten definiert durch alle Atome der Carboxyl- (Asp und Glu) oder Guanidinium-Gruppe (Arg) bzw. dem N ζ des Lysins oder allen Ringatomen des Histidins [119].

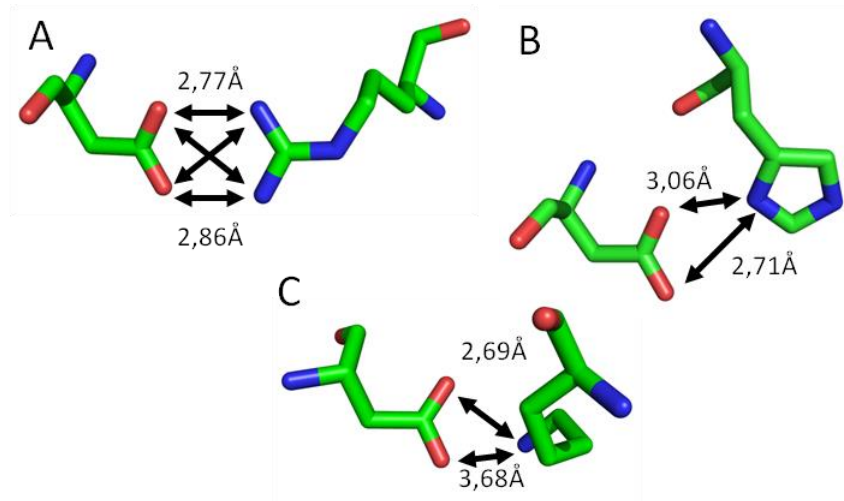


Abb. 7: Mögliche Salzbrücken des Aspartats

Die Abbildung zeigt *Stick*-Repräsentationen der Salzbrücken zwischen Aspartat und Arginin (A), Histidin (B) sowie Lysin (C). Etwaige Interaktionen sind als schwarze Pfeile gezeigt und die dazugehörigen Abstände sind darüber bzw. darunter angegeben. In (A) sind alle möglichen Interaktionen eingezeichnet und die gezeigten Längen beziehen sich auf die oberste bzw. unterste Interaktion.

Unter Nutzung des Ideal-Abstandes von 4,0 Å konnten sämtliche Salzbrücken im Datensatz XRD bestimmt werden. Dabei wurden etwaige multiple Salzbrücken zwischen zwei Aminosäuren (z.B. die vier möglichen Salzbrücken in Abb. 7 (A)) für die statistischen Auswertungen nur einmal gewertet. Insgesamt machen Interaktionen zwischen Aspartat/Glutamat und Arginin bzw. Lysin ungefähr 80% bis 90% aller Salzbrücken aus (Tabelle 11). Des Weiteren ist erkennbar, dass die Interaktionen einer negativ geladenen Aminosäure mit Arginin und Lysin sich ungefähr im Verhältnis 3:2 aufteilen. Das ist besonders interessant wenn bedacht wird, dass Lysin in diesem Datensatz insgesamt deutlich häufiger vorkommt als Arginin (Tabelle A.1). Da Wigley *et al.* anhand von Protein-Ligand Interaktionen in Kristallstrukturen zeigen konnten, dass von Arginin gebildete Salzbrücken deutlich stabiler sind als Salzbrücken von Lysin [120], erscheint die Präferenz von Arginin logisch.

Tabelle 11: Salzbrücken und daran beteiligte Aminosäuren im Datensatz XRD

Im oberen Teil der Tabelle geben die Prozente in Klammern den Anteil der jeweils gefundenen Salzbrücke an der Gesamtheit aller gefundenen Salzbrücken an. Im unteren Teil geben die Prozentzahlen die Verteilung der negativ bzw. positiv geladenen Aminosäuren innerhalb der Salzbrücken an. Zur besseren Vergleichbarkeit mit der Literatur wurden die Interaktionen von Histidin nicht prozentual ausgewertet.

| Salzbrücke | Barlow [117] | Sarakatsannis [119] | Lee [118] | XRD |
|-----------------|---------------|---------------------|------------------|------------------|
| Asp-Arg | 38 (23,3%) | X | 7240 (22,6%) | 20100 (29,3%) |
| Asp-His | 29 | X | X | 6066 |
| Asp-Lys | 34 (20,9%) | X | 8040 (25,0%) | 11526 (16,8%) |
| Glu-Arg | 46 (28,2%) | X | 7337 (22,9%) | 23328 (34,1%) |
| Glu-His | 16 | X | X | 5243 |
| Glu-Lys | 45 (27,6%) | X | 9479 (29,5%) | 13543 (19,8%) |
| Aminosäure | Barlow [117] | Sarakatsannis [119] | Lee [118] | XRD |
| Aspartat | 44,2% | 5272 (50,8%) | 15280 (47,6%) | 37692 (47,2%) |
| Glutamat | 55,8% | 5098 (49,2%) | 16816 (52,4%) | 42114 (52,8%) |
| Arginin | 51,5% | 4408 (44,0%) | 14577 (45,4%) | 43428 (63,4%) |
| Histidin | | 361 | X | 11309 |
| Lysin | 48,5% | 5601 (56,0%) | 17519 (54,6%) | 25069 (26,6%) |

Der Vergleich der erhaltenen statistischen Daten mit früheren Studien zeigt, dass für Aspartat und Glutamat ähnliche prozentuale Anteile nachgewiesen wurden. Bei den positiv geladenen Aminosäuren findet sich jedoch eine deutlich veränderte Aufteilung als bisher postuliert. Im vorliegenden Datensatz werden knapp zwei Drittel aller Salzbrücken von Arginin gebildet während in älteren Studien nur knapp 45% der Salzbrücken durch Arginin gebildet werden [118, 119]. Barlow und Thornton begründeten den höheren Anteil von Arginin in Salzbrücken mit der erhöhten Stabilität von Salzbrücken aufgrund der größeren Anzahl an Ladungsträgern im Arginin im Vergleich zum Lysin [117]. In ähnlicher Weise diskutierten sie, dass Glutamat im Vergleich zu Aspartat eher Salzbrücken ausbilden kann, da es aufgrund seiner längeren Seitenkette besser in der Lage ist, die nötigen Konformationen einzunehmen [117].

Tabelle 12: Prozentuale Verteilung der Sequenzseparationen von Salzbrücken

Die Daten von Sarakatsannis [119] wurden ausgehend von gerundeten Werten aus Abb. 8 berechnet.

| Separation | Sarakatsannis [119] | XRD |
|-------------------------|---------------------|-----|
| 1 Aminosäure | ~3% | 3% |
| 2 Aminosäuren | ~7% | 8% |
| 3 Aminosäuren | ~10% | 9% |
| 4 Aminosäuren | ~11% | 9% |
| 5 Aminosäuren | ~2% | 2% |
| 6 Aminosäuren | ~2% | 2% |
| 7 oder mehr Aminosäuren | ~65% | 67% |

Betrachtet man die Sequenzseparationen von Salzbrücken in einem relativ engem Bereich von maximal 10 bis 20 Aminosäuren (Abb. 8) so erkennt man, dass viele der gefundenen Salzbrücken maximal vier Aminosäuren voneinander entfernt liegen. Bei genauerer Betrachtung der Werte (Tabelle 12) wird jedoch deutlich, dass die beteiligten Aminosäure-Rest in nur knapp 35% aller gefundenen Salzbrücken nah beieinander liegen. Der hier gefundene Anteil von 67% an Salzbrücken mit einer Sequenzseparation von sieben oder mehr Aminosäuren steht im direkten Einklang mit den Daten von Barlow und Thornton [117]. Auch in der Studie von Sarakatsannis und Duan finden sich rund 65% der Salzbrücken mit einer Sequenzseparation von sieben mehr Aminosäuren, trotzdem argumentierten sie, dass Salzbrücken vor allem bei der internen Stabilisierung von Helices eine wichtige Rolle spielen [119].

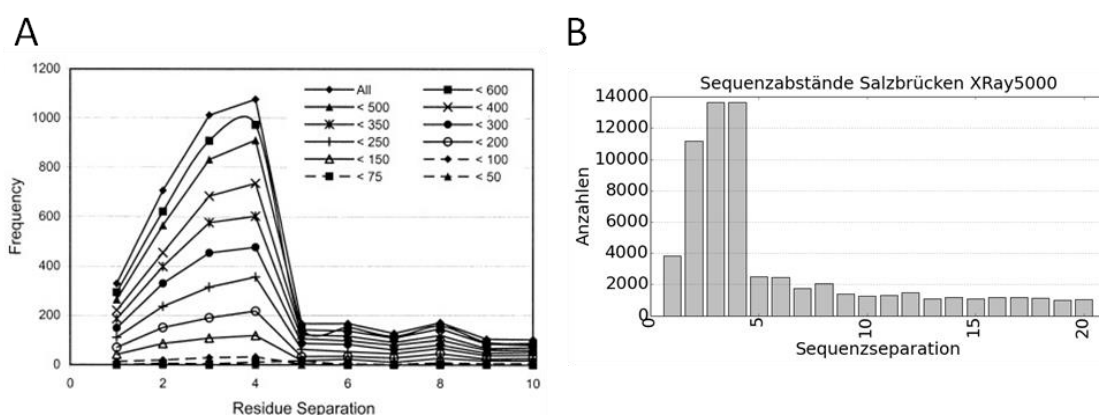


Abb. 8: Sequenzseparationen von Salzbrücken in XRD

Die Diagramme zeigen die gefundenen Sequenzseparationen der Studien von Sarakatsannis und Duan [119] sowie der vorliegenden Arbeit. In beiden Fällen ist ein Maximum im Bereich von drei bis vier Aminosäuren zu erkennen. Bei Separationen von mehr als vier Aminosäuren kommt es zu einem starken Abfall der Anzahl gefundenen Salzbrücken.

2.2.3 Kation- π -Wechselwirkungen

Identifizierung von Kation- π -Interaktionen

Bei Kation- π -Wechselwirkungen interagiert ein geladenes Stickstoff-Atom, z.B. $N\zeta$ von Lysin, mit einem aromatischen π -Elektronensystem [121]. Burley und Petsko konnten anhand von Polarkoordinaten bestimmen, dass der Abstand r zwischen dem Masseschwerpunkt des aromatischen Ringes und dem polarisierten Stickstoff-Atom in einem Bereich von 3,4 Å bis 6,0 Å liegen muss [121]. Sie definierten dazu den Masseschwerpunkt des aromatischen Rings als Koordinatenursprung und legten die x -Achse parallel zur $C\beta$ - $C\gamma$ -Bindung des Aromaten sowie die z -Achse parallel zur 6-fachen Symmetrie-Achse des aromatischen Rings [121]. Um die Orientierung des polarisierten Stickstoffs zur Ringebene zu bestimmen, definierten sie den θ -Winkel zwischen der z -Achse und dem Positionsvektor des Stickstoffatoms innerhalb des Polarkoordinatensystems (Abb. 9 A) [121].

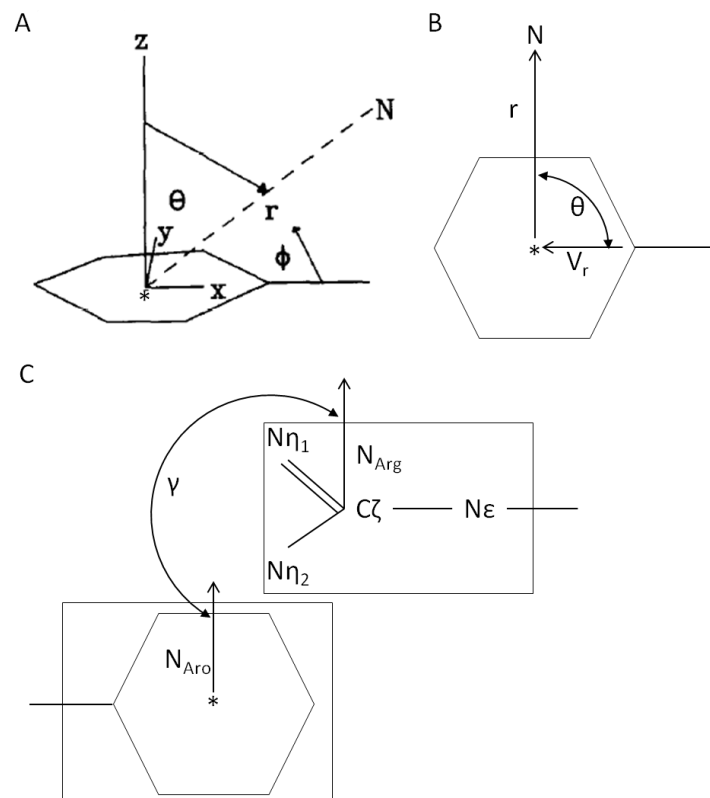


Abb. 9: Schematische Darstellung von Kation- π -Wechselwirkungen

Die Darstellung in (A) wurde der Arbeit von Burley und Petsko [121] entnommen und stellt die von ihnen genutzten Polarkoordinaten dar. In (B) ist das in der vorliegenden Arbeit genutzte System zur Bestimmung von Kation- π -Wechselwirkungen gezeigt. Das Schema in (C) stellt dar, wie die Orientierung der Ebenen (Rechtecke) einer Guanidinium-Gruppe und eines aromatischen Rings in der vorliegenden Arbeit bestimmt wurde. Für mehr Informationen siehe Text.

Strukturbiologische Parameter

In der vorliegenden Arbeit wurde ein verändertes System genutzt. Dabei wurde der Abstand r zwischen dem Masseschwerpunkt des aromatischen Rings und dem N ζ -Atom des Lysins oder dem Masseschwerpunkt der Guanidinium-Gruppe des Arginins bestimmt. Histidin wurde ausschließlich als Aromat behandelt, da die Bestimmung des Protonierungsgrades mit den hier genutzten Strukturen nicht möglich war. Die Masseschwerpunkte wurden dabei als arithmetisches Mittel aller Atome der Ebene berechnet (Tabelle 13, Spalte 2). Lag der berechnete Abstand innerhalb des von Burley und Petsko definierten Rahmens von 3,4 Å bis 6,0 Å [121], konnte von einer Kation- π -Wechselwirkung ausgegangen werden.

Tabelle 13: Zur Bestimmung von Masseschwerpunkten genutzte Atome

Für Tryptophan kamen drei verschiedene Systeme zum Einsatz, die entweder den 5-Ring, den 6-Ring oder das komplette Ringsystem benutzten. Die Atombezeichnungen wurden der PDB entnommen. Die letzte Spalte gibt die Vektoren, an die zur Bestimmung der Normalen der Ebenen genutzt wurden.

| Aminosäure | Atome der Ebene | Vektoren |
|--------------|--|---------------------|
| Arginin | NE, CZ, NH1, NH2 | NE-NH1 und NE-NH2 |
| Histidin | CG, ND1, CD2, CE1, NE2 | CG-CE1 und CG-NE2 |
| Phenylalanin | CG, CD1, CD2, CE1, CE2, CZ | CG-CE1 und CG-CZ |
| Tyrosin | CG, CD1, CD2, CE1, CE2, CZ | CG-CE1 und CG-CZ |
| Tryptophan5 | CG, CD1, CD2, NE1, CE2 | CG-NE1 und CG-CE2 |
| Tryptophan6 | CD2, CE2, CE3, CZ2, CZ3, CH2 | CD2-CZ3 und CD2-CH2 |
| Tryptophan9 | CG, CD1, CD2, NE1, CE2, CE3, CZ2, CZ3, CH2 | CG-CZ3 und CG-CH2 |

Um die Orientierung des Kations zur Ringebene zu bestimmen, wurde analog zu Burley und Petsko [121] ein θ -Winkel definiert. Der θ -Winkel repräsentierte den Winkel zwischen dem Abstandsvektor r und dem Ring-Vektor V_r zwischen Cy-Atom und Masseschwerpunkt des aromatischen Rings (Abb. 9 B). Bei einem Vergleich der θ -Winkel von Burley und Petsko [121] sowie der vorliegenden Arbeit galten dann die in Tabelle 14 gezeigten Verhältnisse.

Tabelle 14: Verhältnisse der verschiedenen θ -Winkel von Burley und Petsko und der aktuellen Arbeit

| Burley und Petsko [121] | Aktuelle Arbeit | Orientierung |
|-------------------------|-----------------|--|
| 0° | 90° | Kation senkrecht über Masseschwerpunkt |
| 45° | 45° oder 135° | |
| 90° | 0° 180° | Kation orientiert sich in Richtung Ende der Seitenkette Kation orientiert sich in Richtung Cy |

Sowohl Flocco und Mowbray [122] als auch Gallivan und Dougherty [123] konnten zwei verschiedene Konformationen für Interaktionen zwischen Arginin und aromatischen Ringen

feststellen (Abb. 10). Um die Orientierung der Ebenen der Guanidinium-Gruppe und des aromatischen Ringes zueinander zu bestimmen, wurde ein zweiter Winkel definiert (Abb. 9 C). Dieser γ -Winkel (Ebenenwinkel) berechnete sich dabei als Winkel zwischen den Normalen der beiden Ebenen (N_{Aro} und N_{Arg}). Die Normalen konnten als Kreuzprodukte von zwei Vektoren innerhalb der jeweiligen Ebene (Tabelle 13, Spalte 3) bestimmt werden.

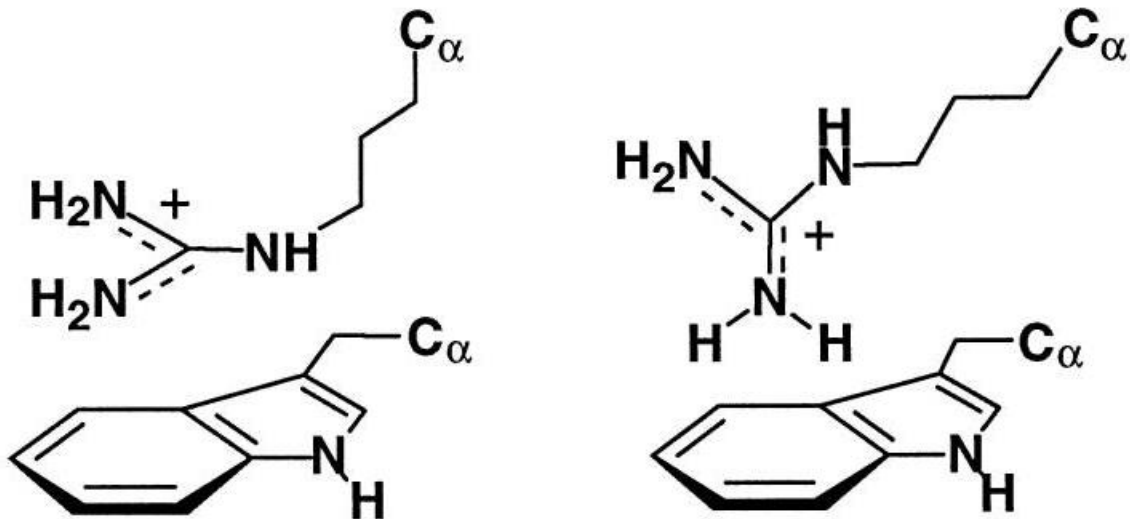


Abb. 10: Verschiedene Orientierungen von Arginin im Bezug zu Aromaten [123]

Die Abbildung wurde der Studie von Gallivan und Dougherty [123] entnommen und stellt die verschiedenen Orientierungen der Guanidinium-Gruppe eines Arginins im Bezug zum aromatischen System dar. Auf der linken Seite ist die planare Orientierung dargestellt und auf der rechten Seite die *T-stacked* Orientierung.

Bisherige Untersuchungen

Im Bezug auf Kation- π -Wechselwirkungen gibt es viele analytische und statistische Studien, von denen einige hier kurz vorgestellt werden sollen. Eine der grundlegendsten Studien wurde von Burley und Petsko [121] angefertigt. Mit Hilfe eines kleinen Datensatzes von 33 Kristallstrukturen und einem Polarkoordinatensystem (Abb. 9 A) konnten sie ein Abstandskriterium von 3,4 Å bis 6,0 Å für Kation- π -Interaktionen definieren [121]. Zusätzlich konnten sie mit Hilfe ihres θ -Winkels (Maximum: 0° - 45°) zeigen, dass ein Großteil der polarisierten Stickstoff-Atome senkrecht über dem Masseschwerpunkt des aromatischen Ringsystems stehen (siehe auch Abb. 11 oder Abb.2 in [121]).

Flocco und Mowbray untersuchten die Interaktionen von Arginin mit aromatischen Aminosäuren anhand von 79 Kristallstrukturen [122]. Sie definierten dabei ein einfaches Abstandskriterium unter Nutzung aller Atome des aromatischen Ringes und aller Atome der

Strukturbiologische Parameter

Arginin-Seitenkette beginnend mit dem C γ -Atom [122]. Sobald mindestens einer der verschiedenen Atomabstände kleiner als 4,6 Å war, nahmen sie eine Kation- π -Wechselwirkung an [122]. Sie konnten so zeigen, dass knapp 33% aller Arginine ihres Datensatzes an Kation- π -Interaktionen beteiligt sind [122]. Zusätzlich konnten sie unter Verwendung eines Ebenenwinkels α ($< 25^\circ$) und eines *Cutoff*-Abstandes von 4,5 Å nachweisen, dass die parallele, planare Orientierung der Guanidinium-Gruppe im Bezug auf die aromatischen Ringe relativ häufig für geringe Abstände vorkommt [122]. Bei Betrachtung der Interaktionspartner des Arginins fanden sie Anteile von 39% für Phenylalanin, 39% für Tyrosin und 22% für Tryptophan [122].

Gallivan und Dougherty [123] sowie Crowley und Golovin [124] verfolgten einen energetischen Ansatz bei der Bestimmung von Kation- π -Wechselwirkungen. In einem ersten Schritte maßen sie den Abstand zwischen dem Zentroid des aromatischen Ringes und dem C ζ -Atoms eines Arginins oder dem N ζ -Atom eines Lysins [123]. Lag dieser Abstand unterhalb des Distanz-*Cutoffs* von 6,0 Å, testeten sie, ob ein Wasser-Molekül in Form einer Kugel mit einem Durchmesser von 2,8Å zwischen den Seitenketten lag [123]. Mit Hilfe von *ab initio*-Berechnungen konnten sie dann Energien für elektrostatische und van-der-Waals-Wechselwirkungen zwischen den beiden Seitenketten zu berechnen [123]. Mit Hilfe dieser Energien konnten sie zwei Grenzwerte für Kation- π -Wechselwirkungen festlegen [123]. Sie akzeptierten alle Interaktionen, deren elektrostatische Bindungsenergie $\leq -2,0$ kcal/mol war sowie Interaktionen, deren elektrostatische Energie zwischen $-2,0$ kcal/mol und $-1,0$ kcal/mol lag, wenn die van-der-Waals-Interaktionsenergie kleiner/gleich $-1,0$ kcal/mol war [123].

Gallivan und Dougherty konnten mit Hilfe des OPLS-Kraftfeldes zeigen, dass in der Gas-Phase Kation- π -Wechselwirkungen von Lysin stärker sind als Interaktionen von Arginin [123]. Des Weiteren konnten sie für die Gas-Phase nachweisen, dass *T-stacked* Interaktionen von Arginin mit aromatischen Aminosäuren energetisch günstiger sind als parallele Orientierungen [123]. Für Interaktionen innerhalb von Proteinstrukturen konnten sie jedoch zeigen, dass die Energie von Kation- π -Wechselwirkungen nahezu unabhängig vom Kation bzw. der Orientierung der Guanidinium-Gruppe ist [123]. Crowley und Golovin konnten mit Hilfe dieses Systems zusätzlich zeigen, dass energetisch günstige *T-stacked* Orientierungen der Guanidinium-Gruppe höhere Distanzen r zwischen dem Masseschwerpunkt des

aromatischen Systems und dem C ζ -Atom des Arginins erfordern als parallele Orientierungen [124].

Statistische Ergebnisse

Ein Vergleich der erhaltenen Kation- π -Wechselwirkungen mit älteren Studien (Tabelle 15) lieferte sehr unterschiedliche Ergebnisse, die vor allem durch unterschiedliche Bestimmungen der Kation- π -Interaktionen zustande kommen. Rund 30% aller in Proteinen vorkommenden Arginin-Reste und 20% der Lysin-Reste sind an Kation- π -Wechselwirkungen beteiligt. Diese Ergebnisse stehen in klarer Diskrepanz zu den Ergebnissen der Untersuchungen von Burley und Petsko [121] sowie Gallivan und Dougherty [123]. Diese Unterschiede resultieren zum einen aus dem sehr geringen Datensatz, der Burley und Petsko zur Verfügung stand [121], zum anderen sorgt die Beschränkung auf ein rein geometrisches Kriterium in der vorliegenden Arbeit dafür, dass etwaige energetisch ungünstige Interaktionen mitgezählt werden. Im Gegensatz dazu betrachten Gallivan und Dougherty nur Wechselwirkungen, deren Bindungsenergie deutlich negativ ist [123]. Interessanterweise stimmt der hier gefundene Anteil von 30,5% für Arginin sehr gut mit den Daten von Flocco und Mowbray überein, die gezeigt haben, dass rund 33% aller Arginin-Reste an Kation- π -Wechselwirkungen beteiligt sind [122].

Tabelle 15: An Kation- π -Wechselwirkungen beteiligte Aminosäuren

Die Prozente in Klammern geben an wie viele der im Datensatz vorhanden Aminosäure-Reste (Tabelle A.1) jeweils an den Interaktionen beteiligt sind.

| Aminosäure | Burley [121] | Gallivan [123] | XRD |
|--------------|---------------|-----------------|------------------|
| Arginin | 44 (46,8%) | 1988 (18,2%) | 22430 (30,5%) |
| Lysin | 45 (26,5%) | 1006 (7,5%) | 16072 (18,7%) |
| Histidin | | | 7275 |
| Phenylalanin | | 915 (10,0%) | 12209 (20,4%) |
| Tryptophan | | 892 (26,1%) | 6065 (27,0%) |
| Tyrosin | | 1187 (14,3%) | 15319 (28,2%) |

Die in Tabelle 16 zusammengefassten möglichen Kombinationen von Kation- π -Wechselwirkungen weisen eine klare Präferenz von Kation- π -Wechselwirkungen zwischen

Arginin und Phenylalanin/Tyrosin (rund 50% aller gefundenen Interaktionen) auf. Insgesamt konnten die Ergebnisse von Gallivan und Dougherty [123] für alle Kombinationen nahezu bestätigt werden. Auffällig ist, dass der prozentuale Anteil an Interaktionen, an denen Tryptophan beteiligt ist, in dieser Untersuchung geringer ist als in den Literaturangaben.

Tabelle 16: Statistische Auswertung aller möglichen Kation- π -Interaktionen

Die Prozentzahlen geben den Anteil der jeweiligen Kombination an allen gefundenen Kation- π -Wechselwirkungen an. Zur besseren Vergleichbarkeit der Ergebnisse wurden die Interaktionen mit Histidin bei der Berechnung der Prozente des Datensatzes XRD nicht beachtet.

| Kombination | Gallivan [123] | XRD |
|-------------|----------------|-----------------|
| Arg-His | | 4705 |
| Arg-Phe | 630 (21,0%) | 8712 (23,1%) |
| Arg-Trp | 609 (20,3%) | 4108 (11,1%) |
| Arg-Tyr | 749 (25,0%) | 9281 (25,2%) |
| Lys-His | | 3220 |
| Lys-Phe | 285 (9,5%) | 4535 (12,3%) |
| Lys-Trp | 283 (9,5%) | 2542 (6,9%) |
| Lys-Tyr | 438 (14,6%) | 7728 (20,9%) |

Mit Hilfe des θ -Winkels kann die Orientierung des Kations im Verhältnis zum aromatischen Ring bestimmt werden. Analog zu den Ergebnissen von Burley und Petsko [121] konnte eine breite Verteilung der θ -Winkel nachgewiesen werden (Abb. 11). Burley und Petsko beschrieben ein Maximum von rund 45° für die Interaktionen von Arginin und Lysin mit aromatischen Aminosäuren [121]. Eine genauere Betrachtung der θ -Winkel der vorliegenden Arbeit konnte zeigen, dass knapp 80% der Werte in einem Bereich von 45° bis 135° liegen und somit ein ähnliches Maximum gefunden wurde.

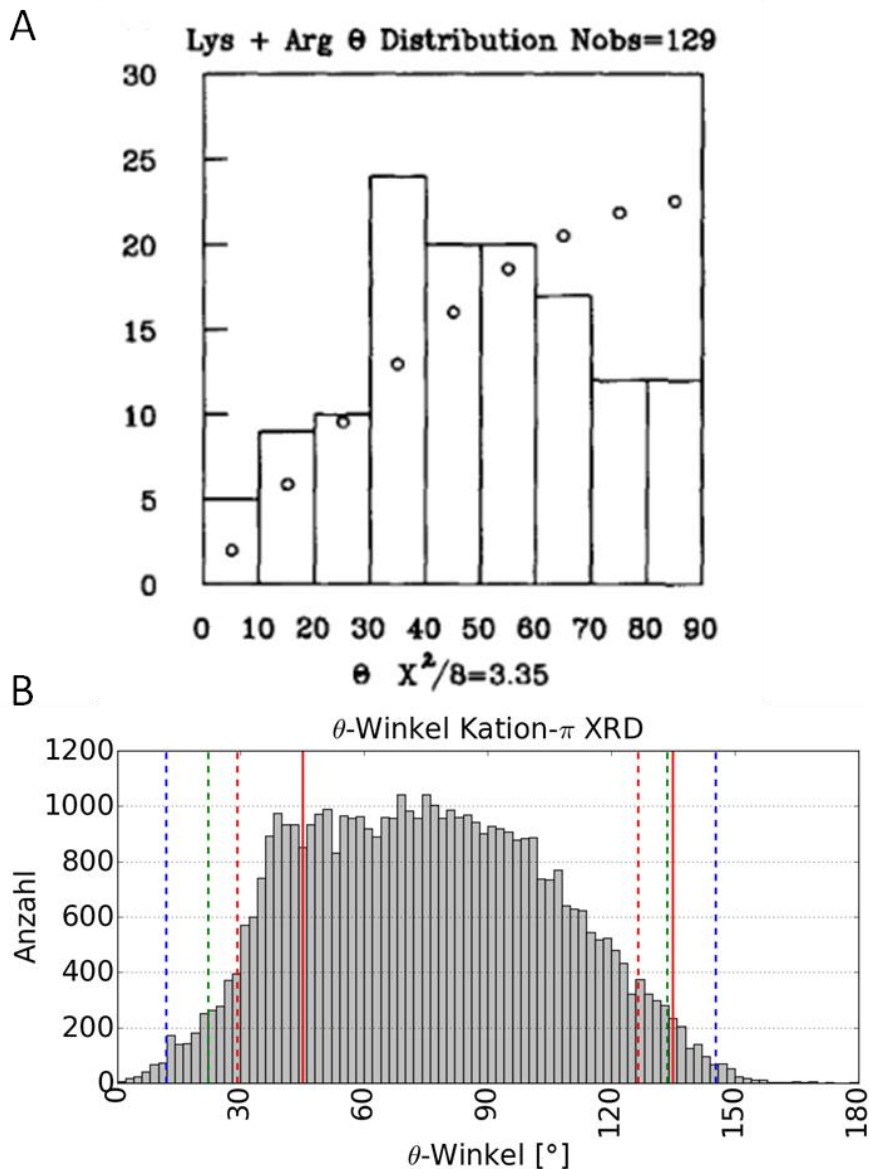


Abb. 11: Verteilung der Theta-Winkel

Abbildung (A) wurde der Studie von Burley und Petsko [121] entnommen und zeigt die berechneten Theta-Winkel von Lysin und Arginin bei Interaktionen mit aromatischen Aminosäuren. Die Punkte entsprechen $\sin(\theta)$ und repräsentieren die zufällige Verteilung von diesen Interaktionen. Das Histogramm in (B) zeigt die in der aktuellen Arbeit berechneten θ -Winkel. Die durchgezogenen roten Linien entsprechen dem ungefähren Maximum von 45° , den Burley und Petsko für ihre Winkel finden konnten [121]. Die gestrichelten Linien entsprechen den Konfidenz-Intervallen von 90% (rot), 95% (grün) und 99% (blau) der gefundenen Werte (Tabelle 17).

Erstaunlicherweise zeigen die hier berechneten θ -Winkel keine Normalverteilung mit einem Maximum im Bereich 90° sondern eine sehr breite Verteilung im Bereich von $\sim 40^\circ$ bis $\sim 110^\circ$ (Abb. 11 B). Eine Berechnung der Konfidenz-Intervalle (gestrichelte Linien in Abb. 11 und

Tabelle 17) konnte nachweisen, dass 90% der berechneten Winkel in einem Bereich von $\sim 30^\circ$ bis $\sim 125^\circ$ liegen. Das 99% Konfidenz-Intervall liegt in einem Bereich von $\sim 12^\circ$ bis $\sim 145^\circ$ und bestätigt, dass die Daten nicht normalverteilt um einen Mittelwert von 90° sind. Dies entspricht der Orientierung des Kations von der $C\beta-C\gamma$ -Bindung des aromatischen Systems weg zum Ende der Seitenkette hin. Insbesondere bei Interaktionen an denen Tyrosin beteiligt ist (Abb. A. 4) zeigt sich dieses Verhalten. Dougherty konnte anhand von *ab initio* 6-31G** Berechnungen zeigen, dass Tyrosin ein zusätzliches negatives Potential, im Vergleich zum π -Elektronensystem, an der Oberfläche im Bereich der Hydroxylgruppe trägt (Abb. A. 5) [125]. Dementsprechend resultiert die präferierte Orientierung der Kationen bei Wechselwirkungen mit Tyrosin in Richtung der Hydroxyl-Gruppe aus diesem negativen Potential.

Tabelle 17: Verschiedene Konfidenz-Intervalle der θ -Winkel des Datensatzes XRD

| Intervall | Untergrenze | Obergrenze | % Winkel |
|-------------------------|--------------|---------------|----------|
| Burley und Petsko [121] | 45° | 135° | 78,9% |
| 90% Konfidenz | $29,2^\circ$ | $126,6^\circ$ | 90% |
| 95% Konfidenz | $22,0^\circ$ | $133,6^\circ$ | 95% |
| 99% Konfidenz | $11,9^\circ$ | $145,4^\circ$ | 99% |

Sonderfall Arginin

Wie weiter oben beschrieben kann Arginin zwei Konformationen bei der Interaktion mit aromatischen Systemen eingehen (Abb. 10). Aus diesem Grund wurden die Ebenenwinkel γ für alle Interaktionen von Arginin mit aromatischen Aminosäuren berechnet (Abb. 12). Die Daten zeigten unabhängig vom Interaktionspartner eine relative Gleichverteilung über den gesamten Wertebereich. Für Interaktionen von Arginin und Histidin konnte eine leichte Präferenz einer *T-stacked* Orientierung ($\gamma=90^\circ$) und für Interaktionen von Arginin mit Tryptophan eine Präferenz einer planaren Orientierung ($\gamma= \pm 180^\circ$) festgestellt werden.

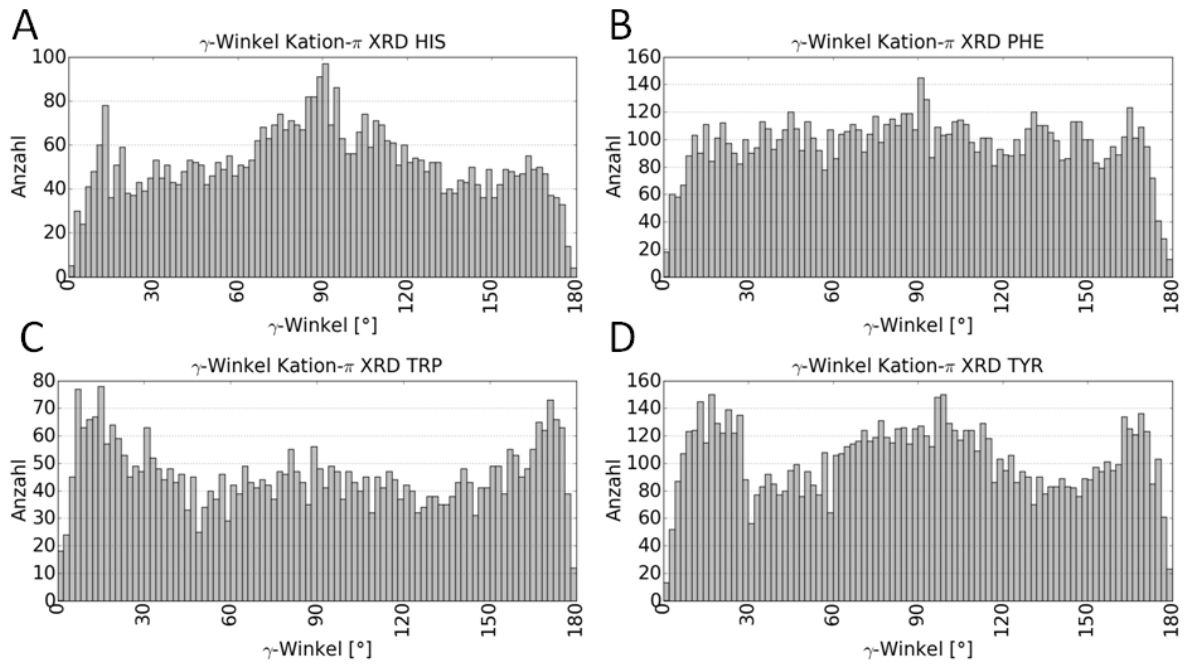


Abb. 12: Verteilung der γ -Winkel bei Kation- π -Wechselwirkungen von Arginin mit verschiedenen aromatischen Aminosäuren

Die vier Histogramme zeigen die Verteilung der berechneten γ -Winkel (Ebenenwinkel) bei Kation- π -Interaktionen von Arginin mit Histidin (A), Phenylalanin (B), Tryptophan (C) und Tyrosin (D). In allen Fällen sind die Winkel relativ gleich verteilt über den gesamten Wertebereich mit leicht unterschiedlichen Maxima.

Flocco und Mowbray konnten zeigen, dass bei kleineren Abständen zwischen der Guanidinium-Gruppe und dem aromatischen Ringsystem planare Orientierungen häufiger vorkommen [122]. Analog dazu konnten Gallivan und Dougherty mit ihren *ab initio* Berechnungen zeigen, dass *T-stacked* Interaktionen zwar stärker sind, eine planare Orientierung jedoch favorisiert wird [123]. Die Ergebnisse der vorliegenden Arbeit (Tabelle 18) zeigen keine Präferenz zugunsten von *T-stacked* Konformationen bei den Wechselwirkungen von Arginin mit aromatischen Systemen. Dies entspricht nicht den theoretischen Ergebnissen von Gallivan und Dougherty bei ihren *ab initio* Berechnungen in der Gas-Phase [123]. Die Diskrepanz der Daten resultiert auch hier aus der Beschränkung auf ein rein geometrisches Kriterium, wodurch auch energetisch ungünstige Interaktionen in die Statistik mit einfließen.

Tabelle 18: Prozentuale Verteilungen der Orientierungen der Guanidinium-Gruppe zum aromatischem Ringsystem

Angegeben sind die prozentualen Anteile der jeweiligen Anordnung im Bezug auf alle gefundenen Interaktionen von Arginin mit aromatischen Ringsystemen. In der ersten Zeile der jeweiligen Studie sind alle Ergebnisse zusammengefasst und danach sind die Aminosäure-spezifischen Anteile angegeben.

| Arbeit | Typ | Planar ($\gamma \geq \pm 150^\circ$) | Schräg | <i>T-Stacked</i> ($60^\circ \leq \gamma \leq 120^\circ$) |
|-----------------|--------------|--|--------|--|
| Crowley [124] | Gesamt | 51% | 28% | 21% |
| | Tyrosin | 22% | 12% | 9% |
| | Phenylalanin | 17% | 11% | 8% |
| | Tryptophan5 | 7% | 3% | 2% |
| | Tryptophan6 | 5% | 2% | 2% |
| Aktuelle Arbeit | Gesamt | 31,5% | 30,7% | 37,8% |
| | Histidin | 4,6% | 5,3% | 7,6% |
| | Phenylalanin | 9,3% | 11,3% | 11,9% |
| | Tryptophan | 5,9% | 4,5% | 4,9% |
| | Tyrosin | 11,7% | 9,5% | 13,5% |

Crowley und Golovin konnten bei den von ihnen gefundenen Interaktionen von Arginin mit aromatischen Ringsystemen zeigen, dass der Abstand der Masseschwerpunkte für *T-Stacked* Konformationen größer sein muss als für planare Konformationen [123]. Die steigende Präferenz von *T-stacked* Anordnungen bei steigendem Abstand, konnte durch eine Analyse der gefundenen Konformationen in Abhängigkeit vom Abstand der Masseschwerpunkte der Guanidinium-Gruppe und des aromatischen Ringes (Tabelle 19), nachgewiesen werden. Dieses Verhalten entspricht den Ergebnissen von Flocco und Mowbray die bei ihren statistischen Untersuchungen ein ähnliches Verhalten nachweisen konnten [122]. Sie begründeten dieses Verhalten mit der Ausbildung von Wasserstoffbrücken zwischen den $N\eta_1$ - und $N\eta_2$ -Atomen des Arginins und dem aromatischen Ring [122]. Des Weiteren diskutierten sie, dass die planare Orientierung eine Ausrichtung von Arginin-Resten in katalytisch aktiven Bereichen von Proteinen ermöglicht, ohne dabei das Wasserstoffbrückennetzwerk rund um das Arginin zu beeinflussen [122].

Tabelle 19: Abstandsabhängigkeit der Konformation der Arginin-Interaktion mit aromatischen Systemen

Die Prozentzahlen in Klammern geben den Anteil der jeweiligen Konformation im dazugehörigen Abstands-Cluster an. Der angegebene Abstand entspricht dem Abstand zwischen den beiden Masseschwerpunkten.

| Abstand | Planar | Schräg | <i>T-Stacked</i> |
|-----------------|-----------------|-----------------|------------------|
| 3,0 Å bis 3,5 Å | 111 (95,7%) | 0 (0,0%) | 5 (4,3%) |
| 3,5 Å bis 4,0 Å | 2562 (90,4%) | 195 (6,9%) | 76 (2,7%) |
| 4,0 Å bis 4,5 Å | 2367 (57,1%) | 1428 (34,4%) | 353 (8,5%) |
| 4,5 Å bis 5,0 Å | 1469 (25,2%) | 2281 (39,1%) | 2078 (35,7%) |
| 5,0 Å bis 5,5 Å | 1069 (16,1%) | 2072 (31,2%) | 3490 (52,7%) |
| 5,5 Å bis 6,0 Å | 847 (11,8%) | 2227 (31,0%) | 4100 (57,2%) |

2.2.4 π - π -Wechselwirkungen

Grundlegende Definition

Statistische Studien von Burley und Petsko (34 Strukturen) [85] sowie McGaughey *et al.* (505 Strukturen) [126] wiesen nach, dass für π - π -Interaktionen der Maximalabstand der Zentroide zweier aromatischer Ringe bei 7,5 Å liegen muss (Abb. 13). Eine Bestimmung aller Interaktionen innerhalb eines Abstandsereiches von 3,0 Å bis 12,0 Å für den Datensatz XRD (Abb. 13 C) stimmte mit den Ergebnissen bisheriger Studien (Abb. 13 A und B) sehr gut überein. Die Verteilung enthält ein (lokales) Maximum im Bereich von 5,0 Å bis 6,0 Å (Idealabstand der Interaktionen) sowie ein Minimum bei rund 7,5 Å. Es handelte sich demnach um eine radiale Verteilungsfunktion für die McGaughey *et al.* feststellten, dass bei Abständen größer 7,5 Å keine direkten Ring-Ring-Kontakte mehr vorhanden sind [126]. Aus diesem Grund werden im Folgenden nur noch Interaktionen betrachtet bei denen die Zentroide der Aminosäuren sich in einer räumlichen Nähe von 3,0 Å bis 7,5 Å befinden.

Strukturbiologische Parameter

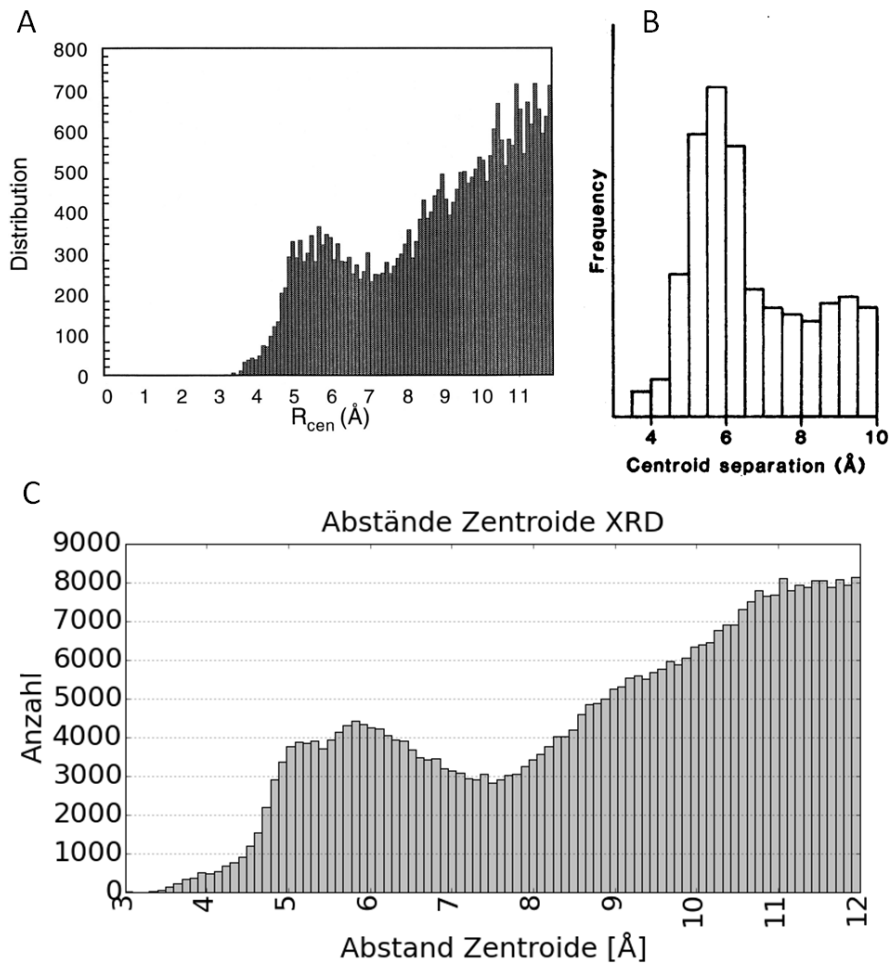


Abb. 13: Vergleich der gefundenen π - π -Wechselwirkungen in Abhängigkeit vom Abstand der Masseschwerpunkte

In der Abbildung sind die Ergebnisse der Studien von McGaughey *et al.* [126] (A), Burley und Petsko [85] (B) und der vorliegenden Arbeit (C) zu sehen. Alle drei Untersuchungen zeigen ein (lokales) Maximum im Bereich von 5,0 Å bis 6,0 Å, sowie einen Anstieg der gefundenen Interaktionen bei Abständen über 7,5 Å.

Konformationen von π - π -Wechselwirkungen

Jorgensen und Severance konnten bei der energetischen Berechnung von Interaktionen von Benzen-Dimeren in der Gas-Phase drei energetisch günstige Konformationen definieren [127]. Mit Hilfe eines OPLS-Kraftfeldes, der Nutzung eines kombinierten Coulomb/Lennard-Jones-Potentials und Monte-Carlo-Simulationen in der Gas-Phase konnten sie die Interaktionsenergien verschiedener Konformationen berechnen [127]. Die drei energetisch günstigsten Konformationen waren dabei *T-stacked*- (-2,3 kcal/mol), *parallel displaced* (-2,1 kcal/mol) und um 30° gedrehte *sandwich*-Konformation (-1,7 kcal/mol) (Abb. 14 a bis C) [127].

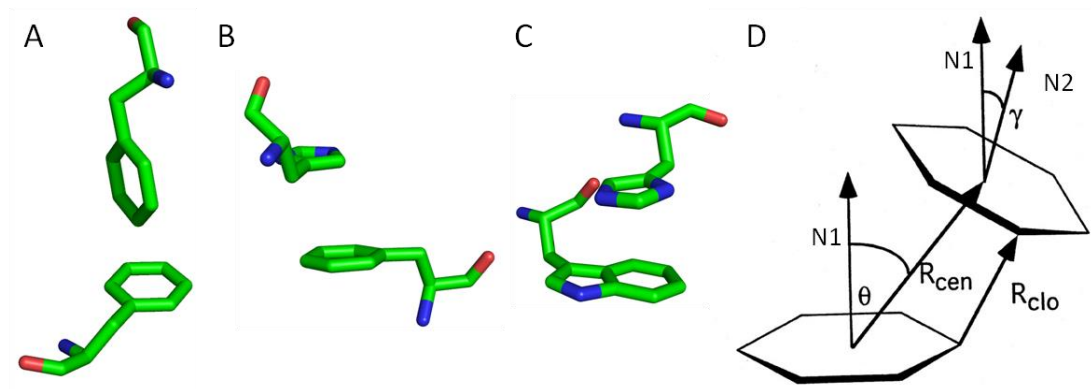


Abb. 14: Mögliche Konformationen von π - π -Wechselwirkungen

Das Bild zeigt zum einen *Stick*-Repräsentationen der drei möglichen Konformationen *T-stacked* (A), *parallel displaced* (B) und *sandwich* (C) von π - π -Wechselwirkungen innerhalb von Proteinstrukturen. Die Abbildung (D) wurde der Arbeit von McGaughey *et al.* [126] entnommen und angepasst. In ihr ist schematisch dargestellt, wie die Orientierung der aromatischen Ringe zueinander bestimmt werden kann.

Die energetisch günstigsten Konformationen bei π - π -Wechselwirkungen können unter Verwendung von verschiedenen Winkeln identifiziert werden (Abb. 14 D und Tabelle 20). Das in der vorliegenden Arbeit genutzte System basiert dabei auf den Festlegungen die McGaughey *et al.* für ihre Analyse von Kristallstrukturen getroffen haben [126]. Am wichtigsten ist dabei der Winkel γ zwischen den Normalen der beiden aromatischen Ringebenen, der die grundsätzliche Lage der Ringsysteme zueinander charakterisiert [126]. Zur genaueren Differenzierung können dann noch die θ -Winkel zwischen den Normalen der Ringebenen und dem Vektor zwischen den Masseschwerpunkten (R_{cen}) herangezogen werden [126]. In Tabelle 20 sind alle möglichen erlaubten Winkelkombinationen der verschiedenen Konformationen zusammengefasst.

Tabelle 20: Idealwerte der Winkel bei π - π -Wechselwirkungen

| Interaktionstyp | γ | θ_1 | θ_2 |
|---------------------------|-----------------------------------|--|-------------------------------------|
| <i>T-Stacking</i> | $80^\circ < \gamma < 100^\circ$ | $0^\circ < \theta_1 < \pm 30^\circ$ | $75^\circ < \theta_2 < 105^\circ$ |
| | | $75^\circ < \theta_1 < 105^\circ$ | $0^\circ < \theta_2 < \pm 30^\circ$ |
| <i>Sandwich</i> | $0^\circ < \gamma < \pm 35^\circ$ | $0^\circ < \theta_1 < \pm 20^\circ$ | |
| <i>Parallel displaced</i> | | $\pm 20^\circ < \theta_1 < \pm 45^\circ$ | |

Abstandsabhängigkeit der Anordnungen

In einer ausführlichen statistischen Studie über π - π -Interaktionsnetzwerke unter Verwendung eines Datensatzes von 7848 Kristallstrukturen mit einer Sequenzidentität von

Strukturbiologische Parameter

maximal 90% konnten Chourasia *et al.* nachweisen, dass die Orientierung von zwei aromatischen Aminosäuren zueinander abhängig von ihrer Zentroid-Distanz ist (Abb. 15) [128]. Sie untersuchten dabei Netzwerke verschiedener Größe, beginnend bei Netzwerken, die aus nur zwei Aminosäuren bestanden (2π) bis hin zu Netzwerken, die zwölf Aminosäuren enthielten (12π) [128]. Ein ähnliches Verhalten konnten McGaughey *et al.* bei der Analyse ihres Datensatzes von 505 nicht-homologen Proteinstrukturen aus der PDB nachweisen [126].

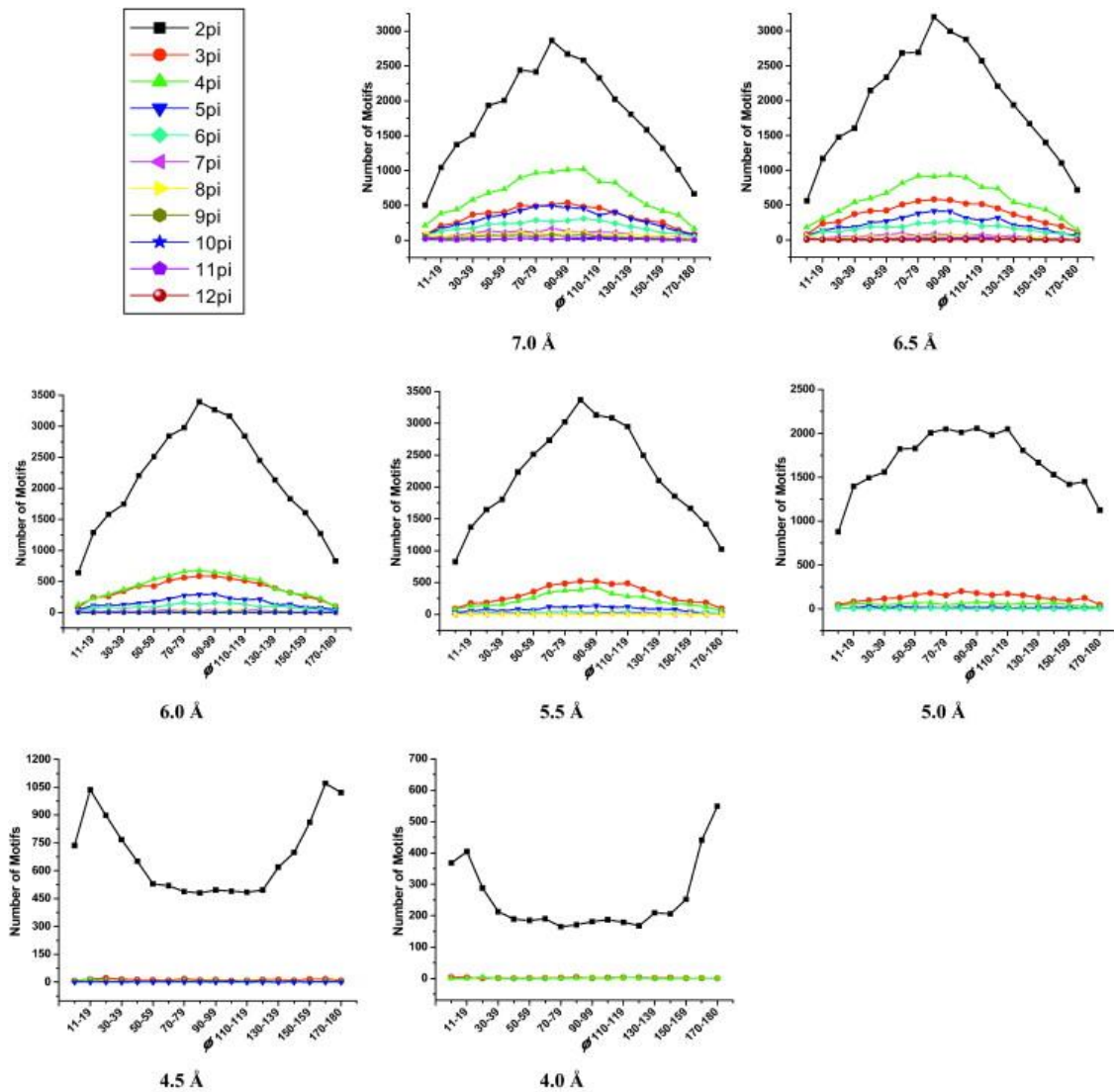


Abb. 15: Ebenenwinkel verschiedener π - π -Interaktionsnetzwerke bei verschiedenen Zentroid-Distanzen [128]

In der Abbildung sind die Anzahlen von verschiedenen gefundenen Interaktionsnetzwerken gegen den Ebenenwinkel bei verschiedenen *Cutoff*-Distanzen aufgetragen. Die verschiedenen farbigen Linien repräsentieren die Ergebnisse verschieden großer Netzwerke. Man kann erkennen, dass sich erst ab einem Mindestabstand von 5,0 Å eine Normalverteilung ausbildet. Bei geringeren Abständen findet man eine inverse Gauß-Verteilung vor.

Strukturbiologische Parameter

Die Verteilungen der γ -Winkel in Abb. 16 stellt nahezu eine Reproduktion der Ergebnisse von Chourasia *et al.* [128] für 2π -Systeme dar (Abb. 15). Bei geringen Abständen von weniger als 4,5 Å (Histogramme A bis C) finden sich vor allem parallele Orientierungen mit γ -Winkeln von 0° bis $\pm 30^\circ$. Erst bei Abständen von mehr als 4,5 Å können auch andere Anordnungen gefunden werden wobei insbesondere die Anzahl an *T-stacked* Interaktionen (γ -Winkel zwischen 80° und 100°) sehr stark zunimmt. McGaughey *et al.* konnten mit Hilfe von Molekülmechanik-Untersuchungen und RFF1-Energieberechnungen an einem Benzen-Dimer zeigen, dass die energetischen Minima von parallelen Orientierungen bzw. *T-stacked* Interaktionen bei Abständen der Zentroide von rund 3,5 Å bzw. 5,0 Å zu finden sind [126]. Dementsprechend bestätigen die hier gefundenen statistischen Abstandsverteilungen diese energetischen Berechnungen. Alles in allem zeigen diese Ergebnisse, dass bei geringen Zentroid-Abständen eine parallele Orientierung der aromatischen Ringe zueinander einer *T-Stacked*-Orientierung eher auftritt.

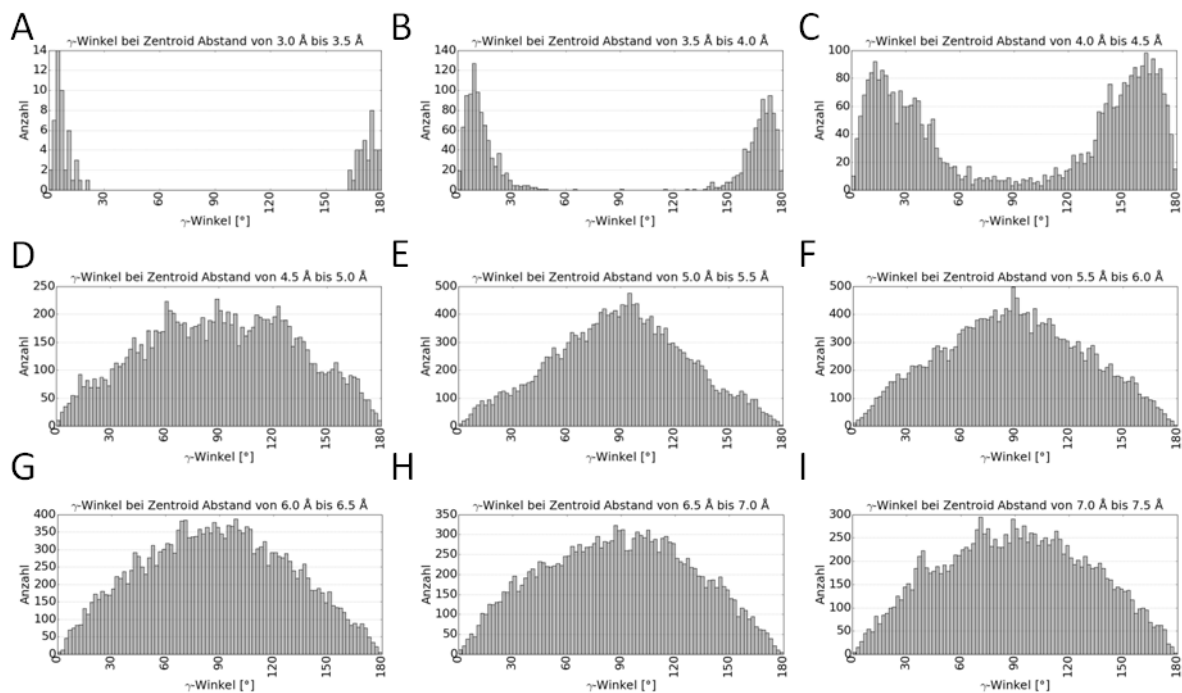


Abb. 16: Ebenenwinkel in Abhängigkeit vom Abstand der Masseschwerpunkte (XRD)

Die neun Histogramme zeigen die Verteilung der berechneten γ -Winkel für verschiedene 0,5 Å große Distanzklassen von 3,0 Å bis 7,5 Å. Es ist klar eine Präferenz von parallelen Konformationen für kleine Abstände (A bis C) zu erkennen. Bei Abständen von mehr als 4,5 Å steigt die Anzahl an anderen Konformationen.

Statistische Auswertung

Ein Vergleich der statistischen Auswertung der gefundenen π - π -Interaktionen innerhalb des Datensatzes XRD mit den Daten von Chourasia *et al.* [128] zeigt relativ starke Ähnlichkeiten (Tabelle 21). Es wurde allerdings ein leicht geringerer Anteil Histidin und Tyrosin sowie ein erhöhter Anteil von Tryptophan im Vergleich zu den Literaturergebnissen gefunden. Im hier genutzten Datensatz setzten sich die aromatischen Aminosäuren aus 20% Histidin, 35% Phenylalanin, 13% Tryptophan und 32% Tyrosin zusammen (Tabelle A.1). Dementsprechend resultieren die Unterschiede zwischen den aktuellen Ergebnissen und denen von Chourasia *et al.* [128] sehr wahrscheinlich nur aus den Zusammensetzungen der Datensätze.

Die Daten in Tabelle 21 zeigen zusätzlich, dass nur rund 20% aller gefundenen π - π -Wechselwirkungen eine der drei energetisch günstigsten Konformationen annehmen. Jaffe und Smith konnten mit Hilfe von quantenmechanischen Berechnungen von Benzen-Dimeren nachweisen, dass sowohl *sandwich*- als auch *T-stacked*-Konformationen Sattelpunkte eines Niedrig-Energie-Übergangs zwischen verschiedenen *parallel-displaced* Konformationen darstellen (Abb. A. 6) [129]. Dementsprechend können auch die Konformationen in den Übergangsbereichen der energetischen Minima (*T-stacked*, *sandwich* und *parallel displaced*) als energetisch günstig angesehen werden [129]. Das heißt die oben beschriebenen Winkel-Kriterien (Tabelle 20) können zwar zur Identifizierung der Konformation genutzt werden, sollten aber kein Ausschluss-Kriterium darstellen.

Tabelle 21: Vorkommen von aromatischen Aminosäuren in π - π -Wechselwirkungen

Die Prozen-te geben die Häufigkeit an, mit der die jeweilige Aminosäure an den Interaktionen beteiligt ist. Für den Datensatz XRD sind die Ergebnisse für alle Interaktionen sowie die Anordnungen nach Anwendung der Winkelkriterien aus Tabelle 20 angegeben. Letztere sind in Spalte 7 nochmals zusammengefasst.

| Aminosäure | Chourasia [128] | XRD Komplet | XRD sandwich | XRD T-stacked | XRD p. displaced | XRD Summe ⁴ |
|--------------|-----------------|------------------|----------------|-----------------|------------------|------------------------|
| Histidin | 13849 (22%) | 19994 (16,5%) | 665 (29,3%) | 1243 (10,9%) | 2960 (23,2%) | 4868 (18,4%) |
| Phenylalanin | 23686 (37%) | 46182 (38,0%) | 671 (29,6%) | 4834 (42,5%) | 4544 (35,5%) | 10049 (38,0%) |
| Tryptophan | 4208 (6%) | 17061 (14,0%) | 345 (15,2%) | 1983 (17,4%) | 1609 (12,6%) | 3937 (14,9%) |
| Tyrosin | 22417 (35%) | 38227 (31,5%) | 585 (25,8%) | 3309 (29,1%) | 3671 (28,7%) | 7565 (28,6%) |

Kombinationen aromatischer Aminosäuren

Anjana *et al.* untersuchten ausgehend von einem Datensatz bestehend aus Kristallstrukturen und NMR-Strukturen (insgesamt 12206) mit einer maximalen Sequenzidentität von 90% mögliche Kombinationen von aromatischen Aminosäuren bei π - π -Wechselwirkungen [130]. Die Analyse dieser Kombinationen in der vorliegenden Arbeit produzierte relativ ähnliche Ergebnisse (Tabelle 22). Die Daten von Anjana *et al.* beziehen sich nur auf bestimmte, nicht näher definierte Ebenenwinkel [130]. Etwaige Unterschiede zu den Ergebnissen der vorliegenden Arbeit resultieren dann sehr wahrscheinlich aus dieser Auswahl (Tabelle der Originalliteratur enthält ~85000 Interaktionen von ~186000 gefundenen Interaktionen).

⁴ #Summe = #sandwich + #T-stacked + #parallel displaced

Tabelle 22: Statistische Auswertung möglicher Kombinationen von aromatischen Interaktionen

Die Prozentzahlen geben den Anteil der jeweiligen Kombination an allen gefundenen Kombinationen an. Für den Datensatz XRD sind die Ergebnisse für alle Interaktionen (nur Abstandskriterium) sowie die drei energetisch günstigsten Konformationen (Winkelkriterien nach Tabelle 20) angegeben. Letztere sind in Spalte 7 nochmals zusammengefasst.

| Kombination | Anjana [130] | XRD Komplett | XRD Sandwich | XRD T-Stacked | XRD P. displaced | XRD Geometrie |
|-------------|------------------|------------------|----------------|-----------------|------------------|-----------------|
| His-His | 5008 (6,2%) | 4397 (5,9%) | 73 (9,9%) | 81 (2,0%) | 304 (7,2%) | 458 (5,1%) |
| His-Phe | 7660 (9,5%) | 4692 (6,3%) | 90 (12,2%) | 283 (7,2%) | 384 (9,1%) | 757 (8,5%) |
| His-Trp | 2791 (3,5%) | 2329 (3,1%) | 76 (10,3%) | 103 (2,6%) | 283 (6,7%) | 462 (5,2%) |
| His-Tyr | 6304 (7,8%) | 4886 (6,6%) | 87 (11,8%) | 177 (4,5%) | 458 (10,9%) | 722 (8,1%) |
| Phe-Phe | 15055 (18,7%) | 20710 (27,8%) | 113 (15,4%) | 1145 (29,0%) | 994 (23,6%) | 2252 (25,3%) |
| Phe-Trp | 9217 (11,4%) | 6529 (8,8%) | 29 (3,9%) | 449 (11,4%) | 298 (7,1%) | 776 (8,7%) |
| Phe-Tyr | 18762 (23,3%) | 13534 (18,2%) | 148 (20,1%) | 687 (17,4%) | 682 (16,2%) | 1517 (17,0%) |
| Trp-Trp | 1931 (2,4%) | 2716 (3,6%) | 18 (2,4%) | 264 (6,7%) | 91 (2,2%) | 373 (4,2%) |
| Trp-Tyr | 6731 (8,3%) | 4558 (6,1%) | 37 (5,0%) | 239 (6,0%) | 211 (5,0%) | 487 (5,5%) |
| Tyr-Tyr | 7217 (8,9%) | 10209 (13,7%) | 65 (8,8%) | 527 (13,0%) | 507 (12,0%) | 1099 (12,3%) |

2.2.5 Hydrophobe Wechselwirkungen

Identifizierung von hydrophoben Wechselwirkungen

Das eine Wechselwirkung zwischen zwei Atome die sich in einer räumlichen Nähe von 3,0 Å bis 6,5 Å befinden möglich ist, wurde von Li *et al.* anhand von statistischen Auswertungen von Kristallstrukturen nachgewiesen [131]. Bei der Analyse der Abstandsverteilung konnten sie zwei Maxima für Abstände von 4,05 Å und 5,35 Å finden. Sie diskutierten, dass diese Abstände identisch zu den Energieminima quantenmechanischer Berechnungen des CCSD-Potentials (*coupled cluster single and double excitation Potential*) [132] eines Ethan-Dimers [133] waren. Anhand der erhaltenen potentiellen Energien von -0,27 kcal/mol bzw. -0,13 kcal/mol kamen sie zu dem Schluss, dass der Energiebeitrag von hydrophoben

Wechselwirkungen zwar relativ klein sei, aufgrund der großen Anzahl von Interaktionen jedoch einen wichtigen Beitrag zur Stabilisierung des Proteins beiträgt [131].

In der vorliegenden Arbeit wurden hydrophobe Wechselwirkungen unter Anwendung dieses Abstandskriteriums durch einfache Abstandsmessungen zwischen zwei Atomen hydrophober Aminosäure-Seitenketten (Tabelle 23) quantitativ bestimmt. Hydrophobe *Cluster* wurden mit Hilfe einer doppelten Suche innerhalb von *k-d* Bäumen [134] identifiziert. Dabei wurde eine *k-d*-Suche für alle Atome hydrophober Seitenketten durchgeführt und für jedes gefundene Nachbar-Atom eine erneute Suche gestartet (siehe Abschnitt 6.2).

Tabelle 23: Genutzte Atome zur Bestimmung von hydrophoben Interaktionen

| Aminosäure | Genutzte Atome |
|--------------|---|
| Alanin | C β |
| Isoleucin | C β , C γ_1 , C γ_2 |
| Leucin | C β , C γ , C δ_1 , C δ_2 |
| Methionin | C β , C γ , S δ , C ϵ |
| Phenylalanin | C β , C γ , C δ_1 , C δ_2 , C ϵ_1 , C ϵ_2 , C ζ |
| Tryptophan | C β , C γ , C δ_1 , C δ_2 , C ϵ_2 , C ϵ_3 , C ζ_2 , C ζ_3 , C η_2 |
| Tyrosin | C β , C γ , C δ_1 , C δ_2 , C ϵ_1 , C ϵ_2 , C ζ |
| Valin | C β , C γ_1 , C γ_2 |

Interaktionskombinationen

Um bestimmen ob bestimmte Aminosäuren besonders häufig miteinander wechselwirken, wurde eine statistische Analyse aller möglichen Kombinationen durchgeführt. Die paarweisen Korrelationen in Tabelle 24, berechnet anhand der Formeln (16) bis (19) in Abschnitt 6.4, machen deutlich, dass es nur sehr wenige präferierte Kombinationen innerhalb der Datensätze gibt. Lediglich die Kombinationen Ala/Ala, Trp/Trp, Tyr/Trp, Tyr/Tyr und Met/Met findet man signifikant häufiger vor als zu erwarten war. Bei den Kombinationen Trp/Trp, Tyr/Trp sowie Tyr/Tyr handelt es sich sehr wahrscheinlich um π - π -Wechselwirkungen, deren Bindungsenergie mit $\sim -2,0$ kcal/mol [127] deutlich günstiger als die $\sim -0,2$ kcal/mol einer hydrophoben Wechselwirkung ist [131].

Tabelle 24: Statistische Auswertung von Kombinationen hydrophober Wechselwirkungen der Datensätze XRD, MBD und SMD

Die Zahlen geben die paarweisen Korrelationen berechnet anhand der Methode von Lifson und Sander [135] (siehe auch Formeln (16) bis (19) in Abschnitt 6.4) an. Jeder Wert gibt dabei an ob eine Kombination häufiger (> 1) oder seltener (< 1) als zu erwarten vorkommt. Signifikant sind Werte $< 0,8$ oder $> 1,2$.

| | Ala | Ile | Leu | Met | Phe | Trp | Tyr | Val |
|-----|-------------|------|------|-------------|------|-------------|-------------|------|
| Ala | 1,27 | | | | | | | |
| Ile | 0,92 | 1,12 | | | | | | |
| Leu | 0,95 | 1,00 | 1,10 | | | | | |
| Met | 1,00 | 0,97 | 0,92 | 1,52 | | | | |
| Phe | 0,89 | 0,99 | 0,98 | 1,07 | 1,10 | | | |
| Trp | 0,97 | 0,90 | 0,94 | 1,04 | 1,06 | 1,49 | | |
| Tyr | 1,01 | 0,97 | 0,89 | 1,06 | 1,07 | 1,22 | 1,24 | |
| Val | 1,01 | 1,02 | 0,98 | 0,94 | 0,95 | 0,92 | 0,95 | 1,10 |

Hydrophobe Cluster

Kalinowska *et al.* konnten bei einer Analyse eines repräsentativen Datensatzes von Proteinstrukturen, basierend auf dem von Devlin vorgeschlagenen Datensatz [136], zeigen, dass nahezu alle Proteinklassen einen hydrophoben Kern besitzen [137]. Sie nutzten dabei die *fuzzy oil drop* Methode [138, 139] die auf der von Kauzmann entwickelten *oil drop* Methode [140] basiert und mit Hilfe von drei-dimensionalen Gaußverteilungen den hydrophoben Kern von Proteinen von der hydrophilen Hülle differenzieren kann [137]. Sie schlossen aus ihren Ergebnissen, dass hydrophobe Kerne ein essentielles strukturelles Merkmal von Proteinstrukturen sind. Aus diesem Grund wurde der Datensatz XRD auf hydrophobe *Cluster* untersucht.

Die Ergebnisse der Analyse von hydrophoben *Clustern* (Tabelle 25) konnte zeigen, dass knapp zwei Drittel aller gefundenen *Cluster* aus zehn oder weniger Aminosäuren bestehen (66,6%). Am häufigsten sind dabei Gruppen von drei, vier oder fünf Aminosäuren die sich zu einem hydrophoben *Cluster* zusammenlagern. Bei diesen kleineren *Clustern* handelt es sich möglicherweise um Bereiche innerhalb von ungeordneten Regionen der Proteinstruktur oder um kleinere β -Faltblätter. Die großen *Cluster* (mehr als 25 Aminosäuren) repräsentieren die hydrophoben Kerne von Proteinstrukturen.

Tabelle 25: Hydrophobe Cluster im Datensatz XRD

In der Tabelle sind sämtliche gefundenen hydrophoben Cluster des Datensatzes XRD zusammengefasst. Der Cluster-Name entspricht der Anzahl an unterschiedlichen Aminosäuren die in einem Cluster vorkommen. Cluster mit einer Größe von mehr als 25 Aminosäuren wurden zusammengefasst.

| Cluster | # | % | Cluster | # | % | Cluster | # | % | Cluster | # | % |
|---------|-------|-----|---------|-------|-----|---------|-------|-----|---------|------|-----|
| 2 | 30186 | 5,6 | 9 | 32233 | 6,0 | 15 | 16282 | 3,0 | 21 | 5222 | 1,0 |
| 3 | 50613 | 9,4 | 10 | 30186 | 5,6 | 16 | 14030 | 2,6 | 22 | 4133 | 0,8 |
| 4 | 50381 | 9,4 | 11 | 27037 | 5,0 | 17 | 11468 | 2,1 | 23 | 3215 | 0,6 |
| 5 | 47155 | 8,8 | 12 | 23428 | 4,4 | 18 | 9839 | 1,8 | 24 | 2694 | 0,5 |
| 6 | 43588 | 8,1 | 13 | 21444 | 4,0 | 19 | 8192 | 1,5 | 25 | 1985 | 0,4 |
| 7 | 39036 | 7,3 | 14 | 18496 | 3,4 | 20 | 6507 | 1,2 | 26 - 45 | 5797 | 1,1 |
| 8 | 35130 | 6,5 | | | | | | | | | |

2.2.6 Wasserstoffbrücken des Protein-Rückgrates

Wasserstoffbrücken zwischen den Atomen des Proteinrückgrates sind essentiell für die Ausbildung von Sekundärstrukturen und liefern einen großen Beitrag zur Faltung des gesamten Proteins [141]. Da in *in-silico* generierten Proteinstrukturmodellen oder experimentell aufgeklärten Proteinstrukturen oftmals Wasserstoff-Atome fehlen [142], musste ein vereinfachter Ansatz zur Bestimmung von Wasserstoffbrücken genutzt werden. Am effektivsten erwies sich der Ansatz von Tosatto [143], bei dem eine Wasserstoffbrücke zwischen den Atomen des Proteinrückgrates nur anhand von drei Abstandskriterien identifiziert werden kann (Abb. 17). Dabei müssen sich ein Carbonyl-Sauerstoff und ein Stickstoff-Atom ($d(O,N)$) in räumlicher Nähe von 2,0 Å bis 4,0 Å befinden. Zusätzlich muss der Abstand zwischen Carbonyl-Kohlenstoff und Stickstoff ($d(C,N)$) größer als der Abstand von Carbonyl-Sauerstoff und Stickstoff sein. Analog muss der Abstand zwischen Carbonyl-Sauerstoff und dem $C\alpha$ -Atom der Partneramino-säure ($d(O,CA)$) größer als der Abstand zwischen Carbonyl-Sauerstoff und Stickstoff sein.

Strukturbiologische Parameter

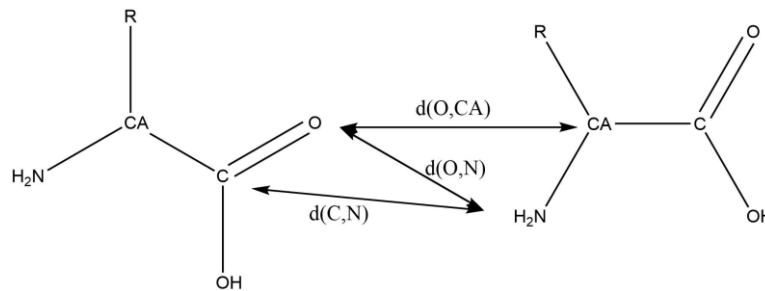


Abb. 17: Distanzkriterien zur Bestimmung von Wasserstoffbrücken

Die Abbildung zeigt schematisch die Abstandskriterien von Wasserstoffbrücken des Rückgrates. Für eine Wasserstoffbrücke muss als erstes $d(O,N)$ zwischen 2,0 Å und 4,0 Å liegen und außerdem müssen $d(O,CA)$ und $d(C,N)$ größer sein als $d(O,N)$.

Ausgehend von diesem vereinfachten Bestimmungsansatz wurden sämtliche Datensätze auf Wasserstoffbrücken überprüft. Dabei wurde für jede Wasserstoffbrücke zusätzlich der C-O-N-Winkel berechnet. In Abb. 18 sind die berechneten O-N-Abstände und C-O-N-Winkel in Form von Histogrammen zusammengefasst. Bei den Abständen ist ein klares Maximum von 3,0 Å erkennbar was mit der Standardlänge einer Wasserstoffbrücke komplett übereinstimmt [144]. Bei den C-O-N-Winkeln sind zwei Maxima im Bereich von $\sim 110^\circ$ und $\sim 155^\circ$ erkennbar. Des Weiteren gibt es keine Winkel unterhalb von 80° , da für einen solchen Winkel der Abstand zwischen dem Stickstoff-Atom und dem Carbonyl-Kohlenstoff ($d(C,N)$) kleiner sein müsste als der Abstand zwischen Carbonyl-Sauerstoff und dem Stickstoff-Atom ($d(N,O)$) (siehe Abb. A. 7).

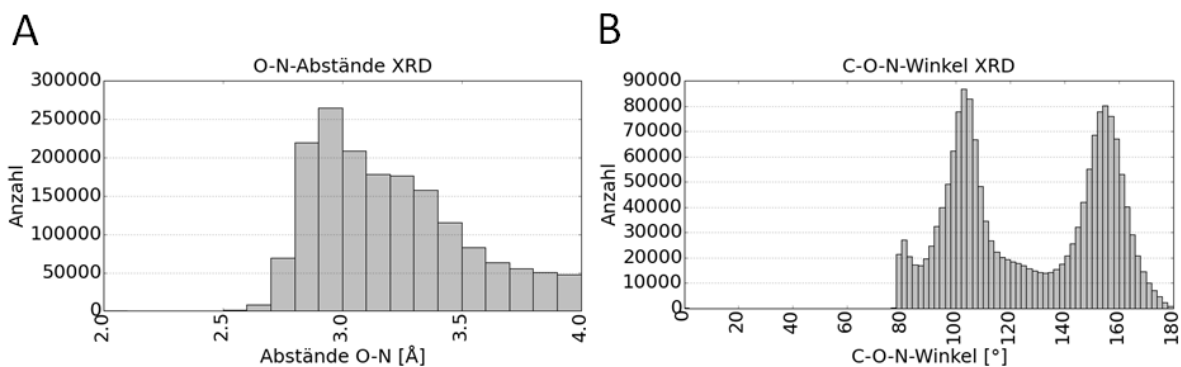


Abb. 18: Abstände und Winkel von Wasserstoffbrücken des Datensatzes XRD

Die Abbildung zeigt die Histogramme der O-N-Abstände (A) sowie der C-O-N-Winkel (B) aller gefundenen Wasserstoffbrücken des Datensatzes XRD. Bei den Abständen ist ein klares Maximum von 3,0 Å erkennbar, während bei den Winkeln zwei Maxima von $\sim 110^\circ$ und $\sim 155^\circ$ sichtbar sind.

Eine Analyse der Sequenzseparationen der Wasserstoffbrücken (Abb. 19 und Tabelle A.2) zeigt, dass rund 80% aller gefundenen Wasserstoffbrücken eine Sequenzseparation von 6

Strukturbiologische Parameter

oder weniger Aminosäuren aufweisen. Da die Strukturen einen helikalen Anteil von rund 35% sowie einen *Coil*-Anteil von rund 42% aufweisen ist dies nicht verwunderlich. Die größeren Sequenzabstände resultieren zu großen Teilen aus β -Faltblättern deren Stränge eine hohe Sequenzseparation aufweisen. Eine genauere Betrachtung der Sequenzseparationen von mehr als 50 Aminosäuren (Tabelle A.2) zeigt, dass es zwar relativ viele weit voneinander entfernte (im Bezug auf die Sequenz) Partner gibt, jedoch der Großteil in einem Bereich bis maximal 400 Aminosäuren zu finden ist. Dementsprechend existieren nur sehr wenige große Proteine (mehr als 1000 Aminosäuren) bei denen es eine Interaktion von N- und C-Terminus gibt. Wahrscheinlich handelt es sich bei den größeren Proteinen um Multi-Domänen-Proteine, deren Termini sich nicht räumlicher Nähe befinden.

Strukturbiologische Parameter

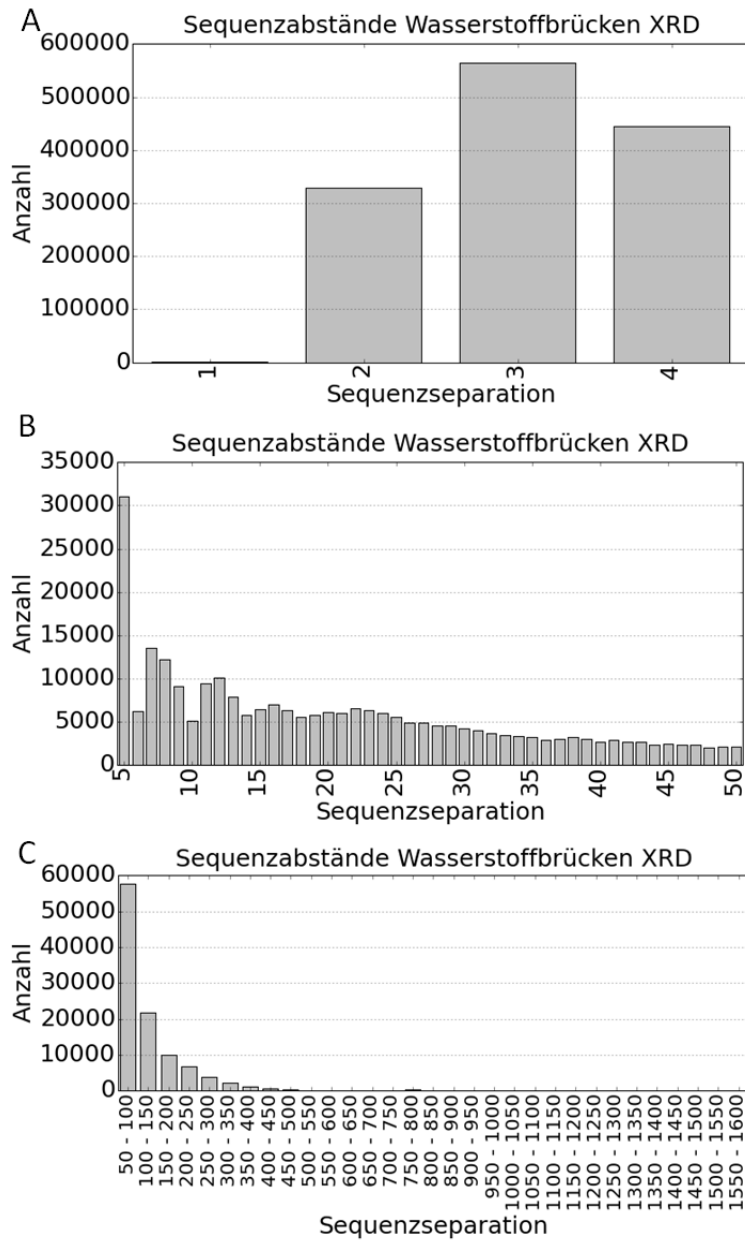


Abb. 19: Sequenzseparationen aller gefundenen Wasserstoffbrücken des Datensatzes XRD
 Die drei Diagramme zeigen die Anzahl an Wasserstoffbrücken mit bestimmten Sequenzseparationen. In (A) sind kurze, in (B) mittlere und in (C) lange Interaktionsdistanzen zusammengefasst. Zur besseren Übersichtlichkeit wurden in (C) die Werte in *Bins* á 50 Aminosäuren zusammengefasst.

2.3 Sekundärstrukturen

Seit der Entdeckung von α -Helices und β -Strängen durch Pauling und Corey [2–4] spielen Sekundärstrukturen bei der Betrachtung von Proteinstrukturen eine essentielle Rolle. Aufgrund ihrer Bedeutung für die gesamte Faltung eines Proteins [141] wurden Algorithmen entwickelt, mit deren Hilfe sie anhand von Atomkoordinaten zugewiesen werden können [145–148]. Seit der Entwicklung von DSSP [145] wurden immer wieder neue Programme wie PREDATOR [147, 148], Stride [146], PALSSE [149], KAKSI [150] oder P-SEA [151] publiziert und ihre Ergebnisse miteinander verglichen [152]. In der vorliegenden Arbeit wurde ebenfalls ein Zuweisungsmechanismus etabliert (siehe Abschnitt 6.3), mit dessen Hilfe eine verbesserte Untersuchung von Sekundärstrukturen ermöglicht werden sollte.

2.3.1 Bestehende Zuweisungsalgorithmen

Problemstellung

Das Ziel von Zuweisungsmechanismen ist es, ausgehend von den Atomkoordinaten einer Proteinstruktur verschiedene Sekundärstrukturelemente zu identifizieren [153]. Um dies zu ermöglichen, gibt es kein einheitliches Konzept, sondern es wurden immer wieder neue Programme entwickelt, um verschiedenen Aspekten von Sekundärstrukturen Rechnung zu tragen [153].

DSSP

Das von Kabsch und Sander entwickelte Programm DSSP [145] ist eines der am meisten genutzten Programme, wenn es um die Zuweisung von Sekundärstrukturen geht [150]. Die Zuweisung beruht dabei auf Wasserstoffbrücken des Proteinerückgrates und nutzt ein Coulomb-Energiepotential zur Bestimmung energetisch günstiger Wasserstoffbrücken [153].

$$E = f\delta_1\delta_2\left(\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)}\right) \quad (2)$$

In der von Lifson *et al.* [154] postulierten Formel zur Berechnung der Energie einer Wasserstoffbrücke (Formel (2)) steht f für den Dimensionsfaktor ($332 \text{ (\AA kcal)}/(e^2 \text{ mol})$), δ_1 ($0,42e$) und δ_2 ($0,20e$) für die Polarladungen der beiden beteiligten Gruppen und $r(AB)$ für den Abstand zwischen Atom A und Atom B [152]. Das Programm erkennt eine Wasserstoffbrücke erst dann an, wenn die berechnete Energie kleiner als $-0,5 \text{ kcal/mol}$ ist

[153]. Da in Kristallstrukturen oftmals die nötigen Wasserstoff-Atome fehlen [142], nutzt DDSP ungefähre Positionen [153]. Dabei wird angenommen, dass die C-O-Doppelbindung und die N-H-Bindung des Partners parallel verlaufen [153]. Die Position des Wasserstoffes wird dann durch die Nutzung eines Vektors mit einer Länge von 1\AA , der Länge einer N-H-Bindung [155], und der Richtung der C-O-Doppelbindung extrapoliert [153]. Durch die Verwendung des Stickstoff-Atoms als Start-Punkt dieses Vektors kann dann die Position des Wasserstoff-Atoms bestimmt werden [153].

Ausgehend von den gefundenen Wasserstoffbrücken identifiziert DSSP insgesamt acht verschiedene Sekundärstrukturen in insgesamt vier Kategorien. Die erste Kategorie umfasst die sogenannte *Turn*-Konformation, die Wasserstoffbrücken zwischen Aminosäure i und Aminosäure $i+n$ beinhaltet [152]. Bei mindestens zwei aufeinanderfolgenden Aminosäuren mit diesem Wasserstoffbrücken-Muster geht das Programm von einer helikalen Struktur aus [153]. Für eine α -Helix (H) müssen es zwei Aminosäuren mit dem Muster $i/i+4$, für eine 3_{10} -Helix (G) zwei Aminosäuren mit dem Muster $i/i+3$ und für eine π -Helix (I) zwei Aminosäuren mit dem Muster $i/i+5$ sein [153]. Sämtliche Aminosäuren, die ein $i/i+n$ Wasserstoffbrückenmuster aufweisen und nicht in Helices liegen werden dann als *Turn* (T) angesehen [153]. In der zweiten Kategorie werden Wasserstoffbrücken zwischen zwei nicht-überlappenden, drei Aminosäure langen Strängen ($i-1/i+1$ und $j-1/j+1$) betrachtet [152]. Diese Stränge formen eine parallele Brücke, wenn es zwei Wasserstoffbrücken der Form $i-1/j$ und $j/i+1$ oder der Form $j-1/i$ und $i/j+1$ gibt [152]. Eine anti-parallele Brücke wird dann gebildet, wenn es zwei Wasserstoffbrücken der Form i/j und j/i oder $i-1/j+1$ und $j-1/i+1$ gibt [152]. Einzelne Brücken werden als B (*Bridge*) gekennzeichnet und aufeinanderfolgende Brücken bilden sogenannte Leitern, die dann als β -Faltblätter (E) identifiziert werden können [152]. Für die dritte Kategorie berechnet DSSP den Winkel $C\alpha_{i-2}-C\alpha_i-C\alpha_{i+2}$ und markiert etwaige noch nicht zugewiesenen Aminosäuren als *Bend* (S) [152]. In der letzten Kategorie werden dann alle restlichen Aminosäuren zusammengefasst und als *Coil* (C) gekennzeichnet [152].

Stride

Das von Frishman und Argos entwickelte Programm Stride [147, 148] arbeitet mit einem ähnlichen Prinzip wie DSSP [153]. Die Wasserstoffbrückenenergie wird dabei jedoch nicht über einen Coulomb-Ansatz [154] sondern über einen empirischen Ansatz [156, 157]

Strukturbiologische Parameter

berechnet. Diese Energie ist dabei abhängig von der N-O-Bindungsenergie und drei verschiedenen Bindungswinkeln [153]. Zusätzlich nutzt das Programm noch Phi/Psi-Torsionswinkel und berechnet Wahrscheinlichkeiten für Helices und Stränge anhand der Abstände der berechneten Torsionswinkel zu den jeweiligen Regionen im Ramachandran-Plot [153].

Die Definition einer α -Helix (H) erfolgt analog zu DSSP (zwei aufeinanderfolgende Aminosäuren mit $i/i+4$ Wasserstoffbrücken) [153]. Diese Definition wird an den Rändern der Helices aufgeweicht um Aminosäuren, die zwar die Torsionswinkelkriterien erfüllen aber nicht ins Wasserstoffbrückenmuster passen, ebenfalls einzuschließen [152]. Für 3_{10} -Helices (G) und π -Helices (I) wird ein identisches System (Wasserstoffbrücken $i/i+3$ für G und $i/i+5$ für I) verwendet [152]. Als β -Stränge (E) werden Bereiche von mindestens zwei aufeinanderfolgenden Aminosäuren gekennzeichnet, die mindestens eine von fünf möglichen Wasserstoffbrückenmustern annehmen [153]. Einzelne Brücken werden hier ebenfalls mit einem B markiert [152]. Die Zuweisung von *Turns* (T) erfolgt aufgrund der Phi/Psi-Torsionswinkel [152] und folgt den Definitionen von Richardson [158] sowie Wilmot und Thornton [159]. Wenn keines der genannten Kriterien erfüllt wurde, wird die entsprechende Aminosäure als *Coil* (C) gekennzeichnet [153].

P-SEA

Das Programm P-SEA von Labesse *et al.* weist Sekundärstrukturen nur anhand der Koordinaten von $C\alpha$ -Atomen zu [151]. Dabei werden nur Helices (Mindestlänge fünf Aminosäuren), Stränge (Mindestlänge drei Aminosäuren) und *Coil* Regionen zugewiesen [153]. Die Zuweisung erfolgt dabei aufgrund der Abstandskriterien $d2i$ (Abstand von $C\alpha_{i-1}$ zu $C\alpha_{i+1}$), $d3i$ ($C\alpha_{i-1}$ zu $C\alpha_{i+2}$) und $d4i$ ($C\alpha_{i-1}$ zu $C\alpha_{i+3}$) sowie dem Winkel τ_i ($C\alpha_{i-1}$ - $C\alpha_i$ - $C\alpha_{i+1}$) und dem Torsionswinkel α_i ($C\alpha_{i-1}$ - $C\alpha_i$ - $C\alpha_{i+1}$ - $C\alpha_{i+2}$) [151].

Eine Helix ist dann definiert als ein Bereich von mindestens fünf Aminosäuren, in dem jede Aminosäure entweder die Abstandskriterien $d3i$ und $d4i$ oder die Winkel-Kriterien τ_i und α_i für Helices (Tabelle 26) erfüllt [151]. Dieser Bereich kann an den Rändern um eine Aminosäure erweitert werden, wenn diese entweder das Abstandskriterium $d3i$ oder das Winkel-Kriterium τ_i für Helices erfüllt [151].

Tabelle 26: Von P-SEA genutzte Kriterien zur Zuweisung von Sekundärstrukturen [151]

| Kriterium | Helix | Strang |
|-------------------------|---------------------------------------|--|
| Winkel τ | $89^\circ \pm 12^\circ$ | $124^\circ \pm 14^\circ$ |
| Torsionswinkel α | $50^\circ \pm 20^\circ$ | $-170^\circ \pm 45^\circ$ |
| Abstand d_{2i} | $5,5 \text{ \AA} \pm 0,5 \text{ \AA}$ | $6,7 \text{ \AA} \pm 0,6 \text{ \AA}$ |
| Abstand d_{3i} | $5,3 \text{ \AA} \pm 0,5 \text{ \AA}$ | $9,9 \text{ \AA} \pm 0,9 \text{ \AA}$ |
| Abstand d_{4i} | $6,4 \text{ \AA} \pm 0,6 \text{ \AA}$ | $12,4 \text{ \AA} \pm 1,1 \text{ \AA}$ |

Für einen β -Strang müssen mindestens drei aufeinanderfolgende Aminosäuren entweder die Abstandskriterien d_{2i} , d_{3i} und d_{4i} oder die Winkelkriterien τ_i und α_i für Stränge (Tabelle 26) erfüllen [151]. Auch hier ist eine Verlängerung an den Enden möglich, wenn die entsprechenden Aminosäuren das Abstandskriterium d_{3i} für Stränge einhalten [151]. Stränge mit einer Länge von drei Aminosäuren werden nur dann akzeptiert, wenn sie Teil eines größeren Faltblattes sind [151]. Dies wird durch die Bestimmung von $C\alpha$ -Kontakten (Abstand $4,2 \text{ \AA}$ bis $5,2 \text{ \AA}$) zu umliegenden Aminosäuren überprüft [151]. Nur wenn die drei Aminosäuren im Strang in der Summe mindestens fünf Kontakte ausbilden, wird er als Teil eines Faltblattes angesehen [151].

KAKSI

Analog zum Programm P-SEA [151] nutzt auch das Zuweisungsprogramm KAKSI von Martin *et al.* $C\alpha$ -Abstandskriterien [150]. Für Helices werden die Abstände $i/i+2$, $i/i+3$, $i/i+4$ und $i/i+5$ beachtet wobei noch eine Unterscheidung zwischen Aminosäuren im Kern und an den Rändern der Helix getroffen wird [150]. Im Gegensatz dazu werden für Stränge Abstände innerhalb eines Stranges ($i/i+2$) sowie Abstände zwischen zwei Strängen i und j ($i+1/j+1$, i/j oder $i/j+2$, $i+2/j+1$) als Zuweisungsgrundlage genutzt [150]. Zusätzlich zu den $C\alpha$ -Abstandsberechnungen definierten Martin *et al.* noch Phi/Psi-Karten, die spezifische Torsionswinkel für Helices bzw. Stränge enthielten [150].

Die Zuweisung der einzelnen Sekundärstrukturen erfolgt dann nach einem heuristischen Verfahren, wobei für Helices Abstands- oder Torsionswinkelkriterien und für Stränge Abstands- und Torsionswinkelkriterien erfüllt sein müssen [150]. Etwaige nicht als Helix definierte Knicke in Helices werden dann unter Berechnung des Unterschiedes der Torsionswinkelpaare von zwei aufeinanderfolgenden Aminosäuren $j/j+1$ ausgeglichen [150].

Das Problem des Zustandsmodells

Ein Vergleich der Ergebnisse von verschiedenen Zuweisungsprogrammen wird durch die unterschiedlichen Zustandsmodelle erschwert. Während DSSP acht verschiedene Sekundärstrukturen zuweist [145], definiert Stride sieben Zustände [146] und P-SEA [151] sowie KAKSI [150] nutzen sogar nur drei Zustände. Cuff und Barton konnten für die Überprüfung von Sekundärstrukturvorhersage-Programmen nachweisen, dass die Verwendung von unterschiedlichen Reduktionsmethoden (Umwandlung von acht Zuständen in drei oder vier Zustände) einen großen Einfluss auf die erhaltenen Ergebnisse hat [160]. Dementsprechend ist es nur logisch, dass ein solcher Einfluss auch beim Vergleich von Sekundärstrukturzuweisungsprogrammen untereinander vorhanden ist.

2.3.2 Vergleich verschiedener Zuweisungsalgorithmen

Für alle Strukturen des Datensatzes XRD wurden Sekundärstrukturen mit dem neu entwickelten Algorithmus (siehe Abschnitt 6.3) und den Programmen DSSP, Stride, P-SEA sowie KAKSI (siehe Abschnitt 6.5 für mehr Informationen über die Zustandsreduktion) zugewiesen. Zusätzlich wurden, wenn möglich, die Angaben der Autoren zu Sekundärstrukturen aus den *pdb-files* entnommen. Die so erhaltenen Ergebnisse wurden dann detailliert miteinander verglichen.

Kurzer Überblick älterer Studien

Als Vergleichsgrundlage der erhaltenen Ergebnisse dienten drei ältere Studien, die im Folgenden kurz erläutert werden sollen. Der erste Vergleich wurde mit der Studie von Martin *et al.* [150] durchgeführt, die im Zuge der Präsentation ihres neuen Programms KAKSI bereits bestehende Mechanismen miteinander verglichen. Etwaige Unterschiede bestimmten sie dabei von prozentualen Übereinstimmungen zwischen den Zuweisungsprogrammen DSSP, Stride, P-SEA, KAKSI sowie den Autorenangaben [150]. Dazu nutzten sie drei verschiedene Datensätze mit Kristallstrukturen (689 Strukturen mit Auflösung $< 1,7 \text{ \AA}$, 624 Strukturen mit Auflösung zwischen $1,7 \text{ \AA}$ und $3,0 \text{ \AA}$ sowie 332 Strukturen mit Auflösung $> 3,0 \text{ \AA}$) sowie einen NMR-Datensatz [150]. Für alle vier Datensätze definierten sie eine maximale Sequenzidentität von 30% [150].

Zhang *et al.* verglichen die Programme DSSP, Stride, P-SEA und KAKSI anhand von 416 strukturell alignierten Proteinpaaren miteinander [161]. Das Alignment führten sie dabei mit Hilfe von CA-Aligner [162] sowie der FSSP-Datenbank [163] durch, da beide Programme nicht

auf Sekundärstrukturen bei der Erstellung des Alignments zurückgreifen [161]. Der so erhaltene Datensatz wies eine maximale Sequenzidentität von 29,6% auf und alle Vergleiche erfolgten anhand von prozentualen Übereinstimmungen [161].

Zu Letzt wurden die erhaltenen Ergebnisse mit einer Untersuchung von Cao *et al.* [164] verglichen. Auch hier fand ein Vergleich von DSSP, Stride, P-SEA und KAKSI anhand eines Datensatzes von 2817 Kristallstrukturen mit einer Sequenzidentität von maximal 30% statt [164]. Der allgemeine Vergleich der Zuweisungsprogramme erfolgte erneut anhand von prozentualen Übereinstimmungen [164]. Zusätzlich verglichen Cao *et al.* noch die Zuweisung von Helices und Strängen mit Hilfe des SOV-Scores, der die Übereinstimmung der Zuweisung zweier Fragmente beschreibt [165, 166].

Ablauf des Vergleichs

Um die verschiedenen Zuweisungsalgorithmen mit dem neu-entwickelten Mechanismus (siehe Abschnitt 6.3) sowie den Autorenangaben zu vergleichen, wurden sämtliche Strukturen des Datensatzes XRD einer Zuweisung mit Hilfe der Programme DSSP, Stride, P-SEA und KAKSI unterzogen (siehe Abschnitt 6.5 für mehr Informationen über die Programme und die genutzte Zustandsreduktion). Zusätzlich wurden die von den Autoren bestimmten Sekundärstrukturen den pdb-files (wenn vorhanden) entnommen. Ein Vergleich der verschiedenen Ergebnisse fand dann nur für Strukturen (5536), für die alle sechs Zuweisungen vorhanden waren, statt.

Der Vergleich der einzelnen Methoden untereinander erfolgte dann anhand der prozentualen Übereinstimmung der gefundenen Sekundärstrukturen. Dabei wurde ein Drei-Zustandsmodell (Helix, Strang, *Coil*) genutzt. Übereinstimmungen von Helices und Strängen wurden ebenfalls in Form von prozentualen Anteilen bestimmt, wobei für jeden Vergleich zwei Werte bestimmt wurden (Formel (3)).

$$\%H = \frac{N_{id}^H}{N_{Prog}^H} \quad (3)$$

In Formel (3) stehen N_{id}^H für die Anzahl an identischen Helix-Zuweisungen und N_{Prog}^H für die Anzahl an Helix-Zuweisungen der einzelnen Programme (DSSP, Stride, P-SEA, KAKSI, VAProTS sowie Autorenangaben). Jeder Vergleich wurde über den gesamten Datensatz durchgeführt, um eine Gesamt-Übereinstimmung zu bestimmen.

Sekundärstrukturgehalt

Strukturbiologische Parameter

In einem ersten Schritt wurden die von den verschiedenen Programmen bestimmten Anteile von Helices und Strängen im Datensatz XRD miteinander verglichen. Die Daten in Tabelle 27 zeigen, dass die Ergebnisse der vorliegenden Arbeit sehr gut mit denen von Martin *et al.* [150] übereinstimmen. Auffällig ist, dass P-SEA deutlich weniger Helices zuweist als die anderen Programme, während bei den Autorenangaben der Helix-Anteil deutlich über den Ergebnissen der verschiedenen Programme liegt. Der erhöhte Anteil an Helices könnte aus einer detaillierteren Analyse der Strukturen durch die Autoren im Vergleich zur reinen informatischen Berechnung der Sekundärstrukturen durch die Programme resultieren.

Tabelle 27: Sekundärstrukturgehalt des Datensatzes XRD bestimmt mit verschiedenen Programmen

In der Tabelle sind die prozentualen Anteile von Helices und Strängen im Datensatz XRD im Vergleich zu den Daten von Martin *et al.* [150] gezeigt. Die Anteile wurden mit Hilfe verschiedener Programme bestimmt bzw. direkt den *pdb-files* entnommen (Autorenangabe).

| Programm | Martin [150] | | Aktuelle Arbeit | |
|----------|--------------|--------|-----------------|--------|
| | Helix | Strang | Helix | Strang |
| DSSP | 35,9% | 22,5% | 36,5% | 23,5% |
| Stride | 36,4% | 22,6% | 37,6% | 22,9% |
| P-SEA | 32,1% | 23,7% | 33,0% | 24,9% |
| KAKSI | 36,8% | 22,0% | 38,5% | 23,4% |
| Autor | 40,5% | 20,3% | 42,4% | 21,8% |
| VAProTS | | | 35,6% | 22,1% |

Allgemeiner Vergleich der Zuweisungsprogramme

In einem ersten Schritt wurde analysiert, wie stark der neue Zuweisungsalgorithmus mit bereits bekannten Programmen übereinstimmt. Dazu wurden Ergebnisse des neuen Programms mit denen von DSSP, Stride, KAKSI und P-SEA sowie den Angaben der Autoren der Kristallstrukturen für den Datensatz XRD verglichen. Anhand der prozentualen Anteile in Tabelle 28 und Tabelle 29 ist erkennbar, dass die in der vorliegenden Studie erhaltenen Ergebnisse sehr stark von denen anderer Studien abweichen.

Tabelle 28: Vergleich verschiedener bekannter Zuweisungsprogramme mit Angaben der Autoren und dem hier genutzten neuen Mechanismus

Strukturbiologische Parameter

Die prozentualen Anteile geben die Gesamtübereinstimmung der verschiedenen Zuweisungsmechanismen zueinander an. Die Spalte Autoren fasst die Ergebnisse der Angaben in den pdb-files zusammen und die Spalte VAProTS die Ergebnisse des neu-entwickelten Zuweisungsalgorithmus.

| Programm | DSSP | Stride | P-SEA | KAKSI | Autoren | VAProTS |
|----------|------|--------|--------|--------|---------|---------|
| DSSP | X | 88,48% | 59,62% | 61,37% | 60,55% | 62,31% |
| Stride | | X | 58,87% | 60,77% | 59,36% | 61,42% |
| P-SEA | | | X | 82,31% | 71,35% | 88,52% |
| KAKSI | | | | X | 77,44% | 86,94% |
| Autoren | | | | | X | 76,03% |
| VAProTS | | | | | | X |

Dabei wird ein Vergleich von verschiedenen Zuweisungsmechanismen durch mehrere grundlegende Probleme erschwert. So stellten schon Cuff und Burton fest, dass die erhaltenen Ergebnisse von dem genutzten Zustandsmodell abhängig sind [160]. DSSP zum Beispiel nutzt in seiner Standard-Ausgabe ein Acht-Zustandsmodell [145], das für den Vergleich mit anderen Programmen in ein Drei-Zustandsmodell überführt werden muss. Cuff und Burton konnten zeigen das die prozentualen Übereinstimmungen von DSSP mit anderen Programmen zwischen den Ansätzen von Rost und Sander [167], Salamov und Solovyev [168] sowie Frishman und Argos [148] nachweisbaren Schwankungen unterworfen waren. Ein weiterer Aspekt der sowohl von Cuff und Barton als auch von Martin *et al.* [150] als problematisch angeführt wurde, waren die unterschiedlichen Mindestlängen, die die einzelnen Programme für die jeweiligen Sekundärstrukturmerkmale festlegen.

Tabelle 29: Prozentuale Übereinstimmungen zwischen verschiedenen Zuweisungsprogrammen in verschiedenen älteren Studien

| Vergleich | Zhang [161] | Martin [150] |
|--------------|-------------|--------------|
| DSSP/Stride | 95,4% | 95,4% |
| DSSP/KAKSI | 83,6% | 82,1% |
| DSSP/P-SEA | 82,1% | 80,1% |
| DSSP/Autor | | 90,8% |
| Stride/KAKSI | 85,0% | 83,5% |
| Stride/P-SEA | 83,3% | 81,8% |
| Stride/Autor | | 89,9% |
| KAKSI/P-SEA | 82,8% | 81,5% |
| KAKSI/Autor | | 83,4% |
| P-SEA/Autor | | 78,1% |

Unterschiede zwischen Helix und Strang-Zuweisungen

Strukturbiologische Parameter

In den meisten Fällen unterscheiden sich die Zuweisungsergebnisse vor allem im Bereich von β -Strängen und ungeordneten Regionen. So konnten Labesse *et al.* [151] bei der Etablierung ihrer Zuweisungsprogramme zwar eine Übereinstimmung von mehr als 90% zu schon bestehenden Programmen nachweisen, bei den β -Strängen lag die Übereinstimmung jedoch in allen Fällen nur noch bei knapp 80%. In Tabelle 30 sind die Ergebnisse der vorliegenden Arbeit denen von Cao *et al.* [164] gegenübergestellt. Auch hier zeigt sich, dass die Übereinstimmungen bei den Helices deutlich höher sind als bei den β -Strängen. Dies resultiert zum einen aus den sehr gut untersuchten geometrischen Merkmalen und Wasserstoffbrückenmustern von Helices [169] die eine klare Abgrenzung von ungeordneten Bereichen ermöglichen. Zum anderen ist vor allem die Endpunktbestimmung von β -Strängen sehr kompliziert, da sich die geometrischen Parameter nur wenig von ungeordneten Regionen unterscheiden [170]. Außerdem benötigen β -Stränge immer mindestens einen Partner-Strang um eine stabile Konformation anzunehmen [169] und die Suche nach diesen Partnern ist für manche Programme schwer durchführbar.

Tabelle 30: Übereinstimmung von Zuweisungen einzelner Sekundärstrukturmerkmale

Die Daten geben den prozentualen Anteil der Übereinstimmung von Helix- und Strang-Zuweisungen beim Vergleich der einzelnen Programme an. Bei den Daten für die aktuelle Arbeit wurde jeweils der größere Anteil (siehe Formel (3)) angegeben.

| Vergleich | Cao [164] | | Aktuelle Arbeit | |
|----------------|-----------|--------|-----------------|--------|
| | Helix | Strang | Helix | Strang |
| DSSP/Stride | 93,7% | 93,0% | 92,8% | 88,6% |
| DSSP/P-SEA | 86,0% | 78,0% | 70,6% | 54,3% |
| DSSP/KAKSI | 88,4% | 88,0% | 70,1% | 56,6% |
| Stride/P-SEA | 86,2% | 78,6% | 71,3% | 53,5% |
| Stride/KAKSI | 88,0% | 87,3% | 69,1% | 55,3% |
| P-SEA/KAKSI | 95,7% | 83,2% | 96,8% | 81,0% |
| DSSP/VAProTS | | | 69,9% | 57,0% |
| Stride/VAProTS | | | 70,6% | 54,3% |
| P-SEA/VAProTS | | | 97,9% | 89,3% |
| KAKSI/VAProTS | | | 95,5% | 87,0% |
| Autor/VAProTS | | | 90,4% | 73,4% |

Alles in allem zeigen die hier vorliegenden Ergebnisse, dass der neu entwickelte Mechanismus im Bereich der Helices und β -Stränge gute bis sehr gute Übereinstimmungen mit den bereits etablierten Programmen P-SEA und KAKSI aufweist. Allerdings sorgt die Neu-Definition der *Turns* (siehe Abschnitt 2.3.5), sowie die Einführung von klaren Mindestlängen

von fünf Aminosäuren für Helices bzw. drei Aminosäuren für β -Stränge für einige Abweichungen. Des Weiteren könnte die von dem neu entwickelten Mechanismus genutzte Zusammenfassung von verschiedenen Helix-Arten (α -Helix, π -Helix und 3_{10} -Helix, siehe nächster Abschnitt) zu einem Sekundärstrukturmerkmal die Daten beeinflussen.

2.3.3 Helices in Proteinstrukturen

Helices allgemein

Helikale Strukturen sind ein immer wiederkehrendes Merkmal in fast allen Proteinstrukturen und insbesondere die α -Helix als essentielles Sekundärstrukturmerkmal ist dabei nahezu ubiquitär vertreten [5]. In der Literatur [171] wird zwischen α -, 3_{10} -, π -, Ω - und Poly-Prolin-Helices unterschieden. Während die α -Helix hinsichtlich ihrer geometrischen Parameter und ihres Wasserstoffbrückenmusters sehr gut definiert ist [171], ist dies für andere Helices, wie Poly-Prolin-Helices [172], π -Helices [173] oder 3_{10} -Helices [174], nicht immer der Fall. Die Kristallstrukturen der genutzten Datensätze enthalten, ausgehend von den Angaben der Autoren, im Bezug auf alle gefundenen helikalen Strukturen einen Anteil von knapp 30% 3_{10} - und rund 70% α -Helices. Die meisten anderen Sonderformen (Poly-Prolin-Helix oder Ω -Helix) sind nur sehr selten bis gar nicht in den untersuchten Strukturen auffindbar (Tabelle 31).

Tabelle 31: Vorkommen verschiedener Helix-Arten innerhalb der Struktur-Datensätze

Die Tabelle fasst alle in den *pdb-files* der einzelnen Datensätze (Abschnitt 6.1) erwähnten Helices zusammen.

| Helixart | XRD | TRD | TED |
|-------------------|---------------|--------------|-------------|
| α -Helix | 60595 (71,9%) | 6729 (70,3%) | 768 (71,3%) |
| 3_{10} -Helix | 23622 (28,1%) | 2849 (29,7%) | 309 (28,7%) |
| Poly-Prolin-Helix | 2 (0,0%) | 0 | 0 |

Ein Vergleich verschiedener geometrischer Parameter von α - und 3_{10} -Helices (Tabelle 32) verdeutlicht, dass sich beide Helices in vielen Bereichen sehr stark ähneln und eine Differenzierung mit den genutzten Parametern nur schwer durchführbar ist. Zu einem ähnlichen Ergebnis kamen auch Crisma *et al.* [171], die feststellten, dass 3_{10} - und α -Helices im Bereich der Torsionswinkel nahezu identisch sind. Aus diesem Grund wurden sie vom neu entwickelten Zuweisungsalgorithmus (siehe Abschnitt 6.3) zu einer Sekundärstruktur zusammengefasst.

Tabelle 32: Geometrische Parameter von α - und 3_{10} -Helices

| Merkmal | α -Helix | 3_{10} -Helix |
|--|---------------------------------------|---------------------------------------|
| Phi | $-64,9^\circ \pm 11,2^\circ$ | $-66,5^\circ \pm 32,7^\circ$ |
| Psi | $-39,5^\circ \pm 10,9^\circ$ | $-15,0^\circ \pm 30,1^\circ$ |
| Tau [175] | $92,3^\circ \pm 4,0^\circ$ | $92,3^\circ \pm 5,9^\circ$ |
| $\text{C}\alpha_i/\text{C}\alpha_{i+3}$ -Abstand | $5,2 \text{ \AA} \pm 0,3 \text{ \AA}$ | $5,7 \text{ \AA} \pm 0,6 \text{ \AA}$ |
| $\text{C}\alpha_i/\text{C}\alpha_{i-3}$ -Abstand | $5,2 \text{ \AA} \pm 0,3 \text{ \AA}$ | $5,7 \text{ \AA} \pm 0,6 \text{ \AA}$ |

Allgemeine statistische Analyse von Helices

Chou und Fasman beschäftigten sich in mehreren Studien mit statistischen Sekundärstruktur-Präferenzen von Aminosäuren [176–179]. Sie berechneten Helix- und Strang-Präferenzen anhand eines Datensatzes von 15 Proteinstrukturen, deren Sekundärstrukturen sie aus der Originalliteratur entnahmen [176]. Dabei fassten sie verschiedene Helix-Arten (α_I -, α_{II} -, 3_{10} - und verdrehte Helices) zu einer Sekundärstruktur zusammen und bestimmten fehlende β -Stränge anhand von Wasserstoffbrücken-Schemata der Autoren [176]. Aufgrund der geringen Datenmenge, die Chou und Fasman bei ihrer Analyse zur Verfügung stand [176], wurden in jüngerer Vergangenheit neuere Studien angefertigt.

Zu diesen neueren Studien gehört eine Untersuchung von Costantini *et al.*, die sich mit dem Einfluss von Strukturklassen auf Aminosäure-Präferenzen beschäftigt haben [180]. Sie nutzten dabei einen Datensatz von 2168 Strukturen aus der PDB mit einer maximalen Sequenzidentität von 25% [180]. Die Zuweisung der Sekundärstrukturen erfolgte dann mit Hilfe von DSSP [145] wobei sie α -, π - und 3_{10} -Helices sowie β -Stränge und einzelne β -Brücken zu Helices bzw. Strängen zusammenfassten [180]. Als Definition von verschiedenen Strukturklassen kamen in ihrer Arbeit zum einen die Definition nach Nakashima [181] und zum anderen die Definition nach Chou [182] zum Einsatz [180].

Einen ähnlichen Ansatz verfolgten Fujiwara *et al.*, die den Einfluss von Faltungstypen auf die Aminosäureverteilung untersuchten [183]. Ihr Datensatz bestand aus 2580 Proteinstrukturen in 51 verschiedenen Faltungsklassen mit einer maximalen Sequenzidentität von 60% [183]. Die Faltungsklassen bestimmten sie mit Hilfe ihrer OLIGAMI-Datenbank [184], die Informationen der SCOP-Datenbank [185] mit Daten zur Oligomerisierung von Proteinen kombiniert [183]. Auch hier kam DSSP [145] zur Bestimmung von Sekundärstrukturen zum Einsatz [183]. Im Gegensatz zur Arbeit von Costantini *et al.*

[180] wurden jedoch nur α - und 3_{10} -Helices zusammengefasst und β -Brücken als *Coil* angesehen [183].

Tabelle 33: Statistische Auswertung von Helix-Präferenzen der einzelnen Aminosäuren für den Datensatz XRD

Die Werte zeigen die statistischen Helix-Präferenzen der Aminosäuren im Bezug auf den gesamten Datensatz. Alle Werte wurden anhand von Formel (15) (siehe Abschnitt 6.4 auf Seite 112) berechnet und Werte größer 1,3 und kleiner 0,7 können als statistisch signifikant angesehen werden.

| Aminosäure | Chou [177] | Costantini [180] | Fujiwara [183] | XRD |
|--------------|------------|------------------|----------------|------|
| Alanin | 1,45 | 1,39 | 1,41 | 1,41 |
| Arginin | 0,79 | 1,17 | 1,21 | 1,22 |
| Asparagin | 0,73 | 0,77 | 0,73 | 0,77 |
| Aspartat | 0,98 | 0,89 | 0,82 | 0,82 |
| Cystein | 0,77 | 0,74 | 0,85 | 0,80 |
| Glutamin | 1,17 | 1,29 | 1,39 | 1,31 |
| Glutamat | 1,53 | 1,35 | 1,26 | 1,38 |
| Glycin | 0,53 | 0,47 | 0,44 | 0,44 |
| Histidin | 1,24 | 0,92 | 0,87 | 0,89 |
| Isoleucin | 1,00 | 1,04 | 1,04 | 1,03 |
| Leucin | 1,34 | 1,32 | 1,28 | 1,31 |
| Lysin | 1,07 | 1,11 | 1,17 | 1,19 |
| Methionin | 1,20 | 1,21 | 1,26 | 1,30 |
| Prolin | 0,59 | 0,50 | 0,44 | 0,49 |
| Phenylalanin | 1,12 | 1,01 | 1,00 | 1,00 |
| Serin | 0,79 | 0,82 | 0,76 | 0,80 |
| Threonin | 0,82 | 0,76 | 0,78 | 0,75 |
| Tryptophan | 1,14 | 1,06 | 1,07 | 1,01 |
| Tyrosin | 0,61 | 0,95 | 0,98 | 0,96 |
| Valin | 1,14 | 0,89 | 0,91 | 0,86 |

Die Daten in Tabelle 33 zeigen große Übereinstimmungen mit den Arbeiten von Costantini *et al.* [180] und Fujiwara *et al.* [183], weichen jedoch für manche Aminosäuren deutlich von den Ergebnissen von Chou und Fasman ab [176]. Diese Abweichungen resultieren aus der unterschiedlichen Aminosäurezusammensetzung des Datensatzes von Chou und Fasman im Vergleich zum Datensatz von Costantini *et al.* [180] und dem hier genutzten (Tabelle A.3).

Vor allem lange, polare Aminosäuren (Glu, Gln, Arg und Lys) sowie unpolare Aminosäuren, die weder aromatisch sind noch am $C\beta$ -Atom eine Verzweigung tragen (Ala, Leu und Met) kommen in Helices vor. Eine Studie von Malkov *et al.* der Abhängigkeit von Sekundärstruktur-Präferenzen von der chemischen Struktur von Aminosäuren kam zu einem

Strukturbiologische Parameter

identischen Ergebnis [186]. Anhand von 1737 Proteinstrukturen mit einer Sequenzidentität von maximal 25% konnten sie zeigen, dass Alanin, Leucin, Glutamin, Glutamat, Arginin, Methionin und Lysin in Helices präferiert vorkommen [186].

Die Präferenz von langen, polaren Aminosäuren in Helices könnte auf die Ausbildung von Salzbrücken zwischen verschiedenen Helices oder innerhalb einer Helix hindeuten. Eine Auswertung aller im Datensatz XRD vorhandenen Salzbrücken unter Einbeziehung der Sekundärstruktur (Tabelle 34) konnte zeigen, dass die meisten Salzbrücken tatsächlich von Aminosäuren in Helices gebildet werden. Hinzu kommt, dass mehr als zwei Drittel dieser Salzbrücken zwischen zwei Aminosäuren in Helices ausgebildet werden.

Tabelle 34: Salzbrücken zwischen verschiedenen Sekundärstrukturen

Die Tabelle fasst alle gefundenen Salzbrücken anhand der Sekundärstruktur der jeweiligen Partner zusammen. Die Zeilen tragen die Sekundärstruktur von Glu oder Asp und die Spalten die Sekundärstruktur von Arg, His oder Lys. Die Prozente geben den Anteil der jeweiligen Kombination an allen Werten der entsprechenden Spalte an. In der letzten Spalte ist die Anzahl von Salzbrücken innerhalb eines Sekundärstrukturmerkmals, z.B. einer Helix, angegeben.

| | Helix | Strang | Turn | Coil | Selbe SSE |
|---------------|------------------|-----------------|----------------|------------------|------------------|
| Helix | 24824 (70,1%) | 2491 (13,9%) | 461 (32,5%) | 7125 (28,5%) | 14113 |
| Strang | 1815 (5,1%) | 8433 (47,0%) | 287 (20,3%) | 3896 (15,6%) | 1521 |
| Turn | 755 (2,1%) | 809 (4,5%) | 206 (14,5%) | 830 (3,3%) | 142 |
| Coil | 8033 (22,7%) | 6204 (34,6%) | 463 (32,7%) | 13174 (52,6%) | X |
| Gesamt | 35427 | 17937 | 1417 | 25025 | 15776 |

Das hohe Vorkommen von Aminosäuren, die unverzweigt am C β -Atom sind, spricht für sterische Probleme bei der Ausbildung der Helix wenn eine solche Verzweigung existiert. Dafür sprechen auch die geringen Werte von Threonin (0,75) und Valin (0,87).

Als Helixbrecher konnten in der vorliegenden Arbeit nur Glycin und Prolin identifiziert werden. Beide Aminosäuren sind in der Lage, abweichende geometrische Konformationen im Vergleich zu den restlichen Aminosäuren anzunehmen und unterbrechen so die regelmäßige Struktur der Helix.

Positionsspezifische Aminosäurevorkommen

Verschiedene Studien beschäftigten sich mit positionsspezifischen Präferenzen von Aminosäuren in Helices. Eine der Ersten wurde von Richardson und Richardson anhand eines Datensatzes von 215 Helices aus 45 Proteinstrukturen angefertigt [187]. Sie untersuchten dabei Helices, die eine Mindestlänge von 8 Aminosäuren aufwiesen und definierten das N- bzw. C-Cap als die erste bzw. letzte Aminosäure, die noch im Zylinder-Bereich der Helix und auf der Spiralbahn der Proteinkette lagen [187]. Auch Aurora und Rose beschäftigten sich in einer statistischen Analyse von 1316 Helices mit mindestens sieben Aminosäuren aus 274 Proteinstrukturen mit den N- und C-Caps [188]. Sie nutzten dabei eine andere Definition und sahen die erste bzw. letzte Aminosäure, die eine $i/i+4$ Wasserstoffbrücke ausbildet, als Cap-Aminosäure an [188]. Ein Vergleich der Ergebnisse von Richardson und Richardson [187] sowie Aurora und Rose [188] mit den Daten der vorliegenden Arbeit (Tabelle 35) zeigt, dass ein Vergleich der erhaltenen Ergebnisse aufgrund der unterschiedlichen Definitionen, Datensatzgrößen und Zuweisungsgrundlagen kaum möglich ist.

Da eine Einordnung der erhaltenen Daten im Bezug auf ältere Studien schwer machbar ist, sollen im Folgenden die einzelnen Positionen innerhalb von Helices separat diskutiert werden (Tabelle 35). Lediglich zwei Aminosäuren zeigen ein relativ positionsunabhängiges Verhalten. Zum einen findet man Glycin fast ausschließlich direkt vor oder nach Helices (Positionen -1 und $n+1$) und zum anderen findet man Prolin fast ausschließlich an der ersten Position von Helices.

An der Position direkt vor der Helix (-1) werden neben Glycin vor allem kurze, polare Aminosäuren bevorzugt. Die Fähigkeit von Serin, Threonin, Asparagin und Aspartat, mit ihren Seitenketten Wasserstoffbrücken zum Proteinerückgrat der Helix auszubilden [187], führt zu dieser Präferenz. Vasudev *et al.* konnten bei einer Analyse von Seitenketten-Wasserstoffbrücken von Asparagin und Glutamin innerhalb von Proteinstrukturen zeigen, dass Asparagin vor allem Wasserstoffbrücken der Form $i/i+2$ oder $i/i+3$ mit seiner Seitenkette eingeht, während Glutamin vor allem Wasserstoffbrücken der Form $i/i-3$ eingeht [189]. Dementsprechend ist Glutamin weniger in der Lage, eine Helix von dieser Position aus zu stabilisieren und kommt deutlich seltener vor.

Strukturbiologische Parameter

Tabelle 35: Positionsspezifische, statistische Präferenzen von Aminosäuren in Helices

Die Position -1 markiert die Aminosäure vor und die Position n+1 die Aminosäure direkt nach der Helix. Die Positionen n bzw. n-1 repräsentieren die letzten beiden Positionen innerhalb der Helix. Positionen 1 und 2 umfassen die erste und die zweite Aminosäure in der Helix. In M sind alle restlichen Aminosäuren der Helix zusammengefasst. In den Studien von Richardson und Richardson [187] bzw. Aurora und Rose [188] sind die Aminosäuren des N-bzw. C-Caps als Positionen -1 und n+1 eingetragen. Da Aurora und Rose nur die Caps von Helices untersuchten, wurden von ihnen keine Werte für die Mitte der Helix definiert [188]. Als statistisch signifikant wurden Werte $\geq 1,5$ oder $\leq 0,5$ angesehen (**fett markiert**).

| Position | Richardson [187] | | | | | | | Aurora [188] | | | | | | | Diese Arbeit | | | | | | |
|----------|------------------|------------|------------|------------|------------|------------|------------|--------------|-------------|-------------|-------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | -1 | 1 | 2 | M | n-1 | n | n+1 | -1 | 1 | 2 | M | n-1 | n | n+1 | -1 | 1 | 2 | M | n-1 | n | n+1 |
| ALA | 0,5 | 1,2 | 1,6 | 1,8 | 1,4 | 1,1 | 0,8 | 0,67 | 1,10 | 1,39 | n. d. | 1,33 | 1,87 | 1,19 | 0,34 | 1,07 | 1,25 | 1,48 | 1,44 | 1,35 | 0,59 |
| ARG | 0,4 | 0,7 | 0,9 | 1,3 | 2,1 | 1,0 | 0,9 | 0,76 | 1,05 | 0,95 | n. d. | 1,39 | 1,66 | 1,45 | 0,54 | 0,96 | 0,85 | 1,24 | 1,64 | 1,27 | 0,72 |
| ASN | 3,5 | 0,7 | 0,7 | 0,9 | 0,9 | 1,2 | 1,6 | 1,28 | 0,72 | 0,67 | n. d. | 0,64 | 0,70 | 1,33 | 2,30 | 0,51 | 0,89 | 0,70 | 0,89 | 1,28 | 1,89 |
| ASP | 2,1 | 0,8 | 2,6 | 2,2 | 0,7 | 0,4 | 0,7 | 1,58 | 1,14 | 1,64 | n. d. | 0,60 | 0,91 | 0,72 | 2,81 | 0,91 | 1,65 | 0,73 | 0,89 | 0,53 | 1,05 |
| CYS | 0,6 | 0,8 | 1,2 | 0,7 | 1,2 | 1,2 | 1,6 | 0,37 | 0,26 | 0,52 | n. d. | 0,44 | 0,33 | 0,44 | 1,01 | 0,42 | 0,41 | 0,88 | 0,76 | 1,02 | 1,03 |
| GLN | 0,4 | 0,7 | 0,8 | 1,3 | 1,6 | 2,1 | 0,9 | 1,05 | 1,31 | 1,60 | n. d. | 1,37 | 1,24 | 1,43 | 0,62 | 0,85 | 1,22 | 1,32 | 1,55 | 1,52 | 1,00 |
| GLU | 0,4 | 2,2 | 2,0 | 0,8 | 1,7 | 0,8 | 0,3 | 0,94 | 2,30 | 2,07 | n. d. | 1,43 | 1,88 | 1,27 | 0,54 | 1,39 | 2,60 | 1,26 | 1,51 | 0,99 | 0,49 |
| GLY | 1,8 | 0,3 | 0,9 | 0,5 | 0,2 | 0,2 | 3,9 | 0,98 | 0,55 | 0,65 | n. d. | 0,20 | 0,33 | 0,74 | 1,57 | 0,47 | 0,74 | 0,42 | 0,34 | 0,38 | 3,96 |
| HIS | 1,1 | 0,7 | 0,7 | 1,0 | 1,8 | 3,4 | 1,3 | 0,83 | 0,83 | 1,36 | n. d. | 1,02 | 0,89 | 1,55 | 1,11 | 0,83 | 1,15 | 0,80 | 0,92 | 1,33 | 1,56 |
| ILE | 0,2 | 0,9 | 0,7 | 1,2 | 0,4 | 0,7 | 0,7 | 0,78 | 1,06 | 0,64 | n. d. | 1,58 | 0,90 | 0,61 | 0,15 | 0,80 | 0,52 | 1,22 | 0,71 | 0,54 | 0,51 |
| LEU | 0,2 | 0,9 | 0,3 | 1,2 | 0,8 | 0,7 | 0,7 | 0,79 | 0,84 | 0,91 | n. d. | 1,63 | 1,65 | 1,36 | 0,25 | 0,96 | 0,57 | 1,44 | 1,20 | 1,43 | 0,73 |
| LYS | 0,7 | 0,6 | 1,0 | 1,1 | 1,9 | 2,0 | 1,3 | 0,84 | 1,08 | 0,80 | n. d. | 1,71 | 1,63 | 1,45 | 0,46 | 1,09 | 1,19 | 1,09 | 1,83 | 1,41 | 0,89 |
| MET | 0,8 | 0,3 | 1,0 | 1,5 | 1,3 | 1,0 | 0,8 | 0,98 | 0,90 | 1,10 | n. d. | 1,76 | 1,35 | 1,35 | 0,28 | 0,94 | 0,54 | 1,48 | 1,05 | 1,23 | 0,71 |
| PRO | 0,8 | 2,6 | 0,5 | 0,3 | 0,2 | 0,0 | 0,7 | 1,12 | 1,67 | 0,94 | n. d. | 0,07 | 0,03 | 0,10 | 1,46 | 3,21 | 0,52 | 0,23 | 0,24 | 0,03 | 0,06 |
| PHE | 0,2 | 0,5 | 0,9 | 1,3 | 0,3 | 0,7 | 0,5 | 0,96 | 0,90 | 1,00 | n. d. | 1,22 | 0,67 | 1,20 | 0,36 | 0,94 | 0,64 | 1,06 | 0,78 | 1,23 | 0,99 |
| SER | 2,3 | 0,7 | 0,8 | 0,6 | 1,6 | 1,7 | 0,8 | 1,25 | 0,81 | 0,69 | n. d. | 0,42 | 0,71 | 1,02 | 2,75 | 0,86 | 1,16 | 0,68 | 0,93 | 1,20 | 0,84 |
| THR | 1,6 | 0,8 | 0,7 | 1,0 | 0,9 | 1,0 | 0,3 | 1,41 | 0,77 | 0,92 | n. d. | 0,57 | 0,50 | 0,82 | 2,03 | 0,74 | 0,87 | 0,71 | 0,66 | 1,03 | 0,59 |
| TRP | 0,3 | 2,1 | 1,7 | 1,5 | 0,4 | 0,0 | 0,0 | 0,94 | 1,26 | 1,10 | n. d. | 1,00 | 1,00 | 0,58 | 0,49 | 1,31 | 0,97 | 1,07 | 0,64 | 0,67 | 0,48 |
| TYR | 0,8 | 1,8 | 0,4 | 0,8 | 0,3 | 1,2 | 0,8 | 0,82 | 0,99 | 0,73 | n. d. | 1,02 | 0,73 | 1,06 | 0,47 | 0,93 | 0,78 | 0,99 | 0,66 | 1,18 | 0,91 |
| VAL | 0,1 | 1,1 | 0,6 | 1,2 | 0,7 | 0,7 | 0,2 | 0,67 | 0,76 | 0,70 | n. d. | 1,08 | 0,51 | 0,46 | 0,17 | 0,76 | 0,53 | 0,99 | 0,62 | 0,53 | 0,50 |

Strukturbiologische Parameter

Als Helixinitiator findet sich in vielen Fällen Prolin. Da in den meisten Helices die Aminosäuren an Position zwei oder drei als erste Donoren von Wasserstoffbrücken auftreten [190], kann ein Prolin durch seine Imido-Gruppe die Exposition einer unverbrückten Amido-Gruppe verhindern [187]. Als besonders ungünstig erweisen sich kurze polare Aminosäuren am Anfang von Helices.

Die zweite Position von Helices häufig durch vor allem negativ geladene Aminosäuren eingenommen. Das stimmt mit einer Arbeit von Shoemaker *et al.* überein, die mit Hilfe von Titrationsexperimenten am isolierten C-Peptid (Aminosäuren 1 bis 13 von Ribonuklease A) zeigen konnten, dass eine negative Ladung am N-Terminus von Helices zu einer Stabilisierung und Ausbildung eines Dipol-Momentes über die Helix beitragen kann [191]. Große, aliphatische, hydrophobe Aminosäuren sowie schwefelhaltigen Aminosäuren sind ungünstig an dieser Position, da sie möglicherweise die Ausbildung einer Helix aufgrund von sterischen Problemen behindern können.

Im mittleren Bereich von Helices sind fast alle Aminosäuren erlaubt (ausgenommen Glycin und Prolin). Anscheinend muss eine Helix nur einmal initiiert werden und kann dann unabhängig von den eingebauten Aminosäuren verlängert werden. Hovmöller *et al.* konnten zeigen, dass Aminosäuren im Inneren von Helices Phi/Psi-Winkel von $\Phi = -63,8^\circ \pm 2^\circ$ und $\Psi = -41,1^\circ \pm 2^\circ$ aufweisen [192]. Dieser sehr enge Bereich spricht dafür, dass Helices nach ihrer Initiierung relativ unabhängig von äußeren Einflüssen sind und nur durch die Einbindung von Glycin oder Prolin durchbrochen werden können.

An der vorletzten und letzten Position von Helices sind vor allem positiv geladene Aminosäuren günstig. So kann ein Ausgleich der negativen Ladung am N-Terminus der Helices erreicht werden. Leicht problematisch scheinen aromatische Aminosäuren sowie Aminosäuren mit einer Verzweigung am C β -Atom zu sein, die eventuell zu sterischen Problemen führen könnten.

Die hohe Präferenz von Glycin an der ersten Position direkt nach der Helix (n+1) resultiert aus seinen starken helixbrechenden Eigenschaften. Richardson und Richardson diskutierten, dass Glycin durch seine L α - oder L $_{310}$ -Konformation in der Lage ist, die Proteinkette sehr stark zu drehen, so dass die Peptidbindungen vor und nach dem Glycin Wasserstoffbrücken zu den Aminosäuren an viertletzter und drittletzter Position ausbilden können [187]. Dadurch würde das regelmäßige Wasserstoffbrückennetzwerk der Helix unterbrochen und eine Fortsetzung wäre nicht mehr möglich [187].

2.3.4 β -Stränge

Statistische Strang-Präferenzen

Analog zur statistischen Auswertung von Helices wurden auch die gefundenen β -Stränge detailliert ausgewertet. Ein Vergleich der erhaltenen Daten mit den Ergebnissen der bereits publizierten Studien von Chou und Fasman [176], Costantini *et al.* [180] sowie Fujiwara *et al.* [183] führte zu relativ ähnlichen Daten (Tabelle 36). So sind neben Cystein vor allem aromatische sowie am C β -Atom verzweigte Aminosäuren bevorzugt in Strängen zu finden. Street und May untersuchten dieses Phänomen mit Hilfe von elektrostatischen Berechnungen von Dipeptiden [193]. Unter Zuhilfenahme der β -Raum Definition von Phi/Psi-Winkeln nach Munoz und Serrano [194] sowie den elektrostatischen Berechnungsansätzen von Gastgeiger und Marsili [195] sowie Rappe und Goddard [196] berechneten sie Entropieänderungen bei Ausbildung von β -Strängen [193]. Sie begründeten das hohe Vorkommen von aromatischen und am C β -Atom verzweigten Aminosäuren damit, dass so sterische Probleme zwischen der Seitenkette und dem Proteinrückgrat vermieden werden können [193].

Tabelle 36: Statische Auswertung von Aminosäure-Präferenzen in β -Strängen

| Aminosäure | Chou [176] | Costantini [180] | Fujiwara [183] | XRD |
|--------------|-------------|------------------|----------------|-------------|
| Alanin | 0,97 | 0,75 | 0,72 | 0,84 |
| Arginin | 0,90 | 0,91 | 0,85 | 0,95 |
| Asparagin | 0,65 | 0,62 | 0,63 | 0,63 |
| Aspartat | 0,80 | 0,55 | 0,63 | 0,59 |
| Cystein | 1,30 | 1,31 | 1,36 | 1,43 |
| Glutamin | 1,23 | 0,76 | 0,72 | 0,76 |
| Glutamat | 0,26 | 0,72 | 0,65 | 0,69 |
| Glycin | 0,81 | 0,65 | 0,67 | 0,60 |
| Histidin | 0,71 | 0,99 | 0,99 | 1,00 |
| Isoleucin | 1,60 | 1,71 | 1,79 | 1,56 |
| Leucin | 1,22 | 1,10 | 1,15 | 1,06 |
| Lysin | 0,74 | 0,83 | 0,76 | 0,80 |
| Methionin | 1,67 | 0,99 | 1,01 | 1,05 |
| Prolin | 0,62 | 0,44 | 0,40 | 0,61 |
| Phenylalanin | 1,28 | 1,43 | 1,40 | 1,41 |
| Serin | 0,72 | 0,85 | 0,81 | 0,95 |
| Threonin | 1,20 | 1,23 | 1,21 | 1,26 |
| Tryptophan | 1,19 | 1,30 | 1,23 | 1,29 |
| Tyrosin | 1,29 | 1,30 | 1,37 | 1,43 |
| Valin | 1,65 | 1,86 | 2,00 | 1,78 |

Polare, nicht basische Aminosäure sowie Glycin und Prolin sind relativ selten in Strängen zu finden. Glycin und Prolin brechen sehr wahrscheinlich die regelmäßige Struktur des Proteinrückgrates und fungieren so als Terminator von Strängen. Asparagin, Aspartat, Glutamin und Glutamat sind wahrscheinlich energetisch ungünstig innerhalb der unpolaren Stränge. Durch die hohe Präferenz von aromatischen Aminosäuren könnte es zu einer energetischen Abstoßung zwischen den π -Elektronensystemen der Aromaten und den negativen Seitenketten von Asn, Asp, Gln sowie Glu kommen.

Positionsspezifische Strang-Präferenzen

West und Hecht konnten mit Hilfe einer statistischen Analyse von 197 Kristallstrukturen zeigen, dass β -Stränge vor allem hydrophobe Aminosäuren enthalten [197]. Ihre Auswertung von binären Pentapeptidmustern (Abfolge von polaren und unpolaren Aminosäuren) ergab, dass unpolare Aminosäuren innerhalb von Strängen oftmals von polaren Aminosäuren eingerahmt werden [197]. Zu einem anderen Schluss kamen Mandel-Gutfreund und Gregoret, die für einen Datensatz von 1911 Kristallstrukturen ein hohes Vorkommen ($\sim 10\%$ aller Stränge) von periodischen Pentapeptiden (polar/nicht-polar/polar/nicht-polar/polar oder umgekehrt) feststellten [198]. Allerdings zeigen ihre Ergebnisse auch, dass $\sim 20\%$ der gefundenen Pentapeptide entweder nur aus nicht-polaren Aminosäuren oder aus nicht-polaren Aminosäuren mit polaren Aminosäuren an den Enden bestehen (Abb. A. 8).

Die Auswertung von positionsspezifischen Präferenzen für β -Stränge mit mindestens neun Aminosäuren (Tabelle 37) konnte die Daten von West und Hecht [197] sowie Mandel-Gutfreund und Gregoret [198] bestätigen. So findet man direkt vor oder nach Strängen kurze polare Aminosäuren sowie Glycin und Prolin. Letztere können durch ihre einzigartige Rückgrat-Geometrie das regelmäßige Muster von β -Strängen durchbrechen. Asparagin und Aspartat hingegen könnten, wie schon bei den Helices diskutiert, an der Ausbildung von Wasserstoffbrücken zwischen ihrer Seitenkette und dem Proteinrückgrat innerhalb des Stranges beteiligt sein. Bestätigt wird diese Vermutung durch das hohe Auftreten von Threonin und Serin an der ersten Position innerhalb der Stränge. Auch diese Aminosäuren sind in der Lage, Wasserstoffbrücken mit ihrer Seitenkette zu bilden. Eventuell dienen diese Wasserstoffbrücken als Alternative zu Wasserstoffbrücken zwischen dem Proteinrückgrat von zwei benachbarten Strängen.

Strukturbiologische Parameter

Neben Serin und Threonin sind vor allem positiv geladene Aminosäuren an der ersten Position im Strang bevorzugt. Im Gegensatz dazu findet man an der letzten Position vor allem kurze polare Aminosäuren (Asn, Asp und Ser). Durch diese Präferenzen kommt es auch hier zu einer Ausbildung eines Dipols über das gesamte Sekundärstrukturelement (SSE).

Tabelle 37: Statistische Auswertung von positionsspezifischen Strang-Präferenzen von β -Strängen mit mindestens neun Aminosäuren für den Datensatz XRD

| Aminosäure | -1 | 1 | 2 | 3 | 4 | M | n-3 | n-2 | n-1 | n | n+1 |
|--------------|-------------|------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Alanin | 0,48 | 0,84 | 0,66 | 0,78 | 0,85 | 0,91 | 0,85 | 0,94 | 0,73 | 0,73 | 0,86 |
| Arginin | 0,91 | 1,31 | 1,17 | 1,09 | 0,93 | 0,88 | 1,00 | 1,07 | 1,14 | 1,01 | 0,71 |
| Asparagin | 1,91 | 0,81 | 0,67 | 0,62 | 0,85 | 0,66 | 0,55 | 0,68 | 0,47 | 1,69 | 1,78 |
| Aspartat | 1,42 | 0,83 | 0,48 | 0,50 | 0,45 | 0,52 | 0,56 | 0,36 | 0,44 | 1,53 | 1,66 |
| Cystein | 0,34 | 0,84 | 1,00 | 1,15 | 1,67 | 0,95 | 1,22 | 1,11 | 1,01 | 1,11 | 0,38 |
| Glutamin | 0,92 | 1,52 | 0,80 | 0,85 | 0,69 | 0,84 | 0,75 | 0,69 | 0,86 | 0,98 | 0,73 |
| Glutamat | 0,81 | 1,05 | 0,72 | 0,69 | 0,72 | 0,73 | 0,62 | 0,69 | 1,00 | 0,73 | 1,07 |
| Glycin | 3,83 | 0,82 | 0,51 | 0,61 | 0,58 | 0,72 | 0,59 | 0,66 | 0,70 | 0,71 | 1,98 |
| Histidin | 1,06 | 1,30 | 1,08 | 1,19 | 0,97 | 0,80 | 0,84 | 0,98 | 1,05 | 1,02 | 0,73 |
| Isoleucin | 0,42 | 0,62 | 1,27 | 1,44 | 1,58 | 1,45 | 1,98 | 1,62 | 1,32 | 0,78 | 0,45 |
| Leucin | 0,39 | 0,59 | 0,99 | 1,00 | 1,18 | 1,07 | 1,08 | 1,12 | 0,85 | 0,60 | 0,51 |
| Lysin | 1,05 | 1,40 | 1,03 | 0,72 | 0,70 | 0,73 | 0,82 | 0,78 | 1,06 | 0,93 | 1,19 |
| Methionin | 0,65 | 1,00 | 1,43 | 0,97 | 1,04 | 1,01 | 1,04 | 0,89 | 0,84 | 0,55 | 0,69 |
| Prolin | 0,78 | 0,86 | 0,97 | 0,37 | 0,36 | 0,40 | 0,37 | 0,48 | 0,67 | 1,72 | 2,22 |
| Phenylalanin | 0,36 | 0,96 | 1,14 | 1,28 | 1,42 | 1,59 | 1,58 | 1,62 | 1,47 | 0,89 | 0,45 |
| Serin | 0,86 | 1,18 | 1,08 | 0,83 | 0,85 | 1,09 | 0,74 | 0,78 | 0,82 | 1,31 | 1,07 |
| Threonin | 1,05 | 1,56 | 1,47 | 1,25 | 1,32 | 1,37 | 1,40 | 1,25 | 1,46 | 1,19 | 1,03 |
| Tryptophan | 0,35 | 1,14 | 1,52 | 0,96 | 1,15 | 1,06 | 1,11 | 1,31 | 1,20 | 0,76 | 0,36 |
| Tyrosin | 0,51 | 1,10 | 1,58 | 2,06 | 1,43 | 1,48 | 1,31 | 1,45 | 1,54 | 0,91 | 0,42 |
| Valin | 0,44 | 0,93 | 1,52 | 2,11 | 1,93 | 1,72 | 1,84 | 1,84 | 1,74 | 1,08 | 0,44 |

Im Inneren der Stränge (Positionen 2 bis n-1) sind häufig vor allem aromatische und am C β -Atom verzweigte, unpolare Aminosäuren zu finden. Stränge können mit Hilfe der häufig zu findenden aromatischen Aminosäuren durch π - π - oder CH- π -Wechselwirkungen zwischen den Aminosäuren eines Stranges stabilisiert werden. Eine detaillierte Analyse von π - π -Wechselwirkungen unter Beachtung der SSEs konnte einen vergleichsweise hohen Anteil (\sim 30%) von Interaktionen mit mindestens einem Partner in Strängen nachweisen (Tabelle 38).

Tabelle 38: π - π -Wechselwirkungen zwischen verschiedenen Sekundärstrukturen

Die Tabelle fasst alle gefundenen π - π -Wechselwirkungen anhand der Sekundärstruktur der jeweiligen Partner zusammen. Die Prozente geben den Anteil der jeweiligen Kombination an allen Werten der entsprechenden Spalte an. In der letzten Spalte ist die Anzahl von π - π -Interaktionen innerhalb eines Sekundärstrukturmerkmals, z.B. einer Helix, angegeben.

| | Helix | Strang | Turn | Coil | Innerhalb einer SSE |
|--------|------------------|------------------|----------------|------------------|---------------------|
| Helix | 24735 (60,5%) | 6172 (18,1%) | 360 (37,4%) | 10553 (30,3%) | 10471 |
| Strang | 5145 (12,6%) | 19337 (56,6%) | 290 (30,1%) | 8480 (24,4%) | 3254 |
| Turn | 568 (1,4%) | 316 (0,9%) | 64 (6,6%) | 288 (0,8%) | 56 |
| Coil | 10443 (25,5%) | 8342 (24,4%) | 249 (25,9%) | 15478 (44,5%) | X |
| Gesamt | 40891 | 34167 | 963 | 34799 | 15776 |

Ungünstig sind dagegen Aspartat, Glutamat, Prolin und Glycin. Die negativ geladenen Aminosäuren könnten zu einer elektrostatischen Abstoßung mit den π -Elektronensystemen der aromatischen Aminosäuren führen. Glycin und Prolin würden wahrscheinlich die Ausbildung einer regelmäßigen Proteinerückgratstruktur verhindern und kommen deshalb nur selten in Strängen vor.

Die Auswertung von Daten für kürzere Stränge (Tabelle A.4) ergab keinen gravierenden Unterschied im Bezug auf die grundsätzliche Aminosäure-Verteilung. Auffällig ist jedoch, dass bei Strängen mit sechs oder weniger Aminosäuren die Ausbildung eines Dipols nicht so stark ausgeprägt ist wie es bei den längeren Strängen der Fall ist. Insbesondere die Präferenz von Lysin an Position 1 innerhalb der Stränge findet sich erst bei Strängen mit mehr als 6 Aminosäuren.

2.3.5 Verfeinerung von *tight Turns*

Grundsätzliche Definition und Arten von tight Turns

Neben α -Helices und β -Strängen gibt es noch weitere wichtige Sekundärstrukturmerkmale in Form von sogenannten *tight Turns* [199]. Per Definition handelt es sich dabei um Regionen mit maximal 6 Aminosäuren innerhalb einer Proteinstruktur, in deren Verlauf sich die Proteinkette um 180° wendet und der Abstand zwischen den $C\alpha$ -Atomen der ersten sowie letzten Aminosäure kleiner als 7\AA ist [200]. Ausgehend von dieser Definition wurden im

Laufe der Jahre verschiedene Typen in Abhängigkeit der enthaltenen Anzahl an Aminosäuren identifiziert. Am häufigsten kommt dabei der β -Turn mit 4 Aminosäuren vor [159], doch auch andere Varianten mit 2 (δ -Turn [158]), 3 (γ -Turn [201]), 5 (α -Turn [202]) oder 6 Aminosäuren (π -Turn [203]) sind möglich. Zusätzlich wurden in den letzten Jahren auch noch einige Sonderformen wie der ω -Turn [204] oder der ε -Turn [205] beschrieben.

Eine Analyse von *Turns*, die mit Hilfe von DSSP und Stride identifiziert wurden, verdeutlichte, dass diese oftmals in ungeordneten Regionen an der Oberfläche von Proteinstrukturen vorkommen. Bei diesen Regionen handelt es sich jedoch um sehr flexible Bereiche, die sich in dynamischen Prozessen stark bewegen und in vielen Fällen ihre *Turn*-Struktur nicht aufrecht erhalten können. Dementsprechend wurde im Rahmen dieser Arbeit versucht, die Definition von *tight Turns* zu verfeinern, um Regionen zu identifizieren, die essentiell für die Faltung des Proteins sind und die im Verlauf von dynamischen Prozessen stabil bleiben.

Betrachtung von Regionen zwischen zwei SSEs

Thornton *et al.* konnten bei einer Analyse von *Loop*-Regionen zwischen zwei Sekundärstrukturmerkmalen Regelmäßigkeiten bei den Phi/Psi-Winkeln feststellen [206]. Die vorliegende Arbeit erweitert diesen Ansatz, indem die klassische Definition von *tight Turns* auf Regionen zwischen Helices und β -Strängen mit maximal 6 Aminosäuren angewandt wurde. So konnte eine Vielzahl von möglichen *Turn*-Kandidaten gefunden werden (Tabelle 39). Die Nomenklatur der hier beschriebenen *Turns* folgt dabei einfachen Regeln. Die Typbezeichnung setzt sich zusammen aus dem SSE vor, dem SSE nach und der Anzahl an Aminosäuren im *Turn*. Ein *Turn* zwischen einer Helix und einem β -Strang mit 3 Aminosäuren, würde dann die Typbezeichnung HE3 zugewiesen bekommen. Ausgehend von dieser Nomenklatur können Hetero-Typen (HE oder EH) sowie Homo-Typen (HH oder EE) definiert werden.

Tabelle 39: Zusammenfassung der hier definierten *Turn*-Typen

In der Tabelle sind die Anzahlen aller möglichen *Turns* im Datensatz XRD zwischen zwei SSEs aufgeführt. In der ersten Spalte sind die verfeinerten *Turns* mit den am ehesten vergleichbaren bereits bekannten *Turn*-Typen aufgeführt (siehe Text für die Diskussion). In den darauffolgenden Spalten sind die allgemeine Anzahl sowie die Anzahl an Regionen, bei denen die C α -Atome der ersten und letzten Aminosäure maximal 7Å voneinander entfernt sind, aufgeführt. In Spalte drei wird die Anzahl an Regionen, für die sowohl der C α -Abstand als auch eine Drehung der Proteinkette um mindestens 120° erfüllt ist, gezeigt. Die vierte Spalte gibt die Anzahl der Regionen an, die zusätzlich das Phi/Psi-Kriterium (Abb. 21) erfüllen. Am Ende der Tabelle wird die mittlere Rotation der Proteinkette im Verlauf der untersuchten Regionen gezeigt.

| <i>Turn</i> -Typ | Anzahl | C α -Abstand | Rotation | Phi/Psi-Cluster | Mittlere Rotation |
|------------------|--------|---------------------|----------|-----------------|-------------------|
| HH1 | 3712 | 3712 | 1368 | 1194 | 146,7° |
| HH2 (β) | 2468 | 2468 | 1796 | 1671 | 146,6° |
| HH3 (α) | 2397 | 2059 | 804 | 688 | 140,5° |
| HE1 | 2937 | 2937 | 1634 | 1486 | 150,3° |
| HE2 (β) | 3858 | 3858 | 2470 | 2270 | 152,2° |
| HE3 (α) | 3527 | 3278 | 2155 | 1975 | 148,0° |
| EH1 | 3320 | 3320 | 1301 | 1147 | 144,7° |
| EH2 (β) | 2450 | 2450 | 1153 | 987 | 143,1° |
| EH3 (α) | 2128 | 1935 | 1000 | 828 | 144,6° |
| EE1 | 6603 | 6603 | 2287 | 1979 | 147,7° |
| EE2 (β) | 7090 | 7090 | 4647 | 4237 | 147,6° |
| EE3 (α) | 9660 | 9104 | 6909 | 6343 | 146,8° |
| EE4 (π) | 4213 | 2049 | 1623 | 1434 | 149,9° |

Die Daten in Tabelle 39 zeigen, wie viele Kandidaten im Allgemeinen gefunden wurden (Spalte 2) und nach verschiedenen Analysen wie der Messung des Abstandes zwischen den C α -Atomen der ersten und der letzten Aminosäure ($\leq 7\text{\AA}$, Spalte 3) sowie der Bestimmung der Drehung der Proteinkette ($\geq 120^\circ$, Spalte 4) noch übrig blieben. Bei *Turns* mit mehr als drei (bzw. vier im Falle von Regionen zwischen zwei β -Strängen) Aminosäuren war die Datenmenge in den meisten Fällen (weniger als 750 Vorkommen) zu gering um eine ausführliche Analyse zu ermöglichen. Aus diesem Grund wurden sie von den weiteren Untersuchungen ausgeschlossen.

Unterschiede der beiden Definitionen

Die Anwendung der einfachen Definition nach Chou [200] auf Regionen zwischen Sekundärstrukturen in der vorliegenden Arbeit führt zu einigen Unterschieden zwischen bereits bekannten *Turn*-Typen und den hier genutzten verfeinerten Typen. Zum einen findet in der vorliegenden Arbeit keine Differenzierung zwischen offenen *Turns*, die keine

Wasserstoffbrücke zwischen der ersten und der letzten Aminosäure ausbilden, sowie geschlossenen *Turns*, mit entsprechender Wasserstoffbrücke, statt. In der Literatur werden im Gegensatz dazu fast ausschließlich geschlossene *Turns* beschrieben [200]. Des Weiteren handelt es sich bei den meisten hier gefundenen Strukturen um *reverse Turns*, bei denen die Seitenketten der beteiligten Aminosäuren nach innen zeigen [207] (Abb. 20). Durch diese Unterschiede ist ein Vergleich der verfeinerten *Turns* mit bereits etablierten Arten nur schwer möglich. Zusätzlich stellen die hier gefundenen *XX1-Turns* eine starke Ausnahme dar, da bei ihnen die Drehung der Proteinkette zu großen Teilen durch die Aminosäuren vor und nach dem *Turn* realisiert wird, während die Aminosäure im *Turn* vor allem zur Stabilisierung der Sekundärstrukturmerkmale davor oder danach dient. Aus diesem Grund werden sie in den folgenden Abschnitten außer Acht gelassen.

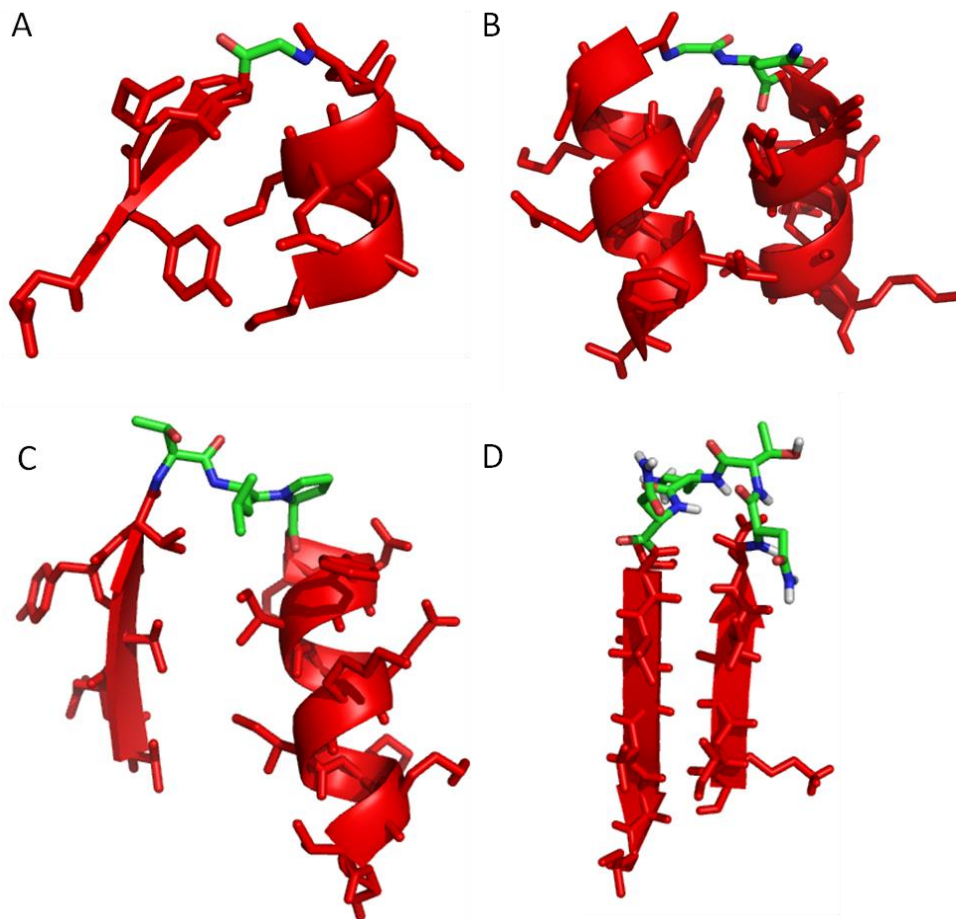


Abb. 20: PyMOL-Darstellungen der neu-definierten *Turn*-Typen

Die Abbildung zeigt PyMOL-Darstellungen eines HE1- (A), eines HH2- (B), eines EH3- (C) und eines EE4-Turns (D). Die Aminosäuren, die den jeweiligen *Turn* ausmachen, sind in grün als *Stick*-Repräsentation dargestellt, während die davor und danach liegenden Sekundärstrukturen als *Cartoon*- sowie als *Stick*-Repräsentation in rot gezeigt sind.

Phi/Psi-Cluster

Nachdem 13 verschiedene *Turns* definiert werden konnten, wurden die Phi/Psi-Winkel der im *Turn* vorkommenden Aminosäuren analysiert. Die erhaltenen Ergebnisse (Abb. 21 und elektronischer Anhang) verdeutlichen, dass es für jeden der dreizehn Typen eindeutige Phi/Psi-Cluster gibt, mit deren Hilfe man etwaige falsch positive Identifizierungen ausschließen kann. Die Ergebnisse stimmen sehr stark mit den Ergebnissen von Hovmöller *et al.* überein, die für Regionen außerhalb von α -Helices und β -Strängen sehr ähnliche Phi/Psi-Winkel-Verteilungen feststellen konnten [192].

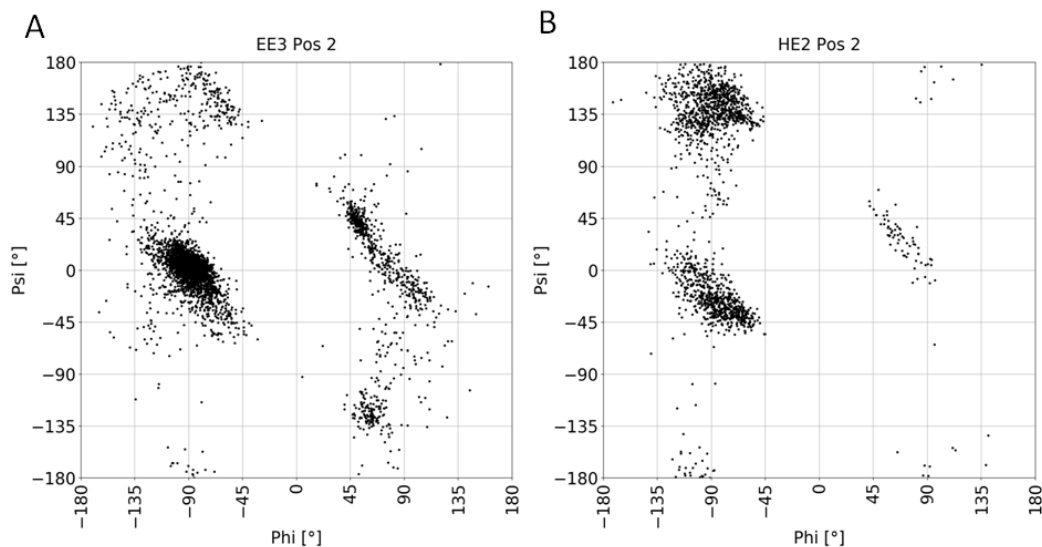


Abb. 21: Beispiele der Phi/Psi-Winkel der Aminosäuren in den *Turns*

Die Diagramme zeigen exemplarisch die berechneten Phi/Psi-Winkel der zweiten Aminosäure im *Turn* EE3 (A) sowie im *Turn* HE2 (B). In beiden Fällen sind deutliche *Cluster* der berechneten Winkelpaare zu erkennen, mit deren Hilfe die *Turns* besser charakterisiert werden können.

Da vor allem Glycin und Prolin deutlich abweichende Phi/Psi-Werte im Vergleich zu den restlichen Aminosäuren besitzen, wurden die Aminosäure-spezifischen Phi/Psi-Winkel in Abhängigkeit von der Position im *Turn* berechnet. Die in Abb. 22 beispielhaft gezeigten *Cluster* für Asparagin und Lysin wurden auch für weitere Aminosäuren gefunden (elektronischer Anhang). Auffällig ist zum einen der relativ große Anteil an Aminosäuren die im Bereich von $\Phi = 45^\circ - 90^\circ$ und $\Psi = 0^\circ - 45^\circ$ liegen. Hovmöller *et al.* diskutierten, dass dieser Bereich, ursprünglich als Bereich der linksgängigen α -Helix bezeichnet, eher als *Turn*-Bereich definiert werden sollte [192]. Zum anderen fällt auf, dass die Rückgrat-Torsionswinkel an Position 1 des *Turns* im α -helikalen Bereich des Ramachandran-Plots, an

Strukturbiologische Parameter

Position 2 größere Psi-Werte bei gleichbleibenden Phi-Werten (im Vergleich zu Position 1) und an Position 3 dann Werte des *Turn*-Bereiches annehmen. Interessanterweise konnten Hovmöller *et al.* für Regionen außerhalb von Helices und Strängen zeigen, dass Asparagin und Lysin (etwas seltener) sehr häufig Werte im Bereich von $\Phi = -135^\circ - -90^\circ$ und $\Psi = 0^\circ - 45^\circ$ annehmen können [192]. Des Weiteren konnten sie zeigen, dass Asparagin nach Glycin die Aminosäure mit den meisten Phi/Psi-Winkeln im *Turn*-Bereich ist [192] und somit anscheinend essentiell für die Ausbildung von *Turns* zu sein scheint.

Insgesamt zeigte sich, dass für *Turns* mit zwei oder mehr Aminosäuren klare Regeln gelten. Grundsätzlich muss eine der Aminosäuren Phi/Psi-Winkel im Bereich von $\Phi = -135^\circ - -90^\circ$ und $\Psi = 0^\circ - 45^\circ$ annehmen, während eine zweite Aminosäuren Torsionswinkel des von Hovmöller *et al.* vorgeschlagenen *Turn*-Bereichs [192] aufweisen muss. Ausgenommen von diesen Regeln sind lediglich Glycin und Prolin, die auch mit anderen Torsionswinkeln in der Lage sind, die Drehung der Proteinkette zu realisieren.

Strukturbiologische Parameter

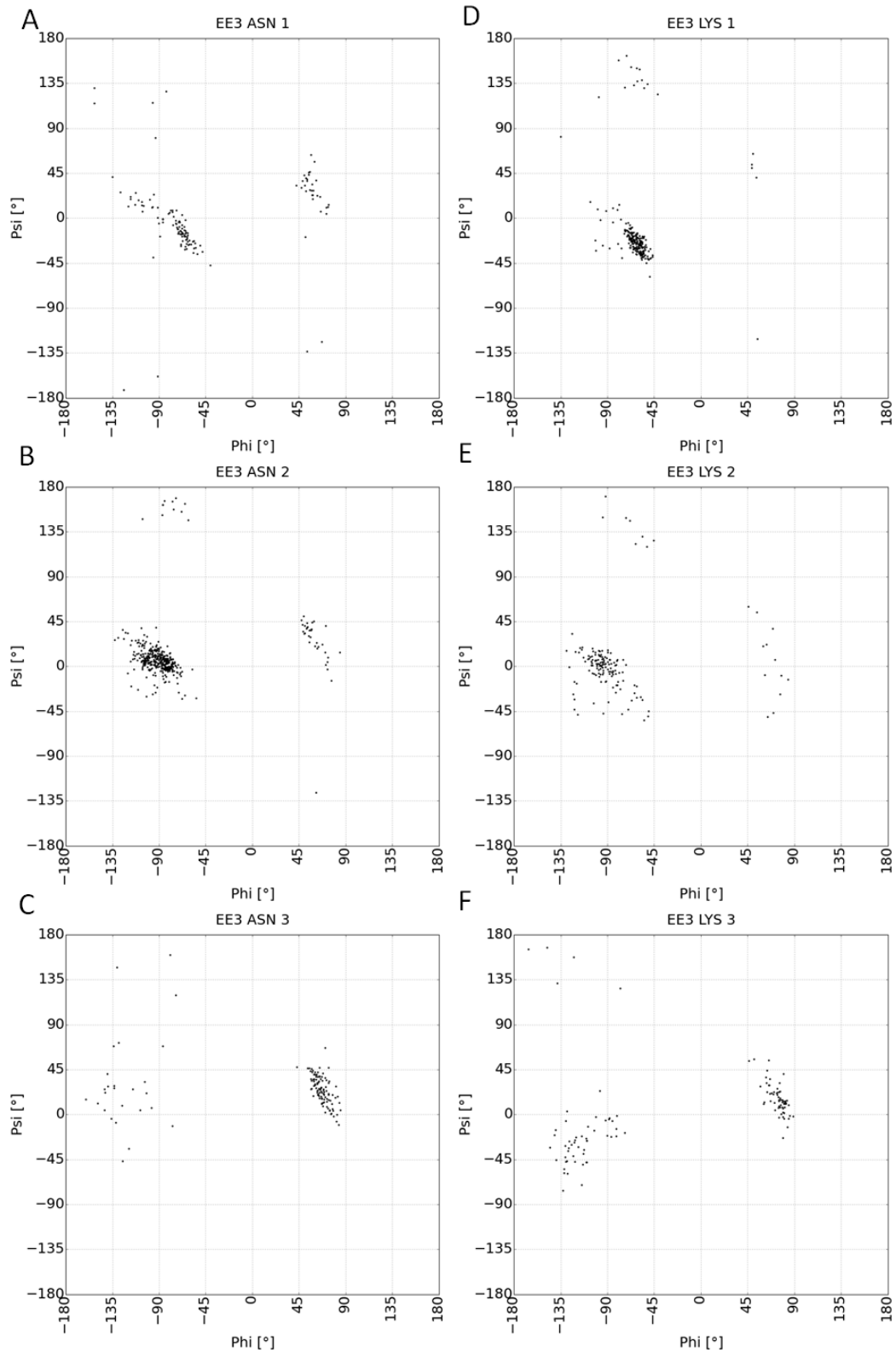


Abb. 22: Phi/Psi-Winkel von Asparagin und Lysin im *Turn* EE3

Die Diagramme zeigen die bestimmten Phi/Psi-Winkel von Asparagin (A - C) und Lysin (D - F) an erster Position (A und D), zweiter Position (B und E) oder dritter Position (C und F) im *Turn* EE3. Für beide Aminosäuren sind ähnliche positionsabhängige *Cluster* der Phi/Psi-Winkel zu erkennen.

Ausgehend von diesen Ergebnissen für die Phi/Psi-Winkel wurde überprüft, ob es spezielle Kombinationen von Phi/Psi-*Clustern* gibt, die die jeweiligen Positionen innerhalb eines *Turns* beschreiben können. Dafür wurden die Phi/Psi-*Plots* in 90° x 90° große Quadrate aufgeteilt und alle möglichen Kombinationen dieser Felder über den Verlauf eines *Turns* wurden gezählt. Für XX2-*Turns* ergeben sich so 256, für XX3-*Turns* 4096 und für EE4-*Turns* 65536 mögliche Phi/Psi-Kombinationen.

Als statistisch signifikant wurden nur Kombinationen angesehen, die in mindestens zehn Prozent aller gefundenen *Turns* vorkommen. Die Auswertung aller gefundenen Kombinationen für die XX2-*Turns* ergab, dass HH2-*Turns* 6, HE2- und EH2-*Turns* je 5 sowie EE2-*Turns* 9 von 256 möglichen Kombinationen annehmen können. Bei den XX3- und EE4-*Turns* finden sich nur 3 von 4096 bzw. 65536 möglichen Kombinationen, die als statistisch signifikant eingestuft werden können. Je nach *Turn*-Typ wurden teilweise bis zu 20 weitere, statistisch insignifikante Kombinationen gefunden.

Statistische Turn Präferenzen

Die statistische Auswertung der XX2-*Turns* ergab, dass vor allem Asparagin, Aspartat und Glycin bevorzugt in diesen *Turns* vorkommen, während aromatische Aminosäuren nur sehr selten in diesen *Turns* zu finden sind (Tabelle 40). Die hohen Vorkommen von Asparagin und Aspartat können durch die Ausbildung von Wasserstoffbrücken zwischen ihrer Seitenkette und dem Rückgrat der Aminosäuren in der dem *Turn* nachfolgenden Sekundärstruktur erklärt werden [189]. Hovmöller *et al.* konnten zeigen, dass beide Aminosäuren sehr häufig Phi/Psi-Winkel in den oben beschriebenen favorisierten *Clustern* annehmen können [192]. Da bei den hier vorliegenden *Turns* die Aminosäure-Seitenketten nach innen zeigen, würden größere Seitenketten zu sterischen Problemen führen. Dementsprechend kommen Glycin sehr häufig und aromatische Aminosäuren nur sehr selten in den hier gefundenen Typen vor.

Tabelle 40: Statistische Auswertung der XX2-Turns

Gezeigt sind statistische Präferenzen (berechnet anhand von Formel (15)). Die **fett** markierten Ergebnisse werden im Text diskutiert.

| Aminosäure | HH2 | | HE2 | | EH2 | | EE2 | |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Alanin | 0,60 | 0,80 | 1,37 | 0,64 | 0,88 | 0,36 | 0,58 | 0,23 |
| Arginin | 0,58 | 1,28 | 0,51 | 0,91 | 1,13 | 1,35 | 0,59 | 0,60 |
| Asparagin | 1,49 | 2,83 | 1,59 | 0,57 | 1,23 | 4,95 | 2,87 | 2,41 |
| Aspartat | 0,37 | 4,76 | 0,56 | 1,42 | 2,51 | 2,79 | 2,41 | 2,42 |
| Cystein | 0,60 | 0,30 | 0,97 | 0,81 | 0,64 | 0,48 | 0,32 | 0,25 |
| Glutamin | 1,96 | 1,09 | 0,63 | 0,56 | 0,54 | 0,31 | 0,86 | 0,59 |
| Glutamat | 1,22 | 0,83 | 0,26 | 0,56 | 0,88 | 0,46 | 1,63 | 0,75 |
| Glycin | 5,88 | 0,36 | 7,10 | 0,73 | 0,31 | 1,02 | 3,29 | 5,08 |
| Histidin | 0,41 | 1,24 | 0,60 | 1,30 | 4,55 | 0,53 | 0,73 | 0,92 |
| Isoleucin | 0,14 | 0,02 | 0,13 | 1,99 | 0,67 | 0,13 | 0,11 | 0,10 |
| Leucin | 0,63 | 0,34 | 0,11 | 1,22 | 0,43 | 0,19 | 0,18 | 0,12 |
| Lysin | 1,12 | 1,00 | 0,69 | 1,36 | 0,78 | 0,44 | 1,08 | 0,92 |
| Methionin | 1,28 | 0,44 | 0,13 | 0,98 | 0,54 | 0,24 | 0,46 | 0,13 |
| Prolin | 0,00 | 1,09 | 0,00 | 0,56 | 1,54 | 0,11 | 1,01 | 0,24 |
| Phenylalanin | 0,35 | 0,16 | 0,81 | 1,06 | 0,71 | 0,27 | 0,35 | 0,18 |
| Serin | 0,47 | 1,32 | 0,53 | 0,25 | 1,65 | 2,81 | 0,72 | 0,75 |
| Threonin | 0,05 | 0,64 | 0,20 | 0,46 | 1,05 | 2,01 | 0,48 | 0,94 |
| Tryptophan | 0,37 | 0,19 | 0,14 | 0,57 | 0,93 | 0,07 | 0,25 | 0,16 |
| Tyrosin | 0,25 | 0,38 | 0,64 | 1,44 | 0,46 | 0,32 | 0,32 | 0,89 |
| Valin | 0,29 | 0,18 | 0,24 | 2,11 | 0,39 | 0,19 | 0,13 | 0,14 |

Die Ergebnisse der statistischen Auswertung der XX3- und des EE4-Turns führte zu ähnlichen Ergebnissen wie bei den XX2-Turns (Tabelle 41). Auch hier zeigte sich ein häufiges Auftreten von Asparagin und Aspartat, die je nach *Turn* an unterschiedlichen Positionen zu finden ist. Bei *Turns*, die vor oder nach einer Helix zu finden sind, wird Asparagin in direkter Nachbarschaft dieser Helices gefunden, während bei EE-Turns Asparagin/Aspartat vor allem in der Mitte des *Turns* vorkommen. Dieses Verhalten resultiert zum einen aus den speziellen Torsionswinkeln, die diese beiden Aminosäuren annehmen können [192] und zum anderen aus ihrer Fähigkeit, Wasserstoffbrücken der Form $i/i+3$ oder $i/i+2$ mit ihren Seitenketten auszubilden (Abb. 23) [189]. Des Weiteren findet sich mindestens eine Position im Verlauf der *Turns*, an der Glycin sehr häufig vorkommt, um die nötige Drehung der Proteinkette zu vervollständigen. Zu guter Letzt sind auch hier aromatische Aminosäuren nur selten zu finden, da sie zu sterischen Problemen innerhalb der *Turns* führen könnten.

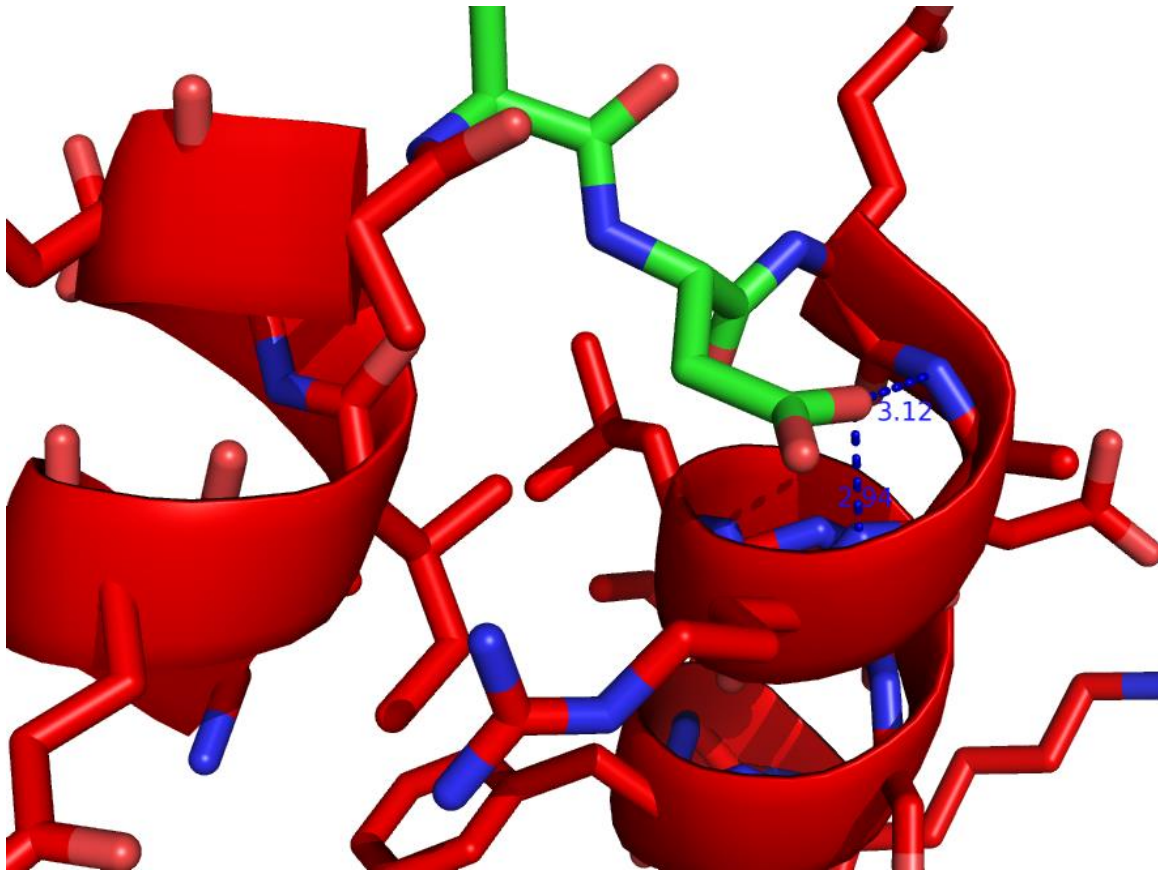


Abb. 23: Wasserstoffbrücken von Aspartat in HH2-Turns

Die Abbildung zeigt die *Stick*-Repräsentation eines HH2-Turns mit Aspartat an zweiter Position. Die Aminosäuren im *Turn* sind in grün und die Aminosäuren der davor bzw. danach folgenden Sekundärstruktur in rot gekennzeichnet. Die gestrichelten blauen Linien repräsentieren mögliche Wasserstoffbrücken des Aspartats mit den dazugehörigen Längen.

Strukturbiologische Parameter

Tabelle 41: Positionsspezifische Aminosäure-Präferenzen in XX3- und EE4-Turns

| Aminosäure | HH3 | | | HE3 | | | EH3 | | | EE3 | | | EE4 | | | |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 4 |
| Alanin | 0,84 | 0,70 | 0,21 | 0,40 | 1,51 | 0,51 | 0,83 | 1,14 | 0,42 | 1,11 | 0,67 | 0,22 | 0,92 | 1,41 | 0,55 | 0,30 |
| Arginin | 0,53 | 0,77 | 0,51 | 0,70 | 0,36 | 1,15 | 0,46 | 0,90 | 0,25 | 0,77 | 0,56 | 0,45 | 0,77 | 1,22 | 0,68 | 0,97 |
| Asparagin | 1,38 | 0,71 | 2,47 | 2,05 | 0,53 | 1,99 | 1,52 | 1,07 | 2,93 | 1,07 | 3,21 | 1,12 | 1,29 | 1,34 | 2,63 | 1,93 |
| Aspartat | 1,47 | 0,85 | 4,34 | 1,13 | 0,45 | 3,19 | 0,93 | 1,94 | 2,86 | 1,59 | 4,43 | 0,79 | 1,42 | 1,24 | 2,47 | 0,96 |
| Cystein | 1,18 | 0,89 | 0,37 | 1,21 | 1,03 | 0,18 | 1,12 | 0,63 | 3,48 | 0,40 | 0,68 | 0,23 | 0,83 | 0,34 | 1,17 | 0,34 |
| Glutamin | 0,64 | 0,79 | 0,55 | 0,70 | 0,41 | 1,85 | 0,40 | 0,89 | 1,53 | 0,73 | 0,87 | 0,55 | 0,71 | 1,11 | 0,71 | 1,31 |
| Glutamat | 0,57 | 0,67 | 0,83 | 0,34 | 0,38 | 1,27 | 0,84 | 0,82 | 0,74 | 1,66 | 0,82 | 0,35 | 0,77 | 1,59 | 0,61 | 0,82 |
| Glycin | 4,40 | 0,60 | 0,65 | 5,96 | 0,62 | 1,18 | 2,84 | 0,88 | 1,52 | 1,12 | 1,44 | 8,74 | 1,10 | 1,16 | 0,95 | 6,04 |
| Histidin | 1,50 | 0,57 | 0,69 | 1,29 | 0,51 | 1,15 | 0,48 | 0,76 | 0,55 | 0,68 | 1,55 | 1,31 | 0,72 | 0,68 | 0,87 | 1,25 |
| Isoleucin | 0,59 | 1,82 | 0,08 | 0,29 | 1,33 | 0,31 | 0,68 | 0,68 | 0,20 | 0,60 | 0,07 | 0,07 | 0,55 | 0,34 | 0,20 | 0,06 |
| Leucin | 0,49 | 1,75 | 0,09 | 0,34 | 0,65 | 0,22 | 0,54 | 0,66 | 0,34 | 0,60 | 0,42 | 0,16 | 1,15 | 0,36 | 0,50 | 0,27 |
| Lysin | 0,68 | 1,08 | 0,39 | 1,28 | 0,63 | 1,50 | 1,08 | 0,79 | 0,47 | 1,21 | 0,81 | 0,74 | 0,86 | 1,64 | 1,16 | 1,51 |
| Methionin | 0,33 | 1,37 | 0,38 | 0,16 | 0,41 | 0,30 | 0,28 | 0,65 | 0,09 | 0,76 | 0,24 | 0,19 | 0,77 | 0,15 | 2,04 | 0,31 |
| Prolin | 0,04 | 1,94 | 1,93 | 0,02 | 4,15 | 1,41 | 1,99 | 1,25 | 0,97 | 2,15 | 0,22 | 0,17 | 2,91 | 1,28 | 0,00 | 0,15 |
| Phenylalanin | 0,34 | 0,98 | 0,25 | 0,47 | 1,98 | 0,17 | 0,69 | 0,65 | 0,34 | 0,29 | 0,34 | 0,19 | 0,66 | 0,23 | 0,38 | 0,15 |
| Serin | 1,49 | 0,82 | 3,06 | 0,52 | 1,07 | 0,74 | 1,48 | 1,48 | 2,27 | 1,46 | 1,20 | 0,59 | 1,13 | 1,58 | 1,70 | 0,64 |
| Threonin | 0,85 | 0,91 | 1,65 | 0,42 | 0,41 | 1,07 | 1,04 | 1,30 | 1,20 | 0,82 | 0,72 | 0,34 | 1,31 | 1,12 | 2,91 | 0,33 |
| Tryptophan | 0,00 | 0,55 | 0,18 | 0,21 | 0,36 | 0,24 | 0,41 | 1,86 | 0,21 | 0,46 | 0,25 | 0,23 | 0,51 | 0,40 | 0,17 | 0,17 |
| Tyrosin | 0,42 | 0,75 | 0,30 | 0,64 | 0,86 | 0,28 | 0,42 | 0,47 | 0,38 | 0,47 | 0,76 | 0,25 | 0,35 | 0,26 | 0,79 | 0,26 |
| Valin | 0,48 | 0,88 | 0,10 | 0,27 | 1,71 | 0,37 | 0,53 | 1,02 | 0,32 | 0,62 | 0,12 | 0,09 | 0,47 | 0,58 | 0,17 | 0,04 |

2.4 Validierung von Proteinstrukturen

Um Proteinstrukturmodelle validieren zu können, ist es wichtig, eine ausführliche Bestimmung von Standard-Parametern durchzuführen. Dabei müssen verschiedene Aspekte beleuchtet werden, auf die in den nachfolgenden Abschnitten eingegangen werden soll.

2.4.1 Überprüfung von einfachen geometrischen Parametern

Eines der wichtigsten Qualitätskriterien von Proteinstrukturen sind einfache geometrische Parameter wie Bindungslängen und -winkel [31, 208], die Planarität von Aminosäure-Seitenketten [209], sowie die Chiralität von C α -Atomen [100]. Um diesem Rechnung zu tragen, wurden Programme wie ProCheck [72] oder WhatCheck [73] entwickelt, mit deren Hilfe eine allgemeine Überprüfung dieser Parameter möglich ist. Allerdings untersuchen diese Anwendungen ausschließlich das Protein-Rückgrat und lassen die Seitenketten der Aminosäuren außer Acht. Des Weiteren wurden die Standard-Parameter bereits zu Beginn der 1990'er Jahre ausgehend von nur sehr geringen Datenmengen [100] oder von kleinen Molekülen [208] bestimmt. Aus diesem Grund wurden sämtliche geometrische Parameter mit Hilfe des großen Datensatzes XRD erneut berechnet.

Tabelle 42: Zusammenfassung der berechneten maximalen RMSDs der Aminosäureseitenketten

Die Zahlen geben die maximal erlaubten RMSDs für alle Bindungslängen (L.) bzw. -winkel (W.) von Aminosäure-Seitenketten einer Art innerhalb einer Struktur an. Die Werte wurden so gewählt, dass maximal 5% der untersuchten Strukturen des Datensatzes XRD einen größeren Wert aufwiesen.

| Aminosäure | RMSD L. | RMSD W. | Aminosäure | RMSD L. | RMSD W. |
|------------|---------|---------|--------------|---------|---------|
| Alanin | 0,025 Å | 1,75° | Leucin | 0,025 Å | 2,50° |
| Arginin | 0,025 Å | 3,00° | Lysin | 0,035 Å | 2,50° |
| Asparagin | 0,030 Å | 2,25° | Methionin | 0,035 Å | 2,50° |
| Aspartat | 0,025 Å | 2,00° | Phenylalanin | 0,025 Å | 1,75° |
| Cystein | 0,025 Å | 2,50° | Prolin | 0,030 Å | 2,25° |
| Glutamat | 0,025 Å | 3,00° | Serin | 0,035 Å | 2,50° |
| Glutamin | 0,030 Å | 2,75° | Threonin | 0,025 Å | 2,50° |
| Glycin | 0,000 Å | 0,00° | Tryptophan | 0,020 Å | 1,75° |
| Histidin | 0,020 Å | 2,25° | Tyrosin | 0,020 Å | 2,00° |
| Isoleucin | 0,025 Å | 2,25° | Valin | 0,025 Å | 2,00° |

Der Vergleich der neu berechneten Daten für die Bindungslängen und -winkel mit denen von Engh und Huber [31] (siehe elektronischen Anhang für ausführliche Daten), sowie denen der

AMBER-Kraftfelder [210–213] zeigte nur sehr geringe Unterschiede im Bereich der zweiten bis dritten Nachkommastelle. Allerdings konnten die Standardabweichungen teilweise drastisch reduziert und so eine genauere Validierung der Bindungen ermöglicht werden. Da in den bisherigen Bewertungsprogrammen [72, 73] die Seitenketten immer außer Acht gelassen werden, ermöglichen die neuen Berechnungen eine Validierung der von den Protein-Modellierungsservern genutzten Seitenketten-Datenbanken. Dabei können sowohl einzelne problematische Bindungen als auch die gesamte Datenbank in Form einer RMSD-Analyse validiert werden [214]. Die berechneten Maximalwerte für die Bindungslängen (Tabelle 42 RMSD L.) und -winkel (RMSD W.) wurden so gewählt, dass mindestens 95% aller untersuchten Kristallstrukturen einen geringeren Wert aufweisen.

Tabelle 43: Mittlere RMS-Distanzen von planaren Gruppen

Hooft *et al.* [209] bestimmten die RMS-Distanzen von Atomen einer planaren Gruppe zu der dazugehörigen Ebene sowohl für PDB-Strukturen (Hooft P) wie auch für kleine organische Moleküle aus der CSD (Hooft C). Dabei wurden für aromatische Systeme nur die Atome des aromatischen Ringsystems zur Etablierung der Ebene genutzt, während für die Daten in der vierten Spalte alle Atome als Teil der Ebene angesehen wurden. Die Daten der vierten Spalte (XRD H) wurden analog des von Hooft *et al.* genutzten Systems berechnet.

| Aminosäure | Hooft C [209] | Hooft P [209] | XRD | XRD H | MBD | SMD |
|-----------------------------|---------------|---------------|----------|----------|----------|----------|
| Arginin | 0,0370 Å | 0,0340 Å | 0,0026 Å | | 0,0011 Å | 0,0036 Å |
| Asparagin | 0,0059 Å | 0,0104 Å | 0,0041 Å | | 0,0015 Å | 0,0056 Å |
| Aspartat | 0,0063 Å | 0,0090 Å | 0,0037 Å | | 0,0013 Å | 0,0082 Å |
| Glutamin | 0,0059 Å | 0,0102 Å | 0,0035 Å | | 0,0012 Å | 0,0041 Å |
| Glutamat | 0,0063 Å | 0,0091 Å | 0,0030 Å | | 0,0009 Å | 0,0068 Å |
| Histidin | 0,0046 Å | 0,0082 Å | 0,0044 Å | 0,0044 Å | 0,0004 Å | 0,0022 Å |
| Histidin-C β | 0,0500 Å | 0,0660 Å | 0,0078 Å | 0,1196 Å | 0,0043 Å | 0,0224 Å |
| Phenylalanin | 0,0089 Å | 0,0091 Å | 0,0052 Å | 0,0052 Å | 0,0020 Å | 0,0117 Å |
| Phenylalanin-C β | 0,0640 Å | 0,0590 Å | 0,0089 Å | 0,0433 Å | 0,0024 Å | 0,0267 Å |
| Tryptophan | 0,0135 Å | 0,0144 Å | 0,0066 Å | 0,0066 Å | 0,0027 Å | 0,0230 Å |
| Tryptophan-C β | 0,0740 Å | 0,0740 Å | 0,0098 Å | 0,0367 Å | 0,0082 Å | 0,0350 Å |
| Tyrosin | 0,0082 Å | 0,0095 Å | 0,0051 Å | 0,0051 Å | 0,0020 Å | 0,0127 Å |
| Tyrosin-C β | 0,0610 Å | 0,0640 Å | 0,0092 Å | 0,0467 Å | 0,0028 Å | 0,0267 Å |
| Tyrosin-O η | 0,0400 Å | 0,0450 Å | 0,0064 Å | 0,0146 Å | 0,0021 Å | 0,0222 Å |
| Tyrosin-C β -O η | | | 0,0107 Å | | 0,0029 Å | 0,0382 Å |

Die Planarität von Seitenketten wurde anhand der RMS-Abstände der Atome einer planaren Gruppe (z.B. aromatische Ringe oder Säuregruppen) zu der dazugehörigen Ebene überprüft [209]. Die neu berechneten Mittelwerte sind in allen Fällen geringer als bereits publizierte Daten (Tabelle 43). Die Hauptgründe für dieses Verhalten liegen zum einen in der deutlich

höheren Datenmenge im Vergleich zur Analyse von Hooft *et al.* [209], und zum anderen in einem etwas veränderten System der Berechnung. Während in der vorliegenden Arbeit die Ebene zur Berechnung auch durch die Substituenten gelegt wurde, haben Hooft *et al.* nur den aromatischen Ring als Ebene definiert und dann den Abstand des Substituenten zu dieser Ebene bestimmt [209]. Besonders auffällig sind die deutlichen Unterschiede bei den aromatischen Ringsystemen wenn die Substituenten (C β oder O η) in die Berechnung mit einbezogen werden. Die erhaltenen Werte zeigen, dass die Substituenten leicht über- oder unterhalb der aromatischen Ringebene stehen. Ein zweiter Berechnungsdurchlauf für den Datensatz XRD unter Anwendung des Berechnungssystems von Hooft *et al.* (Tabelle 43 letzte Spalte) konnte dieses Verhalten bestätigen. Des Weiteren waren die Werte für Arginin deutlich geringer als bei Hooft *et al.*, was auf der Nicht-Beachtung des C δ -Atoms in der vorliegenden Arbeit beruht.

Ein Vergleich der Ergebnisse der Modell-Datensätze mit den Ergebnissen der Kristallstrukturen zeigt, dass sie sich stark voneinander unterscheiden (Tabelle 43). Während die Modelle aus der *ModBase* deutlich näher an der Planarität sind als die Kristallstrukturen, weichen die Modelle von *SwissModel* vor allem bei den aromatischen Aminosäuren sehr stark von der Planarität ab. Hier scheint es ein Problem in der Seitenketten-Datenbank dieses Servers zu geben.

Tabelle 44: Berechnete Zeta-Winkel zur Überprüfung der Chiralität von Aminosäuren

| Aminosäure | Winkel | Zeta |
|---------------|---|--------------|
| Standard | C α -N-C-C β | 34,0° ± 1,4° |
| Prolin | C α -N-C-C β | 38,4° ± 1,1° |
| Ile C β | C β -C γ_1 -C α -C γ_2 | 33,1° ± 1,3° |
| THR C β | C β -O γ_1 -C α -C γ_2 | 34,7° ± 1,4° |

Das letzte geometrische Kriterium zur Validierung von Proteinstrukturen ist der pseudo-Torsionswinkel Zeta, mit dessen Hilfe die Chiralität von sp³-hybridisierten Kohlenstoff-Atomen überprüft werden kann [100]. Er findet bisher vor allem bei der Validierung von C α -Atomen Anwendung und kann zusätzlich Informationen über die Konfiguration (L- oder D-Aminosäuren, siehe Eigenschaften von Aminosäuren) liefern [100]. Durch die Anwendung der Zeta-Winkel-Berechnung auf die C β -Atome von Isoleucin und Threonin kann auch die Chiralität der Seitenketten überprüft werden (Tabelle 44).

Strukturbiologische Parameter

Insgesamt konnte gezeigt werden, dass die meisten bisher genutzten geometrischen Standard-Parameter, z.B. Bindungslängen und -winkel von Engh und Huber [208] oder RMS-Distanzen der Planarität von Hooft *et al.* [209], noch immer ihre Berechtigung haben. Da sie während der Erstellung von experimentellen Proteinstrukturen oftmals starken *Restraints* unterliegen [31], konnten keine großen Unterschiede zu bereits publizierten Daten gefunden werden. Bei der Anwendung dieser Parameter zur Validierung von *in-silico* erstellten Proteinstrukturmodellen, sollten sich dementsprechend auch nur geringe Abweichungen zeigen, da diese in den meisten Fällen auf Kristallstrukturen aufbauen [215]. Trotzdem ist es wichtig, vor allem die Geometrie der Aminosäure-Seitenketten einer intensiven Prüfung zu unterziehen, um systematische Fehler der Modellierungsprogramme und/oder -server ausschließen zu können. Mit den aktuell am meisten verwendeten Programmen ProCheck [72] und WhatCheck [73] ist dies jedoch nicht möglich, da sie nur das Protein-Rückgrat untersuchen und dabei nur Fehler einzelner Bindungen aufzeigen. Sollte es Fehler in den Seitenketten-Datenbanken der Modellierungsserver oder -programme geben, sind diese nur schwer identifizierbar. Dementsprechend liefern die vorgestellten Ergebnisse eine ideale Basis für die Entwicklung neuer Validierungsmechanismen.

2.4.2 Analyse von Torsionswinkeln

Neben den einfachen geometrischen Parametern wie Bindungslängen und -winkel, sind vor allem die Torsionswinkel innerhalb einer Proteinstruktur ein wichtiger Indikator für ihre Qualität. Dabei ist es wichtig, nicht nur die Torsionswinkel des Protein-Rückgrates (Phi, Psi und Omega) zu betrachten, sondern auch ein großes Augenmerk auf die Chi-Winkel der Aminosäure-Seitenketten zu legen. Durch die Untersuchung der Aminosäure-Seitenketten können Probleme im Wasserstoffbrücken-Netzwerk, bzw. sterische Probleme eventuell gelöst werden.

Strukturbiologische Parameter

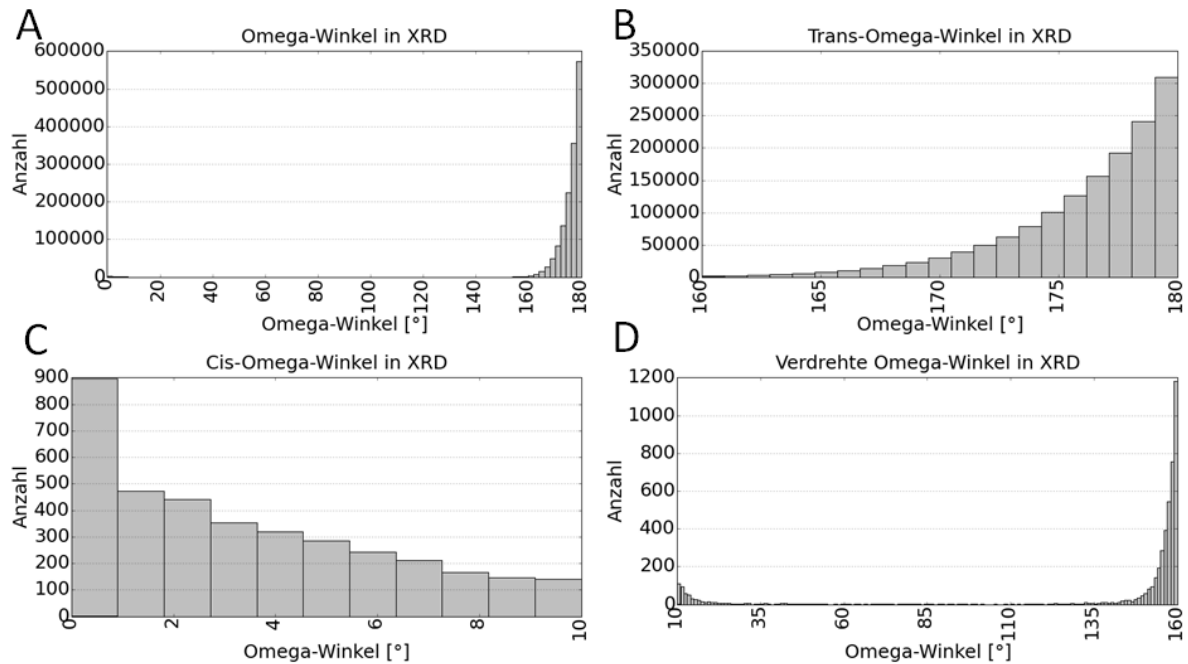


Abb. 24: Omega-Winkel im Datensatz XRD

Die vier Histogramme zeigen die Verteilung des Betrages der Omega-Winkel für den Datensatz XRD. Gezeigt sind die komplette Verteilung (A), die Verteilung der Omega-Winkel im *trans* Bereich (160° bis 180°, B, 99,44% der Werte), im *cis* Bereich (0° bis 10°, C, 0,25% der Werte) sowie im Bereich der verdrehten Winkel (10° bis 160°, D, 0,31% der Werte).

Wie weiter oben beschrieben (Eigenschaften von Aminosäuren), kann der Omega-Torsionswinkel genutzt werden, um ungewöhnliche Konformationen der Peptid-Bindung zu identifizieren. Berkholz *et al.* wiesen im Zuge einer Analyse von hochaufgelösten Kristallstrukturen (Auflösung kleiner als 1,0 Å) nach, dass es viele *trans*-Peptidbindungen, gibt deren Omega-Winkel um mehr als 20° von der Planarität (180°) abweichen [216]. Untersuchungen des Betrages dieses Winkels für den Datensatz XRD zeigten, dass zwei verschiedene Konformationen unterschieden werden können (Abb. 24). Bei diesen handelt es sich um *trans*-Peptidbindungen ($\Omega \geq 160^\circ$) und um *cis*-Peptidbindungen ($\Omega \leq 10^\circ$). Alle Bindungen mit davon abweichenden Werten können als verdrehte Peptidbindungen ($10^\circ < \Omega < 160^\circ$) bezeichnet werden. Während die verdrehten Bindungen als problematisch angesehen werden sollten, kann ein gewisser Teil an *cis*-Peptidbindungen durchaus erlaubt sein, allerdings sollte vermieden werden ihre Bedeutung zu überinterpretieren [107].

Tabelle 45: Peptidbindungskonformationen der verschiedenen Datensätze

| Konformation | XRD | MBD | SMD |
|---------------------------|---------|--------|--------|
| # Peptidbindung | 1482915 | 665000 | 386240 |
| # trans-Peptidbindung | 1474584 | 662622 | 378483 |
| % trans-Peptidbindung | 99,44% | 99,64% | 97,99% |
| # cis-Peptidbindung | 3670 | 1063 | 813 |
| % cis-Peptidbindung | 0,25% | 0,16% | 0,21% |
| # verdrehte Peptidbindung | 4661 | 1315 | 6944 |
| % verdrehte Peptidbindung | 0,31% | 0,20% | 1,80% |

Rund 0,31% aller Peptidbindungen im Datensatz XRD kann als verdreht angesehen werden (Tabelle 45). Dieser Anteil entspricht den Ergebnisse von Berkholz *et al.*, die für Kristallstrukturen mit Auflösungen von 2,5 Å bis 1,2 Å Anteile von 0,20% bis 0,40% nachweisen konnten. Bei den Modell-Datensätzen konnten für die Modelle der *ModBase* sehr ähnliche Ergebnisse gefunden werden, während die Modelle von *SwissModel* deutlich mehr verdrehte Peptidbindungen aufweisen. Möglicherweise liegt hier ein Problem bei der Modellierung des Proteinrückgrates vor.

Strukturbiologische Parameter

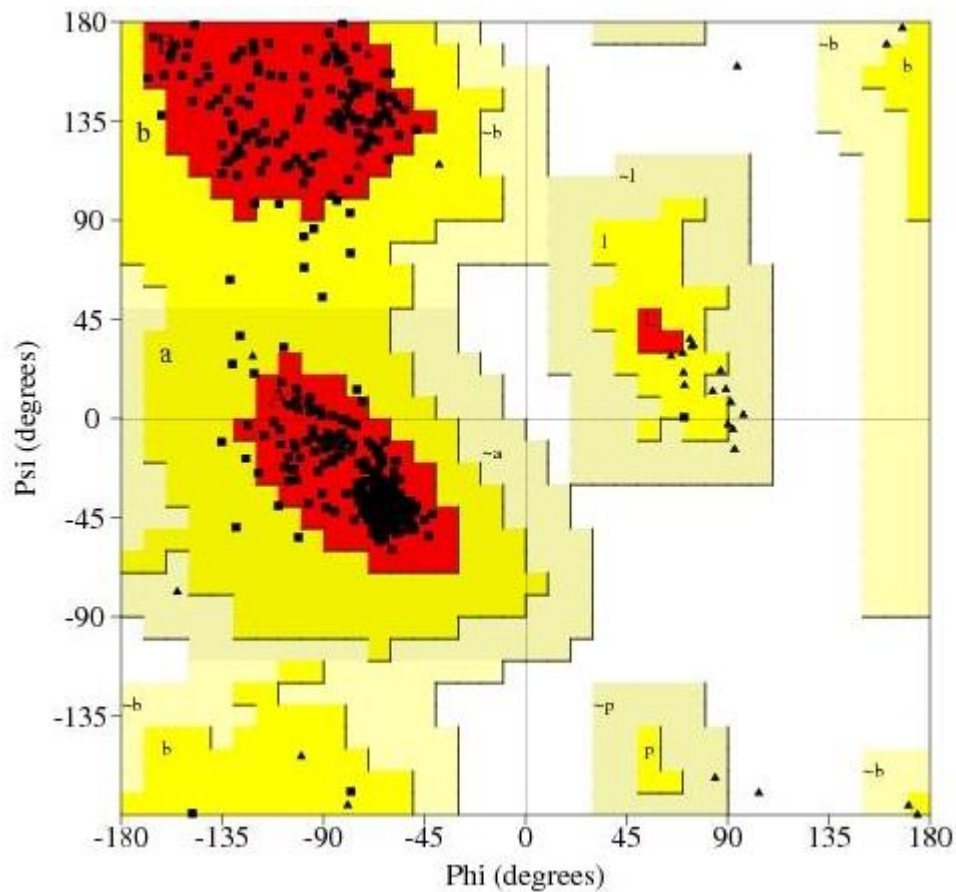


Abb. 25: **Klassischer Ramachandran-Plot**

Die Abbildung zeigt einen klassischen Ramachandran-Plot [217] mit den von Morris [100] definierten Regionen favorisiert (rot), erlaubt (gelb), grenzwertig (hellgelb) und verboten (weiß).

Mit der Entwicklung des Phi/Psi-Plots durch Ramachandran [217–219], der Festlegung von favorisierten und erlaubten Bereichen durch Morris [100] sowie der Weiterentwicklung durch das Richardson-Lab [220, 221] wurde ein wichtiges Instrument zur Überprüfung von Proteinstrukturen (Abb. 25 und Abb. 26) geschaffen.

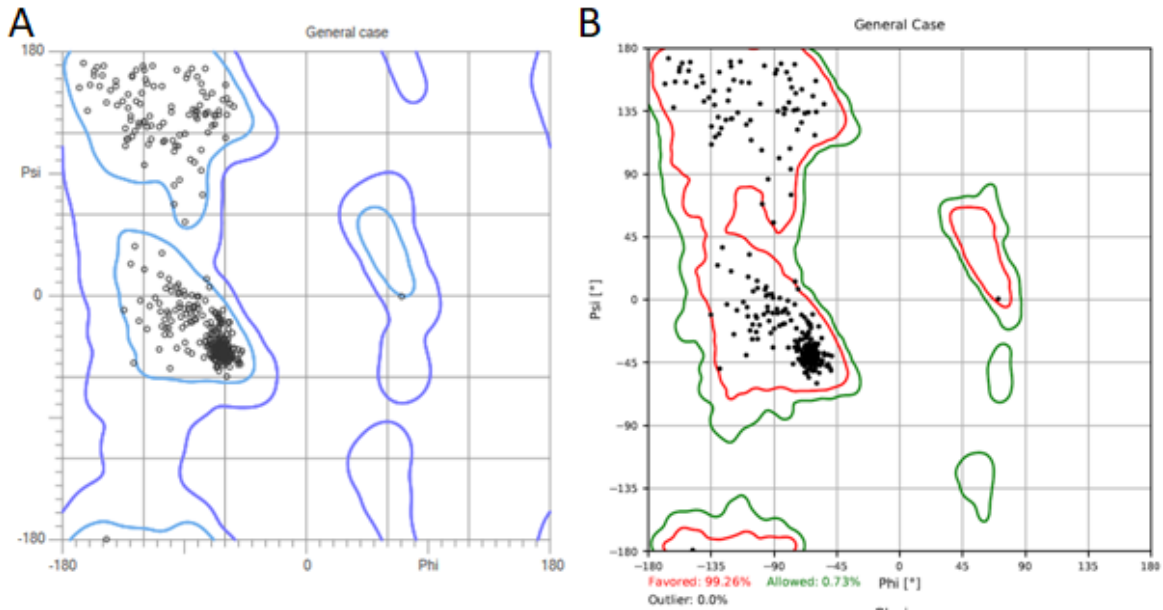


Abb. 26: Vergleich der PhiPsi-Plots von Lovell *et al.* [220] und den in dieser Arbeit durchgeführten Berechnungen

Die beiden Diagramme zeigen die Phi/Psi-Plots für den allgemeinen Fall, die mit Hilfe der von Lovell *et al.* postulierten Dichte-abhängigen Glättungsfunktion erstellt wurden. Es sind deutliche Unterschiede zwischen den erlaubten Bereichen der Originalliteratur (dunkelblaue Kontur in A) und der Neu-Berechnung (grüne Kontur in B) zu erkennen. Außerdem gibt es leichte Unterschiede bei der favorisierten Regionen, vor allem im Bereich $90^\circ < \text{Psi} < 45^\circ$ und $-150^\circ < \text{Phi} < -120^\circ$.

Ausgehend von den Ergebnissen von Lovell *et al.* [220], wurden Dichte-Berechnungen der Phi/Psi-Winkel für alle kanonischen Aminosäuren durchgeführt. Dabei wurde eine geglättete, normalisierte Dichte-Funktion genutzt, die als Summe der Dichten aller Datenpunkte N wie folgt definiert ist:

$$\rho(\phi, \psi) = \sum_{i=1}^N \sigma_i(\phi, \psi) \quad (4)$$

Um eine komplette Dichte zu erhalten wird der Phi/Psi-Plot in gleichmäßige Quadrate ($2^\circ \times 2^\circ$) aufgeteilt und für jeden Eckpunkt dieser Quadrate die Dichte anhand von Formel (4) berechnet. Die Verteilungen werden als normalisierte Kosinus-Maske in Abhängigkeit von den untersuchten Torsionswinkeln definiert. Die Berechnung wird dabei für jeden Datenpunkt i und für alle x_i , die kleiner als α_i , sind durchgeführt.

$$\sigma_i(\phi, \psi) = \frac{\pi}{\alpha_i^2 (\pi^2 - 4)} \left\{ \cos\left(\frac{\pi x_i}{\alpha_i}\right) + 1 \right\} \quad (5)$$

Dabei ist x_i wie folgt definiert:

$$x_i = \sqrt{(\phi - \phi_i)^2 + (\psi - \psi_i)^2} \quad (6)$$

Strukturbiologische Parameter

Zur endgültigen Erstellung der Plots werden zwei Iterationen durchgeführt wobei für α_i in der ersten Iteration für alle Datenpunkte i ein Wert von 10° angenommen wurde. In der zweiten Iteration kann dann α_i wie folgt berechnet werden:

$$\alpha_i = k\rho_1(\phi_i, \psi_i)^{-\lambda/n} \quad (7)$$

Als Konstanten wurden $\lambda = 0,5$, $n = 2$ und $k = 13$ für den allgemeinen *Plot* und $k=16$ für die Sonderfälle genutzt.

Dabei konnten neben den bereits postulierten *Extra-Plots* für Glycin, Prolin und Prä-Prolin (Abb. 27), noch vier weitere in Form von Glutamat, Histidin, Isoleucin und Valin entwickelt werden (Abb. A.9). Ein Blick auf die *Plots* zeigt, dass es nur geringe Unterschiede bei Glycin und Prolin in der Verteilung der favorisierten und erlaubten Regionen gibt (Abb. 27 A, B, D und E). Diese resultieren aus leicht unterschiedlichen Grenzsetzungen für die Konturen bzw. geringen Unterschieden der Datengrundlage. Lovell *et al.* nutzten Grenzwerte von 98% für die favorisierten Bereiche und 99,95% (allgemeine Verteilung) bzw. 99,8% (Sonderfälle) für die erlaubten Bereiche [220]. In der vorliegenden Arbeit wurden für alle Diagramme Grenzwerte von 99,8% für den erlaubten Bereich und 99% für den favorisierten Bereich genutzt. Bei den Prä-Prolinen (Abb. 27 C und F) resultieren etwaige Unterschiede im ersten und vierten Quadranten aus der Inklusion von Glycin bei der Erstellung der Dichten. Dies steht im Gegensatz zum Ansatz von Lovell *et al.*, die bei der Erstellung ihres Diagramms nur nicht-Glycine in den Prä-Prolin-*Plot* integrierten [220].

Strukturbiologische Parameter

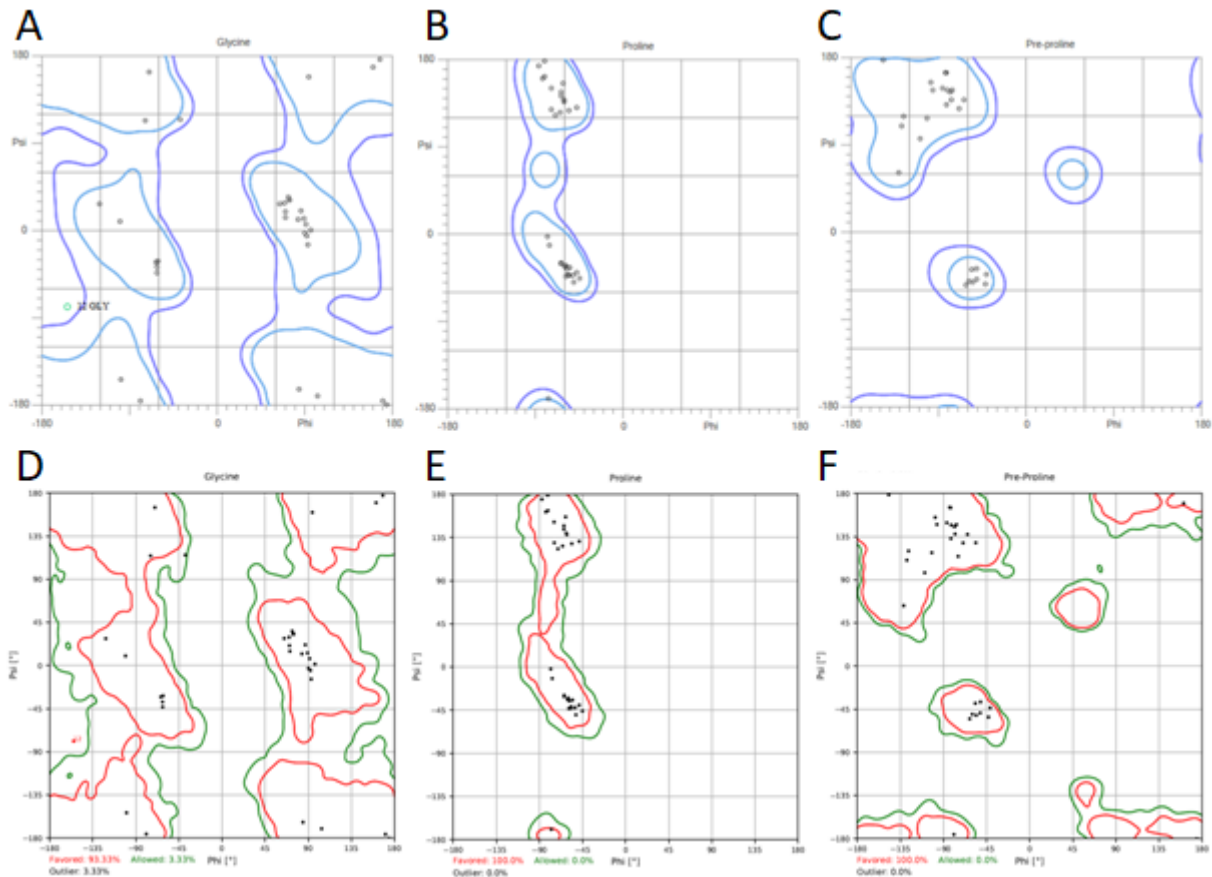


Abb. 27: Phi/Psi-Diagramme der Sonderfälle Glycin, Prolin und Prä-Prolin

In den sechs Diagrammen sind die berechneten idealen Regionen für die Aminosäuren Glycin (A und D), Prolin (B und E) sowie Aminosäuren vor Prolin (C und F) zu sehen. Es sind immer die *Plots* von Lovell *et al.* (obere Zeile) den Neuberechnungen (untere Zeile) gegenübergestellt. Vor allem bei den Prä-Prolinen sind deutliche Unterschiede in den äußersten Bereichen des ersten und vierten Quadranten zu erkennen.

Neben den Torsionswinkeln des Proteinrückgrates sollten auch die Seitenketten-Torsionswinkel (Chi1 bis Chi4) in ähnlicher Form analysiert werden. Dazu wurde die oben genannte Dichte-abhängige Glättungsfunktion auf Chi1/Chi2-, Chi2/Chi3- sowie Chi3/Chi4-Diagramme der Aminosäure-Seitenketten angewandt. Insgesamt konnten so 12 Chi1/Chi2-, 4 Chi2/Chi3- sowie 2 Chi3/Chi4-*Plots* etabliert werden (Beispiel in Abb. 28, komplette Übersicht in Tabelle A.5 und im elektronischen Anhang).

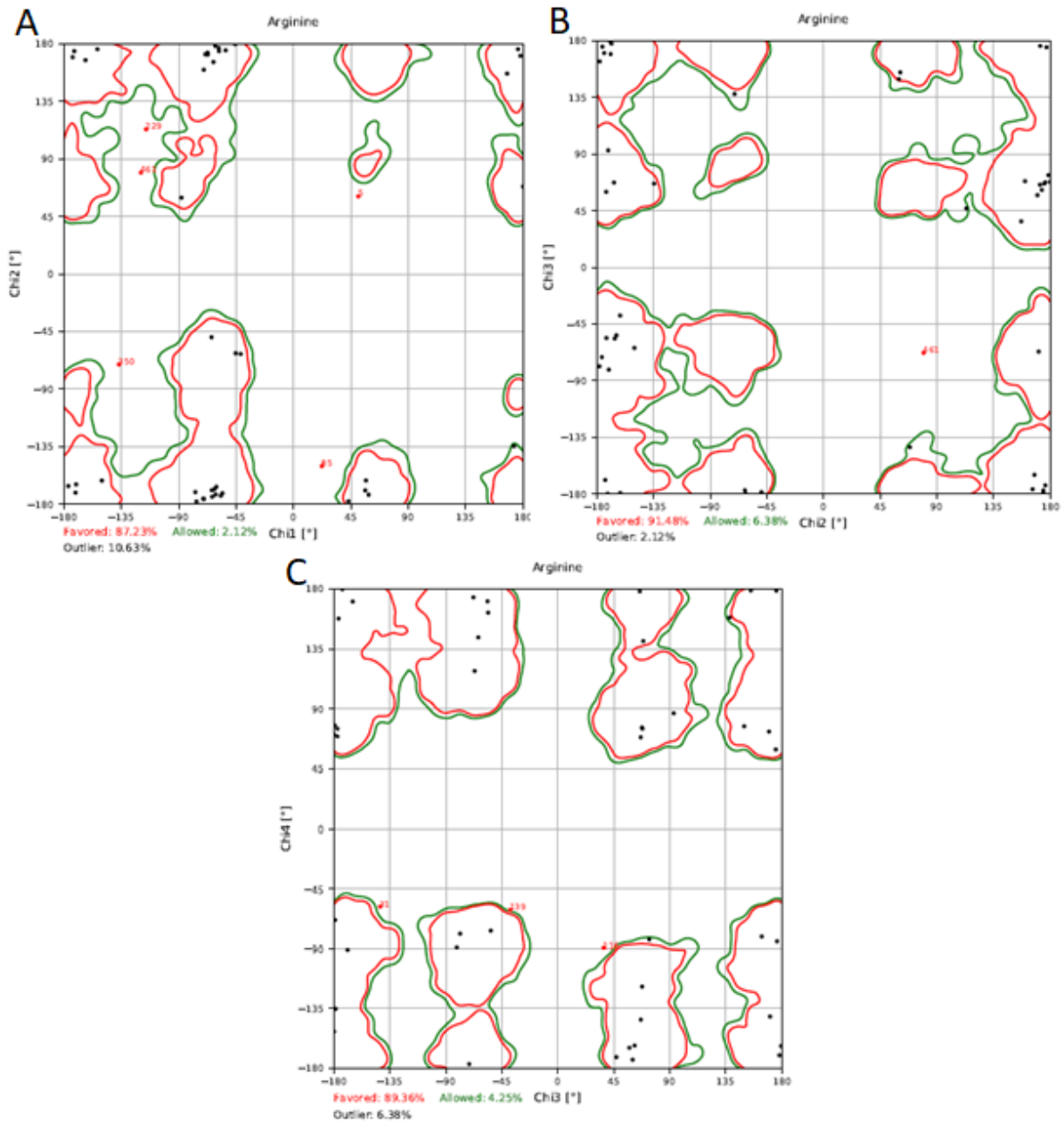


Abb. 28: Seitenketten-Torsionswinkel-Diagramme für Arginin

Die drei Diagramme zeigen die berechneten Chi1/Chi2- (A), Chi2/Chi3- (B) und Chi3/Chi4-Diagramme (C) anhand des Beispiels Arginin. Die Unterschiede zwischen den einzelnen *Plots* sowie die vorliegende Abhängigkeit der Torsionswinkel untereinander sind gut erkennbar. Die erlaubten Regionen (grüne Kontur) liegen in fast allen Bereichen sehr nah an der favorisierten Region (rote Kontur).

Die drei Diagramme in Abb. 28 zeigen, dass es präferierte Kombinationen der Seitenketten-Torsionswinkel gibt. Besonders auffällig sind die nicht populierte Regionen in den mittleren Bereichen beider Achsen (vergleiche Abb. 28 und Abb. 29), die sich auch für andere Aminosäuren so finden lassen (elektronischer Anhang). Insgesamt bestehen die *Plots* aus bis

zu 16 möglichen idealen Kombinationen, deren grundsätzliche Lage für alle Aminosäuren identisch ist. Bei diesen Regionen handelt es sich um alle möglichen Kombinationen der drei idealen Chi-Winkel *gauche*⁻, *gauche*⁺ und *trans*.

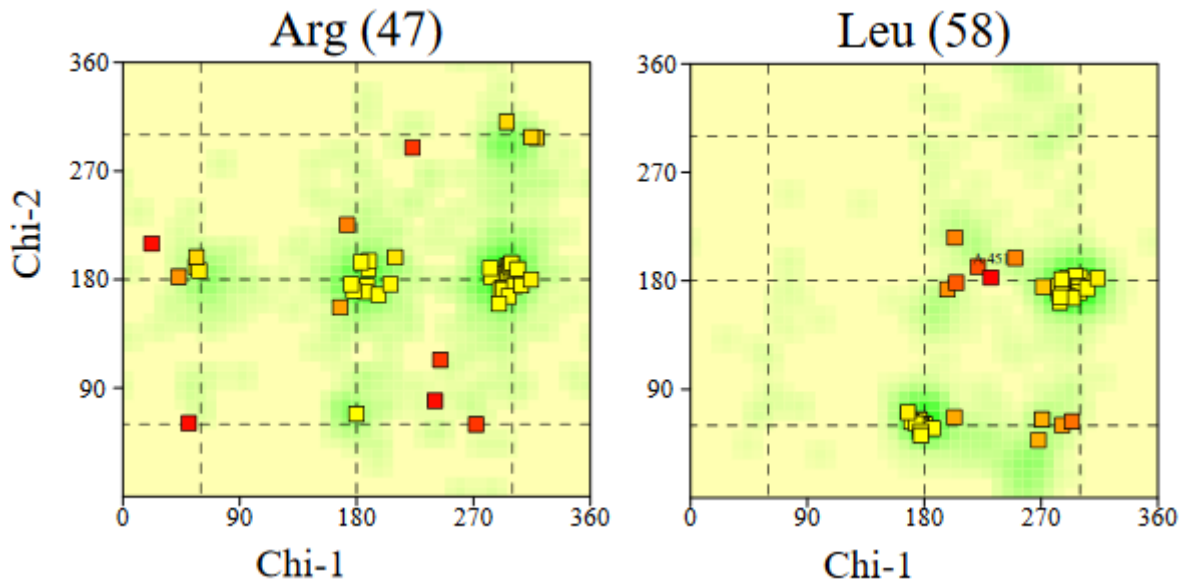


Abb. 29: Chi1/Chi2-Diagramme von Arginin und Leucin erstellt mittels ProCheck

Die beiden Plots zeigen die von ProCheck [72] entwickelten Chi1/Chi2-Diagramme für die Aminosäuren Arginin und Leucin. Obwohl die Winkel-Einteilung etwas anders ist als bei den neu-berechneten Diagrammen, sind zumindest gewisse Ähnlichkeiten zu erkennen. Die favorisierten Regionen sind durch eine Grünfärbung der Zeichenfläche gekennzeichnet.

Eine genauere Betrachtung der einzelnen Chi-Plots zeigt, dass zumindest einige kleine Regelmäßigkeiten existieren. Die Kombination von *gauche*⁺ (~60°) auf der x-Achse und *gauche*⁻ (~-60°) auf der y-Achse ist nur sehr selten, unabhängig von Aminosäure und Diagramm-Typ, als favorisierte Region gekennzeichnet, was mit den Plots von ProCheck durchaus übereinstimmt, denn auch dort ist dieser Bereich (Winkel von 60° für *gauche*⁺ und 300° für *gauche*⁻) als unwahrscheinlich gekennzeichnet (Abb. 29). Des Weiteren scheint ein Chi2-Winkel von ±150° bis ±180° für aromatische Aminosäuren (Phenylalanin, Tryptophan und Tyrosin) energetisch ungünstig zu sein. Da aromatische Aminosäuren bevorzugt eine "*gauche*" (±60°) oder "*perpendicular*" (±90°) Konformation annehmen und die "*eclipsed*" Konformation energetisch eher ungünstig ist [222], entsprechen die hier gefundenen Ergebnisse den stereochemischen Standards.

Ein weiteres Merkmal dieser Diagramme ist, dass die erlaubten Regionen (grüne Konturen) in fast allen Bereichen sehr nah an den favorisierten Regionen (rote Konturen) liegen. Das

spricht für eine insgesamt relativ enge Verteilung der zugrunde liegenden Daten, und deutet somit auf eine klare Differenzierung zwischen den präferierten Winkel-Kombinationen und den energetisch ungünstigen Kombinationen hin.

Insgesamt ermöglichen die vorgestellten neuen *Plots* eine umfassende Bewertung von Proteinstrukturen hinsichtlich ihrer Torsionswinkel. Insbesondere die vier neuen Phi/Psi-*Plots* für die Sonderfälle (Glu, His, Ile und Val, siehe Abb. A.9), von denen zumindest zwei inzwischen auch durch das *Richardson-Lab* bestätigt wurden [221], verdeutlichen den Wert dieses Ansatzes im Bezug auf die Validierung von Proteinstrukturen. Durch die Nutzung des Ansatzes von Lovell *et al.* [220] zur Erstellung von neuen Chi1/Chi2-Diagramme und der Erweiterung dieses Ansatzes auf höhere Seitenketten-Torsionswinkel (Chi2/Chi3- oder Chi3/Chi4-*Plots*) wird ein noch differenzierterer Blick auf diese Bereiche der Proteinstruktur möglich.

2.4.3 Aminosäure-Volumina und sterische Probleme

Das Volumen von Aminosäuren kann entweder mit Hilfe des Voronoi-Ansatzes [223], da dieser eine gute Näherung für die relativ komplizierte Berechnung darstellt [224–227], oder unter Nutzung eines *hard-sphere* Modells durchgeführt werden [228, 229]. Beim Voronoi-Ansatz werden Polyeder rund um die untersuchten Atome erstellt [224]. Die Grenzflächen dieser Polyeder werden dabei durch die umliegenden Atome definiert [224]. Die Kanten dieser Grenzflächen entsprechen einer Senkrechten in der Mitte der Geraden zwischen dem Atomkern des untersuchten Atoms und den Atomkernen der umliegenden Atome [224]. Da eine Implementierung dieses Ansatzes sich als zu kompliziert herausstellte, wurde stattdessen auf das *hard-sphere*-Modell zurückgegriffen.

Tabelle 46: Kovalente und van-der-Waals Radien ausgesuchter Elemente

| Atom | Kovalenter Radius [230, 231] | Van-der-Waals Radius [232] |
|---|------------------------------|----------------------------|
| sp ² -hybridisierter Kohlenstoff | 0,73 Å | 1,70 Å |
| sp ³ -hybridisierter Kohlenstoff | 0,76 Å | 1,70 Å |
| Stickstoff | 0,71 Å | 1,55 Å |
| sp ² -hybridisierter Sauerstoff | 0,57 Å | 1,52 Å |
| sp ³ -hybridisierter Sauerstoff | 0,63 Å | 1,52 Å |
| Schwefel | 1,03 Å | 1,80 Å |

Die Berechnung der van-der-Waals-Volumina von einzelnen Aminosäuren mit Hilfe eines *hard-sphere*-Modells stellte sich als äußerst kompliziert dar, da es zur Überlappung von

Strukturbiologische Parameter

mehreren Kugeln kam und die Volumen-Berechnung solcher Konstrukte sich als nicht trivial erwies [233–236]. Aus diesem Grund wurden stattdessen kovalente Volumina genutzt. Mit Hilfe der kovalenten Radien der einzelnen Atome [230, 231] (Tabelle 46) konnte ein System zur Berechnung des kovalenten Volumens von Aminosäuren entwickelt werden.

Tabelle 47: Berechnete kovalente Volumina der Aminosäure

| Aminosäure | Volumen [\AA^3] | Aminosäure | Volumen [\AA^3] |
|------------|----------------------------|--------------|----------------------------|
| Alanin | 7.303 ± 0.112 | Leucin | 12.874 ± 0.168 |
| Arginin | 15.871 ± 0.203 | Lysin | 14.438 ± 0.231 |
| Asparagin | 10.633 ± 0.131 | Methionin | 15.689 ± 0.289 |
| Aspartat | 10.322 ± 0.130 | Prolin | 11.016 ± 0.167 |
| Cystein | 12.037 ± 0.263 | Phenylalanin | 16.053 ± 0.233 |
| Glutamin | 12.257 ± 0.170 | Serin | 8.490 ± 0.110 |
| Glutamat | 12.023 ± 0.155 | Threonin | 10.399 ± 0.147 |
| Glycin | 5.410 ± 0.087 | Tryptophan | 20.284 ± 0.253 |
| Histidin | 14.120 ± 0.132 | Tyrosin | 17.274 ± 0.227 |
| Isoleucin | 12.912 ± 0.214 | Valin | 11.038 ± 0.169 |

In Tabelle 47 sind die berechneten kovalenten Volumina der zwanzig kanonischen Aminosäuren mit ihren jeweiligen Standardabweichungen zusammen gefasst. Im Zuge der Berechnung dieser Werte stellte sich die Verwendung der publizierten Radien (Tabelle 46) in manchen Fällen jedoch als problematisch heraus. Vergleicht man die berechneten Bindungslängen (Vergleiche Kapitel 2.4.1 und elektronischer Anhang) mit den kovalenten Radien aus Tabelle 46, so fällt auf, dass vor allem bei Kohlenstoffatomen, die eine Einfach- und eine Doppel-Bindung ausbilden, z.B. die C γ -Atome von aromatischen Ringen, der angegebene Radius nicht groß genug ist, um eine Berührung der Orbitale zur Ausbildung einer Einfachbindung zu ermöglichen. Um die nötige Berührung der beiden Atome [237] zu ermöglichen, musste deshalb eine Anpassung der Radien der einzelnen Atome vorgenommen werden (Abb. 30).

Strukturbiologische Parameter

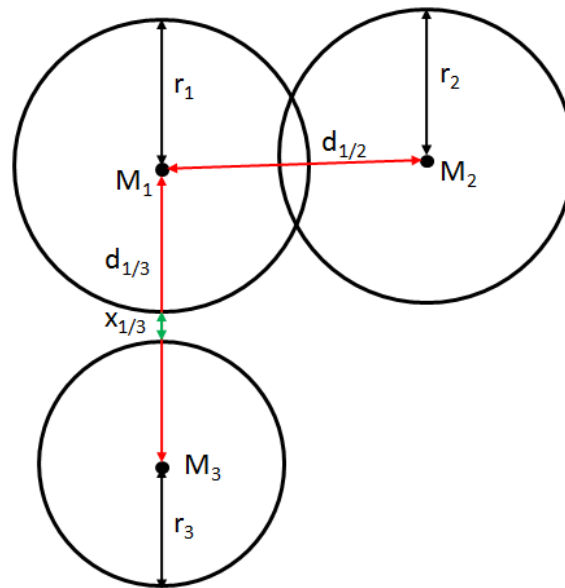


Abb. 30: Schema zur Anpassung kovalenter Radien bei der Volumenberechnung

Die Abbildung zeigt schematisch die Bindung von drei aufeinanderfolgenden Atomen M_1 , M_2 und M_3 , mit den dazugehörigen Bindungslängen $d_{1/2}$ und $d_{1/3}$. Im Falle, dass ein sp^2 -hybridisierter-Kohlenstoff eine Einfach- und eine Doppelbindung aufweist, ist der postulierte kovalente Radius für dieses Atom [230] kürzer als die Einfachbindung $d_{1/3}$ und daraus resultiert eine Lücke zwischen den beiden Atomen ($x_{1/3}$).

Dabei werden für alle kovalenten Bindungen innerhalb der Proteinstruktur bzw. der Aminosäure die Bindungslängen berechnet, und im Anschluss daran die dazugehörigen kovalenten Radien bestimmt. Dabei wird zuerst die Differenz aus der berechneten Bindungslänge, z.B. 1,532 Å für CA-CB von Aspartat, und der theoretischen Bindungslänge (Summe der kovalenten Radien der beteiligten Atome), z.B. 0,76 Å + 0,76 Å = 1,52 Å, gebildet. Ausgehend von dieser Differenz, hier 0,012 Å, kann nun berechnet werden, wie sehr der jeweilige Radius angepasst werden muss. Dazu muss zuerst das Verhältnis der beiden Literaturwerte zueinander bestimmt, in unserem Fall 1:1, d.h. beide Radien müssen gleichermaßen angepasst werden, und dann die Differenz entsprechend dieses Verhältnisses auf die jeweiligen kovalenten Radien addiert werden. Im vorliegenden Fall müsste also 0,006 Å auf die Literaturwerte addiert werden, so dass man zu realen kovalenten Radien von je 0,766 Å kommt. Diese Berechnung wird für alle Bindungen eines Atoms wiederholt, so dass z.B. ein Atom mit drei Bindungen am Ende drei kovalente Radien zugewiesen bekommt. Da für alle Bindungen eine Berührung der Orbitale vorhanden sein muss [237], wird das Volumen des Atoms dann aus dem maximalen angepassten kovalenten Radius r berechnet.

Dieses Volumen entspricht dem Volumen einer Kugel mit dem Radius r , abzüglich aller Kugelschnitte die durch die Überlappung entlang der Bindungen entstehen (Abb. 31).

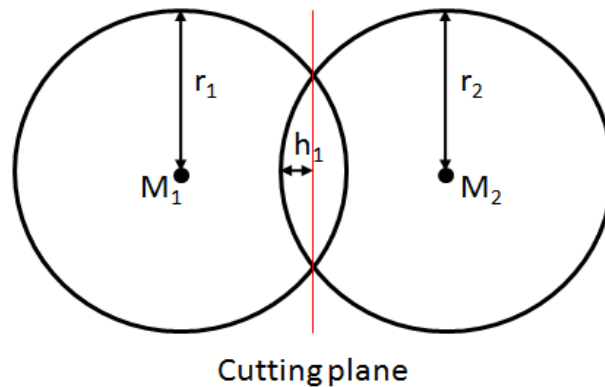


Abb. 31: Schematische Darstellung der Volumen-Berechnung überlappender Kugeln

Das Schema zeigt die für die Berechnung des Volumens von zwei überlappenden Kugeln nötigen Werte. Mit Hilfe von einfachen mathematischen Berechnungen kann das Volumen der Kugelkappen bestimmt und von den Volumina der Kugeln subtrahiert werden.

Aufgrund der Nutzung des kovalenten Volumens im Gegensatz zum van-der-Waals-Volumen in früheren Arbeiten, ist ein Vergleich der Ergebnisse nur in Form von normierten Volumina möglich (Tabelle 48). Grundsätzlich sind die Ergebnisse der verschiedenen Analysen sehr ähnlich, mit Ausnahme der Ergebnisse von Finney, obwohl verschiedene Methoden zum Einsatz kamen. Während Pontius *et al.* [227] sowie Finney [225] den klassischen Voronoi-Ansatz [223] nutzten, verwendeten Harpaz *et al.* [238] sowie Chotia [239] eine angepasste Version des Voronoi-Ansatzes unter Beachtung der van-der-Waals-Radien der einzelnen Atome [224]. Richards begründete die Anpassung des Voronoi-Ansatzes mit dem sogenannten Vertex-Fehler [224, 228], jedoch konnten Gerstein *et al.* mit Hilfe von Moleküldynamik-Simulationen mit ENCAD zeigen, dass dieser Fehler nur sehr geringe Auswirkungen auf die Berechnung der Volumina hat [240]. Die starken Abweichungen der Ergebnisse von Finney resultieren aus der Einbeziehung von Wasser-Molekülen in die Berechnung, mit deren Hilfe angepasste Voronoi-Polyhedra berechnet wurden [225]. Alles in allem entspricht das Verhalten der hier berechneten kovalenten Volumina den Ergebnissen von früheren Berechnungen von van-der-Waals-Volumina.

Tabelle 48: Vergleich von normierten Aminosäure-Volumina verschiedener Studien

In der Tabelle sind die auf Glycin normierten Volumina von vier älteren Studien sowie den Ergebnissen der vorliegenden Arbeit gezeigt. Dazu wurde für jede Aminosäure das berechnete Volumen durch das Volumen von Glycin geteilt, um einen Vergleich von van-der-Waals- und kovalentem Volumen zu ermöglichen. Die Zahlen in Klammern sind die tatsächlich berechneten Volumina der einzelnen Aminosäuren. Bei der Studie von Finney [225] wurde das theoretische Volumen als Summe von Volumina des Proteinerückgrates und der entsprechenden Seitenkette genutzt.

| Aminosäure | Harpaz [238] | Chotia [239] | Finney [225] | Pontius [227] | XRD |
|--------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|--------------------------------|
| Alanin | 1,41 (90,1 Å ³) | 1,38 (91,5 Å ³) | 1,52 (85,8 Å ³) | 1,36 (91,5 Å ³) | 1,35 (7,3 Å ³) |
| Arginin | 3,02 (192,8 Å ³) | | 3,73 (209,8 Å ³) | 2,91 (196,1 Å ³) | 2,93 (15,9 Å ³) |
| Asparagin | 2,00 (127,5 Å ³) | 2,04 (135,2 Å ³) | 2,31 (130,2 Å ³) | 2,05 (138,3 Å ³) | 1,97 (10,6 Å ³) |
| Aspartat | 1,84 (117,1 Å ³) | 1,88 (124,5 Å ³) | 2,34 (131,6 Å ³) | 2,00 (135,2 Å ³) | 1,91 (10,3 Å ³) |
| Cystein | 1,78 (113,3 Å ³) | 1,77 (117,7 Å ³) | 1,83 (103,1 Å ³) | 1,69 (114,4 Å ³) | 2,22 (12,0 Å ³) |
| Glutamin | 2,34 (149,4 Å ³) | 2,43 (161,1 Å ³) | 2,74 (154,4 Å ³) | 2,32 (156,4 Å ³) | 2,27 (12,3 Å ³) |
| Glutamat | 2,21 (140,8 Å ³) | 2,34 (155,1 Å ³) | 2,70 (151,8 Å ³) | 2,29 (154,6 Å ³) | 2,22 (12,0 Å ³) |
| Glycin | 1,00 (63,8 Å ³) | 1,00 (66,4 Å ³) | 1,00 (56,3 Å ³) | 1,00 (67,5 Å ³) | 1,00 (5,4 Å ³) |
| Histidin | 2,50 (159,3 Å ³) | 2,52 (167,3 Å ³) | 2,77 (156,2 Å ³) | 2,42 (163,2 Å ³) | 2,61 (14,1 Å ³) |
| Isoleucin | 2,58 (164,9 Å ³) | 2,54 (168,8 Å ³) | 2,88 (162,4 Å ³) | 2,41 (162,6 Å ³) | 2,39 (12,9 Å ³) |
| Leucin | 2,58 (164,6 Å ³) | 2,53 (167,9 Å ³) | 2,80 (157,7 Å ³) | 2,42 (163,4 Å ³) | 2,38 (12,9 Å ³) |
| Lysin | 2,66 (170,0 Å ³) | 2,58 (171,3 Å ³) | 3,04 (171,1 Å ³) | 2,41 (162,5 Å ³) | 2,67 (14,4 Å ³) |
| Methionin | 2,63 (167,7 Å ³) | 2,57 (170,8 Å ³) | 2,85 (160,2 Å ³) | 2,46 (165,9 Å ³) | 2,90 (15,7 Å ³) |
| Prolin | 1,93 (123,1 Å ³) | 1,95 (129,3 Å ³) | 2,23 (125,6 Å ³) | 1,83 (123,4 Å ³) | 2,04 (11,0 Å ³) |
| Phenylalanin | 3,03 (193,5 Å ³) | 3,06 (203,4 Å ³) | 3,62 (203,9 Å ³) | 2,95 (198,8 Å ³) | 2,97 (16,1 Å ³) |
| Serin | 1,48 (94,2 Å ³) | 1,49 (99,1 Å ³) | 1,80 (101,1 Å ³) | 1,51 (102,0 Å ³) | 1,57 (8,5 Å ³) |
| Threonin | 1,88 (120,0 Å ³) | 1,84 (122,1 Å ³) | 2,20 (123,9 Å ³) | 1,87 (126,0 Å ³) | 1,92 (10,4 Å ³) |
| Tryptophan | 3,63 (231,7 Å ³) | 3,58 (237,6 Å ³) | | 3,51 (237,2 Å ³) | 3,75 (20,3 Å ³) |
| Tyrosin | 3,09 (197,1 Å ³) | 3,07 (203,6 Å ³) | 3,46 (194,9 Å ³) | 3,11 (209,8 Å ³) | 3,19 (17,3 Å ³) |
| Valin | 2,18 (139,1 Å ³) | 2,13 (141,7 Å ³) | 2,25 (126,4 Å ³) | 2,05 (138,4 Å ³) | 2,04 (11,0 Å ³) |

Zusätzlich zur Berechnung von Aminosäure-Volumina wurde in der vorliegenden Arbeit das *hard-sphere*-Modell genutzt um sterische Probleme zu bestimmen. Dazu wurden die van-der-Waals-Radien der Atome genutzt [232] und alle nicht-kovalent gebundenen Atome (mehr als vier Bindungen voneinander entfernt), deren Kugeln sich mehr als 0,4 Å überlappen, als sterische Hinderungen definiert [241]. Analog zu Davis *et al.* [77] kann dann ein *Clashscore* berechnet werden, definiert als Anzahl an sterischen Problemen pro 1000 Atome, um eine Aussage über die sterische Qualität der untersuchten Struktur treffen zu können. Nicht-kovalent gebundene Atome, deren kovalente Radien sich überlappen, wurden zusätzlich als massive sterische Probleme definiert.

2.4.4 Energiepotentiale

Alle in der vorliegenden Arbeit etablierten Energiepotentiale wurden mit Hilfe des Ansatzes von Melo und Feytmans [242] berechnet. Dabei kann die Energie für eine Interaktion mit Hilfe von Formel (8) berechnet werden. Die so erhaltene Energie ist dann abhängig vom Interaktionspaar i, j , der Distanzklasse l und der Sequenzseparation k .

$$\Delta E_k^{i,j}(l) = RT \ln(1 + M_{i,j,k} \sigma) - RT \ln\left[1 + M_{i,j,k} \sigma \frac{f_k^{i,j}(l)}{f_k^{xx}(l)}\right] \quad (8)$$

Die darin vorkommenden Variablen M_{ijk} , $f_k^{ij}(l)$ und $f_k^{xx}(l)$ können dann mit Hilfe der Formeln (9) bis (11) berechnet werden. Für den Faktor RT wurde 0,582 kcal/mol gewählt. Die darin vorkommende Variable $f(i, j, k, l)$ entspricht der Anzahl an gefundenen Interaktionen von Atom i und Atom j bei einem Abstand l und der Sequenzseparation k . In Formel (9) werden alle Interaktionen des Paares i, j (z.B. ALA und ARG) bei einer bestimmten Sequenzseparation (z.B. $k=5$) für alle Distanzklassen l aufaddiert, um so die Gesamtzahl an Interaktionen eines bestimmten Paares zu bekommen.

$$M_{i,j,k} = \sum_{l=1}^n f(i, j, k, l) \quad (9)$$

Mit Hilfe der Variable M_{ijk} kann nun die Interaktion des Paares i, j bei einer Sequenzseparation k innerhalb einer bestimmten Distanzklasse l anhand von Formel (10) normiert werden.

Strukturbiologische Parameter

$$f_k^{i,j}(l) = \frac{f(i,j,k,l)}{M_{i,j,k}} \quad (10)$$

So erhält man die relative Häufigkeit des Auftauchens einer bestimmten Interaktion innerhalb der aktuellen Distanzklasse. Formel (11) beschreibt dann die relative Häufigkeit aller Interaktionen in der aktuellen Distanzklasse l.

$$f_k^{xx}(l) = \frac{\sum_{i=1}^n \sum_{j=1}^n f(i,j,k,l)}{\sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n f(i,j,k,l)} \quad (11)$$

Dazu wird die Summe aller Interaktionen innerhalb der Distanzklasse l (Zähler) durch die Gesamtzahl aller Interaktionen (Nenner) geteilt. Unter Nutzung des von Melo und Feytmans [242] entwickelten Ansatzes (siehe oben) wurden verschiedene Potentiale entwickelt, mit deren Hilfe eine energetische Validierung von Proteinstrukturen ermöglicht werden sollte. Neben klassischen Interaktionspotentialen wurden dabei auch weitere Formen getestet, die nachfolgend erläutert und diskutiert werden sollen. Es wurde grundsätzlich versucht, auf schon bestehende Berechnungen (Abschnitt Intramolekulare Interaktionen) zurückzugreifen, um eine Verknüpfung der Wechselwirkungsanalyse mit den Bewertungsalgorithmen zu schaffen.

Tabelle 49: Standard-Parameter der entwickelten Energiepotentiale

| Potential | l | k | σ |
|--|-------|---|----------|
| Cα-Interaktionspotential | 8,5 Å | 4 | 1/50 |
| Cβ-Interaktionspotential | 7,5 Å | 4 | 1/50 |
| Cα-Sekundärstrukturabhängiges Interaktionspotential | 7,5 Å | 3 | 1/50 |
| Cβ-Sekundärstrukturabhängiges Interaktionspotential | 7,5 Å | 3 | 1/50 |
| All-Atom-Interaktionspotential | 7,5 Å | 6 | - |
| Cα-Lösungsmittelpotential | 8,0 Å | - | - |
| Cβ-Lösungsmittelpotential | 8,0 Å | - | - |

Um eine Bewertung mit Hilfe von Interaktionspotentialen zu ermöglichen, müssen die Parameter für Distanzklasse l, Sequenzseparation k sowie den Anpassungsfaktor σ ermittelt werden. Für die einfachen C α /C β -Interaktionspotentiale wurde ein Anpassungsfaktor von 1/50, eine Sequenzseparation von 4 Aminosäuren sowie eine maximale Distanz von 8,5 Å für das C α -Interaktionspotential bzw. 7,5 Å für das C β -Interaktionspotential (Tabelle 49) genutzt.

Strukturbiologische Parameter

Neben den einfachen Aminosäure-abhängigen Interaktionspotentialen wurden noch Sekundärstruktur-abhängige Interaktionspotentiale entwickelt. Dabei kamen zwei verschiedene Systeme, ein Drei-Zustandsmodell (Helix, Strang und *Coil*) sowie ein Vier-Zustands-Modell (Helix, Strang, *Turn* und *Coil*) zum Einsatz, um den neu-definierten *Turns* Rechnung tragen zu können. Unter Nutzung der selben Sequenzseparationen, Anpassungsfaktoren und fast identischen maximal-Distanzen (Tabelle 49), konnten insgesamt vier Potentiale entwickelt werden.

Ausgehend von den $C\alpha$ - und $C\beta$ -Potentialen wurde ein *All-Atom*-Interaktionspotential analog zu ANOLEA [242] entwickelt. Dabei wurden alle Atomarten der kanonischen Aminosäuren innerhalb von Proteinstrukturen als eigener Atomtyp angesehen, dementsprechend repräsentieren z.B. ASN_OD1 und ASN_OD2 unterschiedliche Atomtypen. Als Parameter wurden eine Sequenzseparation von 6 Aminosäuren, eine Maximal-Distanz von 7,5 Å sowie ein Anpassungsfaktor von 1/50 gewählt.

Mit Hilfe eines vereinfachten Berechnungsansatzes (Formel (12)) wurden Solvatationspotentiale entwickelt. Dabei wurde sich zunutze gemacht, dass die Solvation einer Aminosäure durch die Anzahl an anderen Aminosäuren in ihrer direkten Umgebung bestimmt werden kann [143, 243]. Ausgehend von dieser Annahme wurden zwei Solvatations-Potentiale anhand von Formel (12) erstellt. Die Anzahl der Aminosäuren in der Umgebung wurde dabei entweder anhand von $C\alpha$ - oder $C\beta$ -Atomen bestimmt.

$$\Delta E(r) = -RT \ln\left(\frac{f^{AA}(r)}{f(r)}\right) \quad (12)$$

Dabei ist $f(r)$ die Wahrscheinlichkeit aller Aminosäuren eine bestimmte Anzahl "r" an umgebenden Aminosäuren zu besitzen und $f^{AA}(r)$ ist dieselbe Wahrscheinlichkeit des Aminosäuretyps AA [143, 243].

$$f(r) = \frac{\sum_{ALA}^{VAL} n^{AA}(r)}{\sum_{ALA}^{VAL} \sum_{r=0}^{20} n^{AA}(r)} \quad (13)$$

Der erste Term berechnet sich dabei aus der Anzahl der gezählten Aminosäuren mit Nachbaratomen "r" geteilt durch die Anzahl aller gezählten Aminosäuren.

$$f^{AA}(r) = \frac{n^{AA}(r)}{\sum_{r=0}^{20} n^{AA}(r)} \quad (14)$$

Der zweite Term (Formel (14)) berechnet sich dann aus der Anzahl von Aminosäuretyp AA mit "r" Nachbarn geteilt durch die Anzahl aller gefundenen Aminosäuren vom Typ AA.

2.4.5 Überprüfung von Sekundärstrukturmerkmalen

Obwohl Sekundärstrukturen essentiell für die Faltung von Proteinen sind [5], wurden bisher nur wenige Versuche unternommen ihre strukturelle Qualität zu überprüfen. Die meisten Untersuchungen in diesem Bereich beschränken sich entweder auf die Übereinstimmung von vorhergesagter und modellierter Sekundärstruktur [244, 245], die Qualität von Sekundärstrukturvorhersagen [246–248] oder die Entwicklung von energetischen Funktionen mit Hilfe von Sekundärstrukturen [249, 250]. Nachdem alle bekannten Validierungskriterien analysiert und erweitert wurden, wurden deshalb in einem letzten Test Bewertungsparameter für Sekundärstrukturmerkmale etabliert. Dabei wurden vor allem neue geometrische Untersuchungen durchgeführt.

Tabelle 50: Geometrische Parameter zur Validierung von Sekundärstrukturmerkmalen

| Parameter | Beschreibung | Erlaubter Bereich |
|---------------------|--|--|
| CAp3-Distanz | Mittelwert der CA-CA+3 Abstände über das gesamte Merkmal | Helices: 2.52Å – 2.74Å Strands: 4.88Å – 5.16Å |
| Streckung | Mittelwert der CA-CA+2 Abstände über das gesamte Merkmal | Helices: 5.38Å – 5.59Å Strands: 6.35Å – 7.05Å |
| Vektorwinkel | Mittelwert der Winkel zwischen dem Vektor über das Merkmal und den Vektoren zwischen zwei aufeinanderfolgenden C α -Atomen im Merkmal | Helices: 62.89° – 66.77° Strands: 24.53° – 33.45° |
| Steigung | Quotient aus Länge der Helix und Anzahl der darin vorkommenden Aminosäuren | Helices: 1.33Å – 1.51Å |
| CB-Torsion | Torsionswinkel CB-CA-CA-CB von zwei aufeinanderfolgenden Aminosäuren im Merkmal | Helices: 55.24° – 76.26° Strands: 126.21° – 175.71° |
| Rotation | Berechnet als $\arccos((1-4*\cos^2((\Phi+\Psi)/2))/3)$ für jede Aminosäure in der Helix [175] | Helices: 94.11° – 107.11° |

Bei der Etablierung von geometrischen Kriterien für die Validierung von Sekundärstrukturen musste darauf geachtet werden, Parameter zu nutzen, die nicht im Zuge der Zuweisung von Sekundärstrukturen (Abschnitt 6.3) angewendet wurden. Dementsprechend wurden Abstands- und Winkel-Berechnungen durchgeführt, um aussagekräftige und gut differenzierende Kriterien zu etablieren. Nach ausführlichen Berechnungen und Tests konnten sechs Kriterien für Helices sowie vier Kriterien für β -Stränge identifiziert werden, mit denen eine Validierung möglich ist (Abb. 32, Abb. 33 und Tabelle 50).

Strukturbiologische Parameter

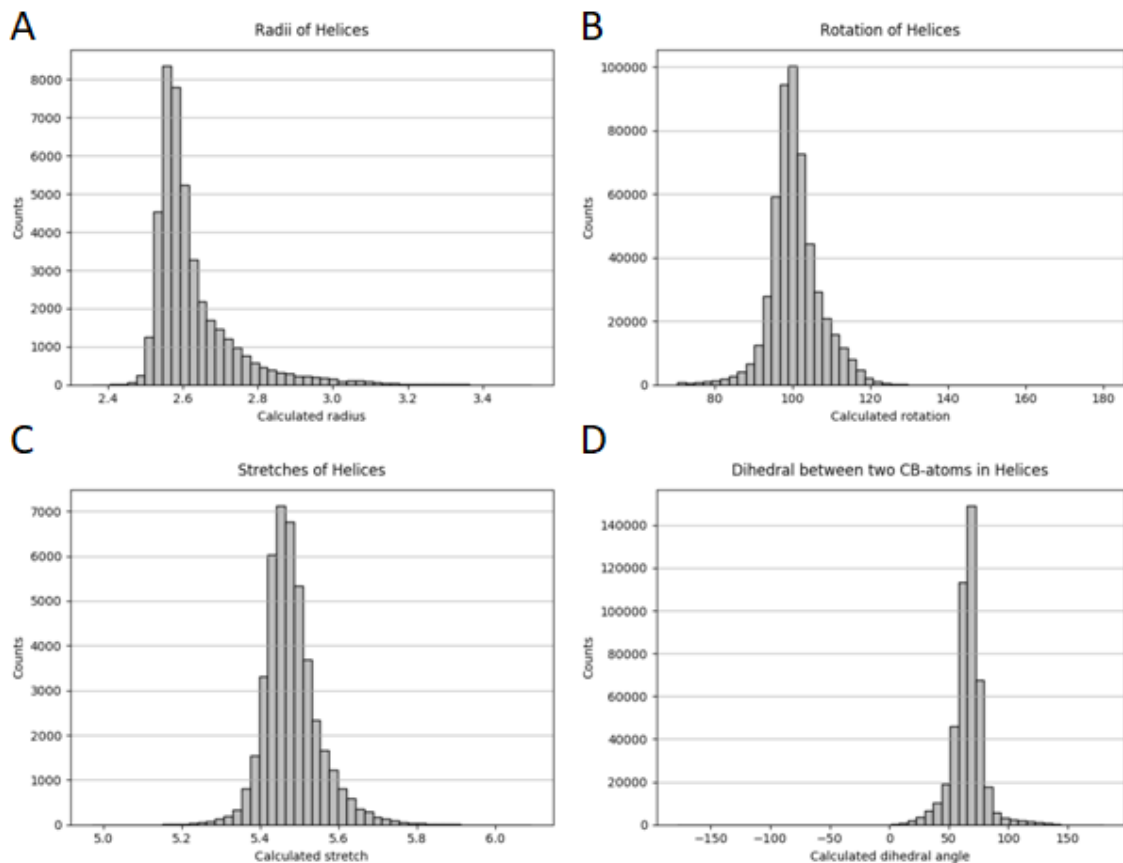


Abb. 32: Geometrische Kriterien zur Validierung von Helices

Die vier Histogramme zeigen die Verteilung von vier geometrischen Parametern, die zur Validierung von Helices herangezogen werden. Darunter sind die CAp3-Distanz (A), die Rotation (B) und die "Streckung" (C) einer Helix sowie der Torsionswinkel zwischen zwei aufeinanderfolgenden C β -Atomen innerhalb einer Helix (D) (Vergleiche Tabelle 50). Alle vier Kriterien weisen schmale Verteilungen mit klar zu erkennenden Maxima auf.

Vor allem die Abstandskriterien CAp3-Distanz und Streckung differenzieren sehr stark zwischen gut und schlecht gefalteten Helices oder β -Strängen, da ihre Verteilungen in den meisten Fällen sehr schmal sind und klar erkennbare Maxima besitzen (Abb. 32 und Abb. 33 A bzw. C). Auch die Rotation von Aminosäuren [175] wurde für Helices sowie β -Stränge berechnet (B) und kann zur Validierung von Helices herangezogen werden. Bei den Strängen zeigte sich, dass die angenommenen Phi/Psi-Winkel sehr stark schwanken und somit eine Nutzung dieses Parameters hier nicht möglich ist. Aufgrund der Regelmäßigkeit von Sekundärstrukturen wurden auch verschiedene Torsionswinkel über mehrere Aminosäuren innerhalb der Merkmale getestet. Dabei konnte zumindest ein positives Ergebnis in Form des CB-CA-CA-CB-Winkels gefunden werden (D). Das Besondere an diesem Kriterium ist, dass die

Strukturbiologische Parameter

Maxima der beiden Verteilungen (Helix oder β -Strang) sehr unterschiedliche Werte annehmen und somit spezifisch für das jeweilige Merkmal sind.

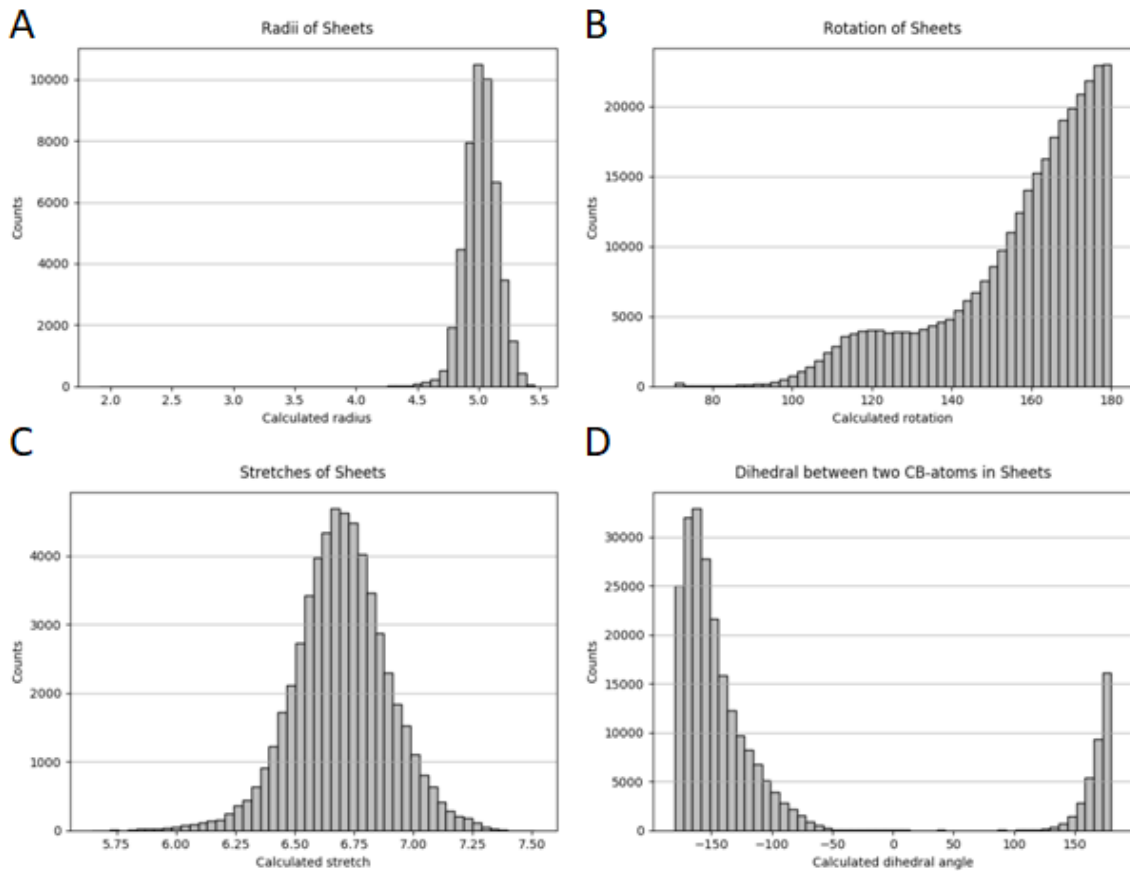


Abb. 33: Mögliche Validierungskriterien von β -Strängen

Die vorliegende Abbildung zeigt die Histogramme von vier geometrischen Kriterien, die im Zuge der Suche nach Parametern für die Validierung von β -Strängen analysiert wurden. Zu sehen sind die CAp3-Distanz (A), die "Rotation" (B) und die Streckung (C) eines β -Stranges sowie der Torsionswinkel zwischen zwei aufeinanderfolgenden C β -Atomen innerhalb eines Strangs (D).

3. Informatische Implementierung

Unter Verwendung der oben beschriebenen Parameter und Berechnungsalgorithmen konnten verschiedenen Analyse- und Bewertungsprogramme entwickelt werden. Diese Programme wurden in zwei verschiedene Funktionsserver integriert, die im Folgenden kurz beschrieben werden sollen. Zusätzlich wurden noch drei Daten-Server etabliert, die eine ausführliche Dokumentation, statistische Auswertungen der einzelnen Datensätze sowie sämtliche genutzten Parameter in übersichtlicher Form aufbereiten.

3.1 Die Funktionsserver VAProTS und VAProIN

Im Zuge dieser Arbeit wurden die beiden Funktionsserver VAProTS (*Validation and Analysis of Protein Tertiary Structures*) und VAProIN (*Validation and Analysis of Protein Interaction Networks*) erstellt, mit deren Hilfe Proteinstrukturmodelle umfassend analysiert und validiert werden können. Beide Server bieten dem Nutzer eine Vielzahl an Auswahlmöglichkeiten, um die übergebenen Strukturen ganz nach seinen Wünschen zu analysieren. Des Weiteren bieten beide Server eine Vielzahl an verschiedenen Ausgaben, sowohl *online* als auch in Form von txt-, csv- oder pdf-Dateien an, die in Form eines Archivs herunterladbar sind und individuell ausgewählt werden können.

Der VAProTS-Server besteht aus drei verschiedenen Modulen, die eine Sekundärstrukturzuweisung, eine umfassende Analyse und eine ausführliche Validierung der Strukturen ermöglichen. Die Zuweisung von Sekundärstrukturen erfolgt dabei anhand des in der vorliegenden Arbeit entwickelten Mechanismus und kann individuell angepasst werden. Dazu stehen dem Nutzer drei Arten der Zuweisung zu Verfügung, ungefiltert, d.h. kurzer Zuweisungslauf ohne Anwendung von Mindestlängen, gefiltert, d.h. zu kurze Sekundärstrukturmerkmale werden ausgeschlossen, und detailliert, d.h. Durchführung von mehreren zusätzlichen Tests (Tabelle 53). Des Weiteren kann die Zuweisung der neu-definierten *Turn*-Typen deselektiert werden. Das zweite Modul erstellt eine umfassende Analyse der Proteinstruktur und nutzt dabei einen Großteil der oben beschriebenen Analyse Kriterien (Kapitel 2.2 Analytische Untersuchungen von Proteinstrukturen). Zusätzlich enthält es eine detaillierte Sekundärstrukturzuweisung, inkl. der neu-definierten *Turns*. Die einzelnen Kriterien können dabei vom Nutzer deaktiviert werden und ermöglichen so eine

spezifische Analyse der hochgeladenen Strukturen. Eine Auswahl der Standard-Einstellungen führt zum Start eines großen Validierungsmoduls, das zusätzlich sowohl das Analyse-Modul als auch das Zuweisungs-Modul mit Standard-Einstellungen beinhaltet. Die Validierung erfolgt in Form von fünf Unter-Modulen, die die oben beschriebenen Validierungskriterien (Abschnitt 2.4 Validierung von Proteinstrukturen) aufgreifen. Am Ende erhält der Nutzer für jedes Kriterium einen Zahlenwert, der die Gesamtqualität in Bezug auf dieses Kriterium, z.B. Torsionswinkel, repräsentiert, sowie mehrere Zahlenwerte, die einzelne Teile des Kriteriums, z.B. Omega-Winkel, repräsentieren. Sämtliche berechneten *Scores* werden dem Nutzer in Form von übersichtlichen Tabellen auf den Ergebnisseiten des Servers übergeben. Zusätzlich können noch verschiedene Datei-Ausgaben erstellt werden. Diese reichen von unformatierten, tabellarischen csv-Dateien über graphische Ausgaben in pdf-Dateien bis hin zu Zusammenfassungen und detaillierten Listen der Ergebnisse in Form von txt-Dateien.

Im Zuge der Implementierung von VAProTS stellte sich vor allem die übersichtliche Darstellung der verschiedenen gefundenen intramolekularen Interaktionen in Form von Interaktionskarten als problematisch dar. Aus diesem Grund wurden die verschiedenen Interaktions-Module von VAProTS (siehe auch Abschnitt 2.2 Intramolekulare Interaktionen) nochmals in einem zweiten Funktionsserver zusammengefasst und um Wasserstoffbrücken, hydrophobe Wechselwirkungen und Aminosäurekontakte erweitert. Mit Hilfe des Servers VAProIN können Nutzer sich nun interaktive Interaktionskarten erstellen lassen und online betrachten. Sämtliche gefundenen Wechselwirkungen werden dabei in Form von kleinen (50 Aminosäuren x 50 Aminosäuren) quadratischen Karten visualisiert. Dazu werden pro Bereich vier verschiedene Karten (Aminosäurekontakte, Wasserstoffbrücken, hydrophobe Wechselwirkungen sowie sonstige Wechselwirkungen) erstellt, in denen die jeweiligen Interaktionen durch kleine, farbige Quadrate gekennzeichnet werden (Abb. 34). Bewegt der Nutzer die Maus über ein solches Quadrat werden ihm detaillierte Informationen, z.B. welche Aminosäuren beteiligt sind, in Form eines *Tooltips* angezeigt. Zur besseren Orientierung innerhalb der Karten sind die x- und die y-Achse mit den Aminosäurenummern der ersten und letzten Aminosäure des aktuellen Bereichs sowie einer schematischen Sekundärstrukturzuweisung versehen. Komplette Interaktionskarten sind in Form von txt-Dateien als Archiv direkt von der Ergebnisseite herunterladbar.

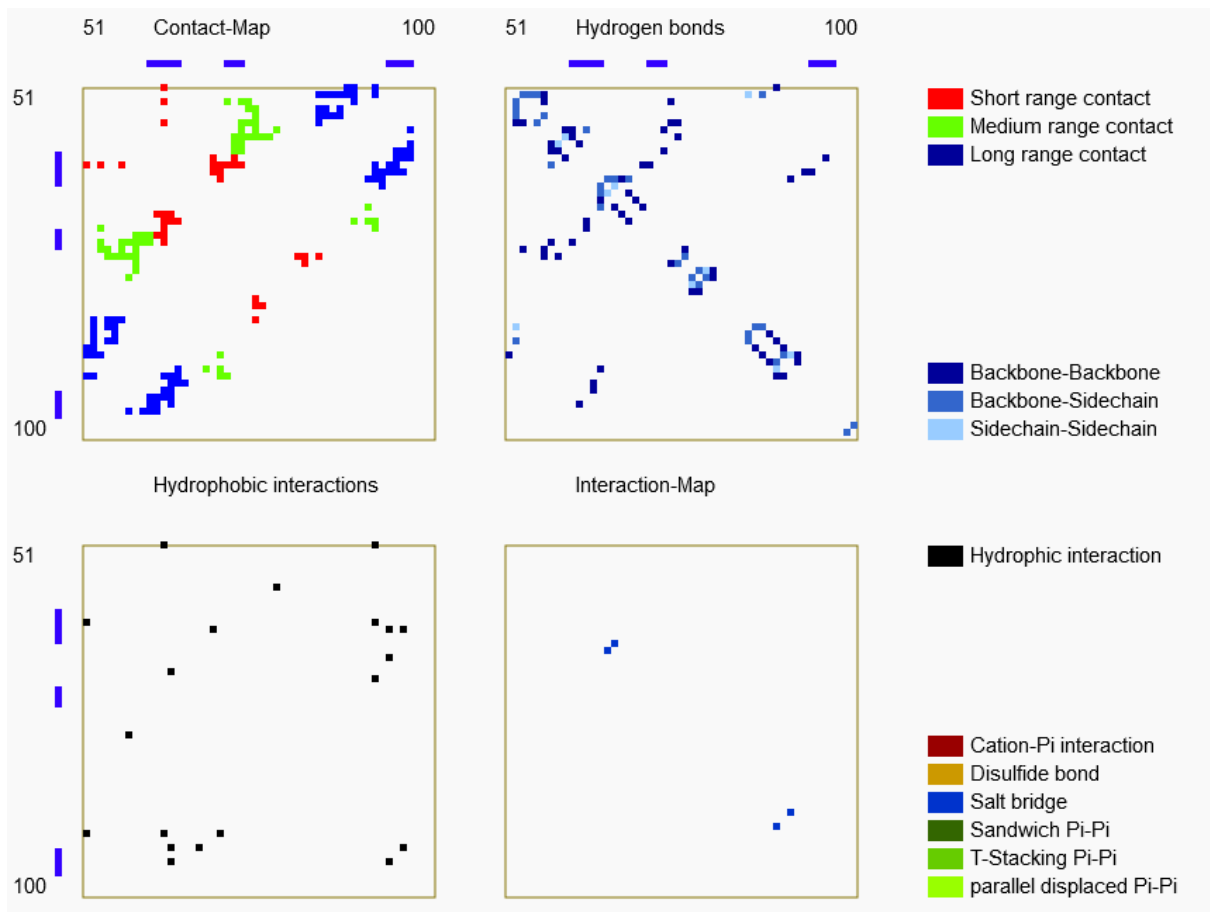


Abb. 34: Interaktive Interaktionskarten des Servers VAProIN

Die Abbildung zeigt die vier vom Server VAProIN erstellten interaktiven Karten für Aminosäurekontakte (links oben), Wasserstoffbrücken (rechts oben), hydrophobe Wechselwirkungen (links unten) und die restlichen Wechselwirkungen (rechts unten). Am rechten Rand befindet sich die dazugehörige Legende und die blauen Balken auf der x- und y-Achse stellen die schematische Sekundärstrukturzuweisung (in diesem Fall drei β -Stränge) dar. Zahlen am oberen und rechten Rand geben die genaue Position der Karten innerhalb der Proteinstruktur an. Durch die Gegenüberstellung aller Interaktionen ermöglichen diese Karten die Identifikation von miteinander wechselwirkenden Regionen innerhalb des Proteins.

3.2 Datenserver VADoc, VASat und VAPara

Um den Nutzern der beiden Funktionsserver möglichst viele Informationen über die zugrunde liegenden Algorithmen, Standard-Parameter sowie die Verwendung der Weboberfläche an die Hand zu geben, wurde eine ausführliche Dokumentation erstellt. Diese ist unter dem Namen VADoc (*Validation and Analysis Documentation*, vaprots.biochemtech.uni-halle.de/Wiki/Start.php) online verfügbar und enthält einzelne Bereiche für jeden Server im erstellten Netzwerk. Die einzelnen Seiten bestehen dabei aus

einer Textspalte und einer kleineren Spalte mit verschiedenen Bildern zur Illustration der verschiedenen Aspekte des jeweiligen Servers.

Da die Wahl der Parameter von entscheidender Bedeutung bei der Entwicklung von Bewertungsservern ist, wurden sämtliche berechneten und von den Funktionsservern genutzten Parameter auf der Webseite VAPara (*Validation and Analysis Parameters*, vaprots.biochemtech.uni-halle.de/Parameter/Start.php) bereitgestellt. Des Weiteren wurden csv-files und Archive der einzelnen Datensätze erstellt und können, genau wie die Parametersätze, von Nutzern heruntergeladen werden.

Zu guter Letzt wurden alle zur Erstellung der Funktionsserver genutzten Datensätze statistisch ausgewertet und ihre Ergebnisse online auf VASat (*Validation and Analysis Statistics*, vaprots.biochemtech.uni-halle.de/Statistics/Start.php) bereitgestellt. Ein großer Teil der berechneten Daten entfällt dabei auf die Aminosäureverteilungen, mit deren Hilfe etwaige Präferenzen innerhalb der einzelnen Sekundärstrukturen bestimmt werden konnten. Außerdem wurden sämtliche Strukturen noch einer kompletten Bewertung mit Hilfe von VAProTS unterzogen und die Ergebnisse der Analysen von Validierungen können ebenfalls eingesehen werden.

4. Zusammenfassung

In der vorliegenden Arbeit wurde ein Server zur Analyse, Bewertung und Verfeinerung von dreidimensionalen Proteinstrukturen entwickelt, der als Alternative zu bereits bekannten Validierungsprogrammen dienen kann. Durch die Bereitstellung verschiedener Analyse- und Bewertungsprogramme sollte damit die Validierung von Proteinstrukturen, die aktuell aufgrund von fehlenden Korrelationen und einer Vielzahl von schlecht dokumentierten Programmen relativ kompliziert ist, vereinfacht und umfassender gestaltet werden.

Ausgehend von verschiedenen Datensätzen bestehend aus hochaufgelösten Kristallstrukturen aus der PDB mussten dafür verschiedene grundlegende Parameter berechnet und überprüft werden. Beginnend mit einfachen Analyse-Kriterien wie Aminosäurezusammensetzungen von Proteinstrukturen konnte eine qualitative Untersuchung von Proteinen im Allgemeinen ermöglicht werden. Durch die Etablierung von Analysefunktionen für *cis*-Peptidbindungen, D-Aminosäuren sowie verschiedene Typen von intramolekularen Wechselwirkungen, war es möglich, umfassende Analysen von Proteinstrukturen durchzuführen.

Da sie essentiell sowohl für die Faltung als auch die Stabilität von Proteinstrukturen sind, wurde ein neuer Mechanismus zur Zuweisung von Sekundärstrukturen entwickelt. Die schrittweise Identifizierung von Helices und β -Strängen sorgt für eine genauere Abgrenzung zwischen ungeordneten Regionen und den Sekundärstrukturelementen. Die Übertragung der klassischen *tight Turn* Definition auf Regionen zwischen Sekundärstrukturelementen ermöglichte eine Neu-Definition von 13 *Turn*-Typen, deren Relevanz anhand von Phi/Psi-*Clustern* und statistischen Daten nachgewiesen werden konnte.

Um eine umfassende Validierung von Proteinstrukturen zu ermöglichen, ist es notwendig, Standard-Parameter ausgehend von bekannten Strukturen zu ermitteln. Die erhaltenen Ergebnisse für einfache geometrische Parameter konnten verdeutlichen, dass, obwohl es nur geringe Unterschiede zwischen den bereits publizierten Werten und den hier berechneten Werten gab, eine Überprüfung der alten Daten sinnvoll ist. Durch die Erweiterung der Bewertungsroutinen auf Aminosäure-Seitenketten konnte zudem ein Mehrwert im Bezug auf die geometrische Bewertung von Proteinstrukturen gewonnen werden.

Zusammenfassung

Durch die Anwendung einer Dichte-abhängigen Glättungsfunktion bei der Erstellung von Phi/Psi-, Chi1/Chi2-, Chi2/Chi3- und Chi3/Chi4-Diagrammen konnten insgesamt 22 neue Validierungsroutinen für Torsionswinkel etabliert werden. Des Weiteren konnten durch die Nutzung eines *hard-sphere* Modells sowohl kovalente Aminosäure-Volumina berechnet als auch eine Suche nach sterischen Hinderungen etabliert werden. Durch die Definition von massiven sterischen Problemen konnte außerdem eine Erweiterung der bisherigen Analyse von ungünstigen Kontakten innerhalb von Proteinstrukturen eingeführt werden.

Die Berechnung von Energie-Funktionen mit Hilfe eines wissensbasierten Ansatzes konnte genutzt werden, um sieben neue Potentiale zu entwickeln. Durch die Einbindung von Sekundärstruktur-Abhängigkeiten, verschiedenen Interaktionszentren und der Lösungsmittelzugänglichkeit ist eine komplexe, energetische Überprüfung möglich.

Mit Hilfe von bereits etablierten und neuen geometrischen sowie energetischen Parametern konnte eine qualitative und quantitative Bewertung von einzelnen Sekundärstrukturelementen ermöglicht werden. Diese neue Form der Bewertung erlaubt sogar Rückschlüsse auf die Lage, Faltung und Umgebung von Sekundärstrukturen.

Unter Nutzung der berechneten Standard-Parameter konnten zwei Funktionsserver sowie drei Datenserver implementiert werden, mit deren Hilfe eine umfassende Analyse und Bewertung von Proteinstrukturen möglich ist. Die erstellten Interaktionskarten sowie die umfassenden Bewertungsergebnisse der beiden Server VAProTS und VAProIN ermöglichen es Nutzern, problematische Bereiche in ihren Proteinmodellen zu finden. Durch die Erstellung einer umfangreichen Dokumentation, der Bereitstellung aller berechneten Standard-Parameter sowie die statistische Auswertung aller Datensätze wurde der Grundstein für ein umfassendes Server-Netzwerk gelegt.

Zusammenfassend wurden in der vorliegenden Arbeit Server zur Analyse, Validierung und Verfeinerung von Proteinstrukturen erstellt. Durch die Nutzung vieler verschiedener Analyse-Parameter, inklusive übersichtlicher Karten zur Veranschaulichung von intramolekularen Wechselwirkungen, können fast alle wichtigen analytischen Aspekte von Proteinstrukturen schnell und einfach bestimmt werden. Die umfangreichen Validierungsprogramme für geometrische, sterische sowie energetische Aspekte liefern einen umfassenden Einblick in die Qualität der untersuchten Modelle. Durch die verschiedenen Datenserver, konnte der Beginn eines umfangreichen bioinformatischen Netzwerkes gelegt werden.

5. Ausblick

Die in der vorliegenden Arbeit etablierten Funktions- und Datenserver ermöglichen es Nutzern, ausführliche Validierungen sowie Analysen ihrer Proteinstrukturen durchzuführen. Trotzdem gibt es noch Kriterien, die bisher keine Verwendung innerhalb der Programme gefunden haben. So existiert bisher keine Analyse der Lösungsmittel-zugänglichen Oberfläche von Proteinen bzw. Aminosäuren [251]. Des Weiteren sollte versucht werden, van-der-Waals-Volumina der Aminosäuren zu berechnen, um mit ihrer Hilfe eine Differenzierung zwischen Aminosäuren im Inneren und an der Oberfläche von Proteinstrukturen zu ermöglichen [229]. Zwar ermöglichen die Server die parallele Validierung/Analyse von bis zu 30 Proteinstrukturen, eine Vergleichsmöglichkeit von Strukturen mit derselben Aminosäuresequenz auf struktureller Ebene fehlt jedoch. Dementsprechend wäre die Implementierung eines Servers zum qualitativen und quantitativen Vergleich von Proteinstrukturen, vor allem für Arbeitsgruppen, die sich mit *in silico* Modellierung von Proteinstrukturen beschäftigen, eine große Hilfe.

Im Bereich von Nukleinsäurestrukturen gibt es nur wenige Ansätze zur Validierung dieser Strukturen [221]. Aus diesem Grund wäre es wünschenswert, weitere Funktionsserver zu entwickeln, die sich mit der Validierung von Nukleinsäure-Strukturen, kleinen organischen Molekülen (Liganden) sowie Protein-Komplexen beschäftigen.

Des Weiteren könnten Erkenntnisse, die durch die Berechnung von Standard-Parametern für solche Strukturen gewonnen wurden, genutzt werden um mögliche Protein-Protein Interaktionen oder Wechselwirkungen zwischen Proteinen und/oder Liganden/Nukleinsäuren vorherzusagen bzw. zu validieren.

Zu guter Letzt könnten sämtliche Validierungs- und Analyse-Ergebnisse, die im Zuge der Entwicklung der einzelnen Server für experimentell bestimmte Proteinstrukturen gewonnen wurden, noch in eine umfassende Datenbank eingebettet werden, mit deren Hilfe die Qualität der Strukturen schnell und umfassend bestimmt werden kann. Dies wäre vor allem im Bereich der Homologie-Modellierung eine große Hilfe, da Nutzer schnell überprüfen können ob die von ihnen ausgewählten *Templates* für eine Modellierung nutzbar sind [215].

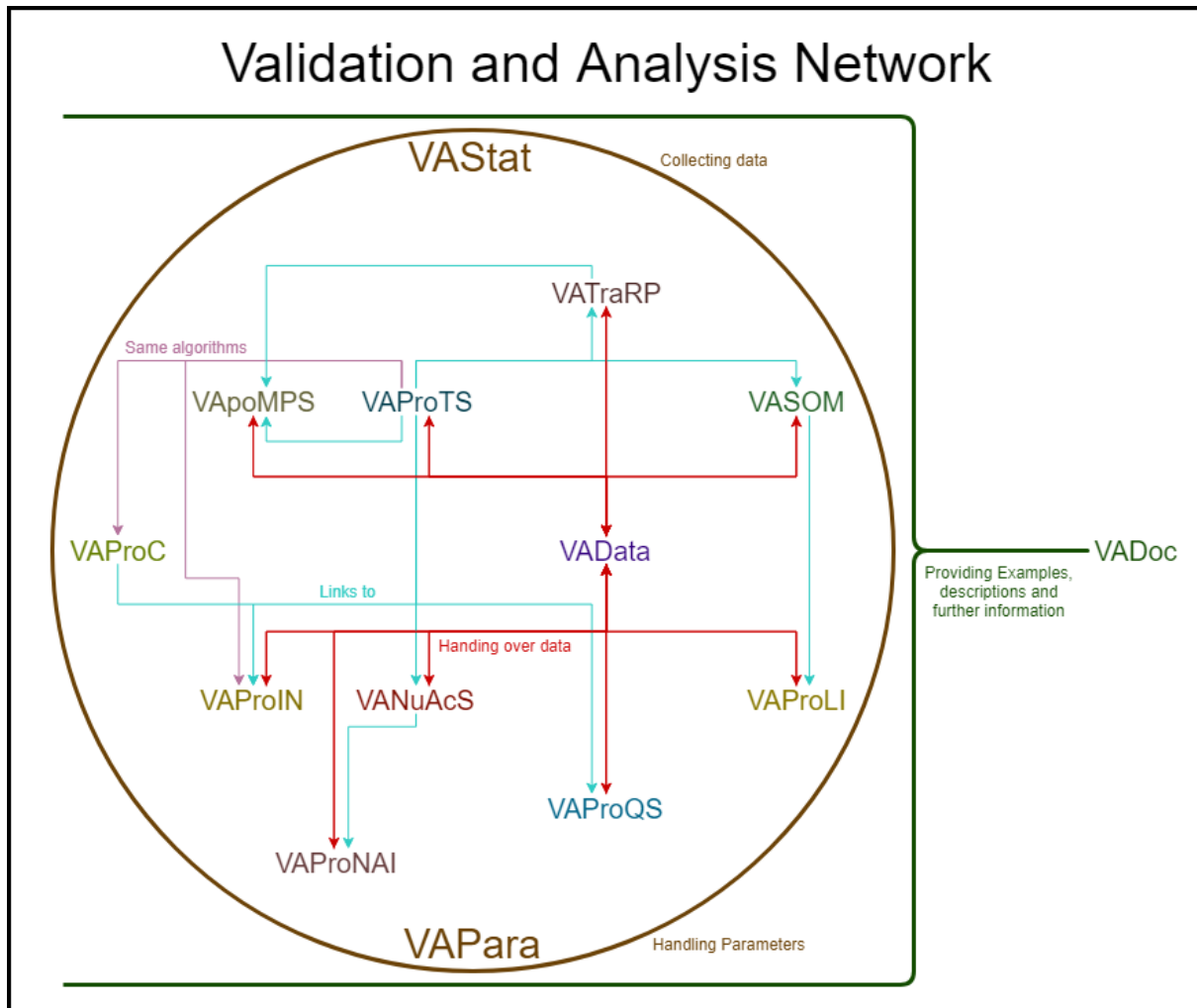


Abb. 35: Graphische Repräsentation des Bewertungsserver-Netzwerkes

Die Abbildung zeigt eine Zusammenfassung aller möglichen zukünftigen Funktions- und Daten-Server des zu entwickelnden Netzwerkes. Die Pfeile geben die Beziehungen der einzelnen Server/Programme zueinander an. Zu diesen Beziehungen gehören die Übergabe von Daten (rot), das gegenseitige Verlinken (hellblau) sowie das Teilen von Algorithmen (pink). Siehe Tabelle A.6 für die Bedeutung der einzelnen Abkürzungen, sowie kurze Beschreibungen der geplanten Server.

Ausgehend von den vielen Möglichkeiten, die im strukturbiologischen Bereich noch vorhanden sind, könnte ein umfassendes Netzwerk von Servern und Programmen entwickelt werden (Abb. 35), mit dessen Hilfe eine Vielzahl an bioinformatischen Routinen aus einer Hand angeboten werden können.

6. Material und Methoden

6.1 Datensätze

Insgesamt wurden in der vorliegenden Arbeit drei Datensätze verwendet, mit deren Hilfe die Standard-Parameter bestimmt werden konnten. Es wurde zum einen ein großer Datensatz XRay5000 (Suchdatum: 27.4.2017) erstellt, in dem mehr als 5000 Kristallstrukturen mit einer maximalen Sequenzidentität von 30% vorhanden waren. Zum anderen wurden der Trainings- und Test-Datensatz erstellt, die sich aus mehreren kleineren Datensätzen (Suchdatum: 20.5.2016) zusammensetzten (Tabelle 51 Spalten 2 bis 4). Dabei wurden nur Strukturen mit einer Kette in der asymmetrischen Einheit ausgewählt und per Zufall ein Zehntel der Strukturen ausgewählt. Diese zufällig ausgewählten Strukturen dienten als Test-Datensatz für die Sekundärstrukturzuweisung, während die restlichen Strukturen als Trainingsdatensatz für die Etablierung von Parametern für die Sekundärstrukturzuweisung genutzt wurden. Für alle Datensätze wurden monomere Proteine verwendet und Strukturen, in denen Atome oder Aminosäuren fehlten, wurden ausgeschlossen.

Tabelle 51: Suchkriterien in der PDB zur Erstellung der Datensätze

| Kriterium | Domänen | XRay100 | XRay500 | XRay1000 | XRay5000 |
|--------------------|----------|-----------|------------|----------|----------|
| Kettenlänge | 40 < 100 | 101 < 500 | 501 < 1000 | > 1000 | - |
| Auflösung | < 1.6 Å | < 1.3 Å | < 2.0 Å | < 2.0 Å | < 2 Å |
| R-free | ≤ 0.25 | ≤ 0.20 | ≤ 0.25 | - | ≤ 0.25 |
| Strukturen | 210 | 566 | 342 | 82 | 5545 |

Zusätzlich wurden zwei Modell-Datensätze erstellt, bestehend aus *in silico* generierten Modellen von Proteinen aus *Mycobacterium tuberculosis*. Dazu wurden alle Strukturen aus der *ModBase* (10158 Modelle) und dem *SwissModel Repository* (5446 Modelle) heruntergeladen (Stand: 3.6.2019), die aus diesem Organismus stammen. Anschließend wurden Modelle, die eine Sequenzidentität von 25% bis 35% zu ihrem Template aufwiesen, ausgewählt (3196 für *Modbase* und 1245 für *SwissModel*).

6.2 Suche nach hydrophoben Interaktionen mit Hilfe von *k-d* Bäumen

Um eine schnelle und effektive Suche von möglichen Interaktionen innerhalb von Proteinstrukturen zu ermöglichen wurde auf *k-d*-Bäume zurückgegriffen [134]. Dabei

werden sämtliche Atome einer Proteinstruktur in einen binären Suchbaum überführt und etwaige Nachbarn können effizient anhand ihrer Koordinaten identifiziert werden. Bei der Suche nach hydrophoben Wechselwirkungen wurde vereinfachter ein *k-d* Baum erstellt, der nur die Atome hydrophober Seitenketten (Tabelle 23 auf Seite 43) enthielt. Anschließend wurde eine Nachbarsuche für jedes im Baum enthaltene Atom durchgeführt. Wenn eine hydrophobe Interaktion identifiziert wurde, folgte eine weitere Nachbarsuche ausgehend von dem gefundenen Interaktionspartner. Durch diese doppelte Iteration der Suchen konnten etwaige hydrophobe *Cluster* innerhalb der Proteinstruktur analysiert werden.

6.3 Zuweisung von Sekundärstrukturen

Der in dieser Arbeit genutzte Mechanismus zur Zuweisung von Sekundärstrukturen besteht aus insgesamt vier Schritten. Zuerst erfolgt eine Zuweisung von Helices anhand von CA±3-Abständen, dem Winkel über drei aufeinanderfolgende Cα-Atome τ [175] sowie den Phi/Psi-Torsionswinkeln [171] (Tabelle 52).

Tabelle 52: Wertebereiche der verschiedenen Helixkriterien

| Parameter 1 | Bereich | Parameter 2 | Bereich |
|-------------|---|--------------|---|
| Tau | $80^\circ \leq \text{Tau} \leq 100^\circ$ | CA±3-Abstand | $\leq 6,3\text{Å}$ |
| Tau | $100^\circ \leq \text{Tau} \leq 110^\circ$ | CA±3-Abstand | $\leq 6,0\text{Å}$ |
| Phi | $-130^\circ \leq \text{Phi} \leq -40^\circ$ | Psi | $-70^\circ \leq \text{Psi} \leq 20^\circ$ |

Im Anschluss an die Identifizierung von helikalen Strukturen werden Cα_i-N_j-Abstandsmessungen zwischen den verbliebenen Aminosäuren durchgeführt, um eine Zuweisung von β-Strängen zu ermöglichen. Zwei Aminosäuren *i* und *j* werden dabei als Kandidaten für eine Strangzuweisung angesehen, wenn ihre Cα-N-Abstände kleiner als 6,0 Å sind. Im Anschluss daran erfolgte eine Verifizierung der potentiellen Stränge durch eine vereinfachte Berechnung von Wasserstoffbrücken (N-O-Abstand kleiner als 3,5 Å) zwischen den einzelnen Strängen. Nur Abschnitte, die mindestens eine potentielle Wasserstoffbrücke pro Aminosäure zu anderen als Strang gekennzeichneten Aminosäuren aufweisen, werden übernommen.

Tabelle 53: Sonderfälle bei der Zuweisung von β -Strängen

| Problem | Lösung |
|--|---|
| Kurzes Faltblatt mit 3 Aminosäuren pro Strang, bei nur zwei vorhandenen Strängen | Testen der Drehung der Proteinkette zwischen den beiden Strängen. Liegt diese bei mindestens 120° liegt ein Faltblatt vor. |
| Falsch negative Zuweisung eines Stranges? | Überprüfung des Tau-Winkels (105° bis 120°) und des $CA\pm 3$ -Abstandes ($\geq 9,7\text{\AA}$) |
| Lücken in Strang-Zuweisung / Kettenbrüche in Strängen | Überprüfung ob Aminosäuren davor und danach bestimmte Parameter erfüllen |
| Zu wenig Wasserstoffbrücken im Strang | Aufheben der Strangzuweisung |
| Strang mit weniger als 3 Aminosäuren | Aufheben der Strangzuweisung |

Aufgrund von Sonderfällen und Problemen an den Termini bzw. bei kleinen Strängen erfolgt nach der Strang-Zuweisung eine detaillierte Untersuchung der zugewiesenen Bereiche. Unter Nutzung verschiedener Test-Verfahren (Tabelle 53) wird eine genauere Zuweisung ermöglicht.

Der letzte Schritt bei der Bestimmung von Sekundärstrukturen innerhalb von Proteinstrukturen ist die Suche nach *tight Turns*. Dabei handelt es sich um Regionen, in deren Verlauf sich die Proteinkette um 180° wendet, die nicht mehr als sechs Aminosäuren umfassen und in denen der Abstand zwischen den $C\alpha$ -Atomen der ersten und der letzten Aminosäure kleiner als $7,0\text{\AA}$ ist [200]. In der vorliegenden Arbeit kommen 13 neu-definierte *Turn*-Strukturen zum Einsatz, die durch die Anwendung der *Turn*-Kriterien auf Bereiche zwischen Helices und/oder β -Strängen sowie eine Überprüfung der Phi/Psi-Torsionswinkel identifizierbar sind. Für mehr Informationen siehe 2.3.5 Verfeinerung von *tight Turns*.

6.4 Statistische Auswertungen

Alle positionsspezifischen oder Struktur-spezifischen statistischen Auswertungen dieser Arbeit wurden anhand von Formel (15) berechnet. Der Wert $n_{AA,x}$ steht dabei für die Anzahl von Aminosäure AA in Sekundärstrukturelement x oder an Position x und der Wert n_{AA} für die Anzahl an Aminosäure AA in der Referenzmenge, z.B. dem gesamten Datensatz.

$$\frac{n_{AA,x}}{\sum_{Ala}^{Val} n_{AA,x}} / \frac{n_{AA}}{\sum_{Ala}^{Val} n_{AA}} \quad (15)$$

Um zu testen, ob eine bestimmte Aminosäure-Kombination signifikant ist, können analog zu den Studien von Lifson und Sander [135] paarweise Korrelationen anhand von Formel (16) berechnet werden.

$$g_{i,j} = \frac{n_{i,j}}{E_{i,j}} \quad (16)$$

Dabei steht $n_{i,j}$ für die Anzahl der Kombination von Aminosäure i und Aminosäure j , und $E_{i,j}$ entspricht der erwarteten Anzahl an Kombinationen wenn die Verteilung zufällig erfolgt. Sie kann mit Hilfe Formel (17) berechnet werden

$$E_{i,j} = \frac{n_i n_j}{n} \quad (17)$$

Dabei stehen n_i und n_j für die Anzahl von Interaktionen an denen die Aminosäure i bzw. j beteiligt ist (18) und n für die Anzahl von allen gefundenen Interaktionen (19).

$$n_i = \sum_{j=1}^{20} n_{i,j} \quad (18)$$

$$n = \sum_{i=1}^{20} \sum_{j=1}^{20} n_{i,j} \quad (19)$$

6.5 Sekundärstrukturzuweisung mittels externer Programme

Um die Qualität des neu entwickelten Sekundärstrukturzuweisungsalgorithmus zu überprüfen, wurden dessen Ergebnisse mit denen älterer Programme verglichen (Tabelle 54). Da alle Programme unterschiedliche Ausgaben verwenden, wurden die Ergebnisse auf Drei- oder Vier-Zustandsmodelle reduziert. Die genaue Zuordnung der Ausgaben der einzelnen Programme zu den jeweiligen Sekundärstrukturmerkmalen ist in der letzten Spalte von Tabelle 54 zusammengefasst.

Tabelle 54: Externe Sekundärstrukturzuweisungsprogramme

In der letzten Spalte ist angegeben, wie die Ausgaben der einzelnen Programme zusammengefasst wurden um jeweils ein Drei- oder Vier-Zustandsmodell der Zuweisung zu erhalten.

| Programm | Version | Anmerkungen |
|--------------|-----------|--|
| DSSP [145] | 2.0.4 | Helix: H, G , I Strang: B, E <i>Turn</i> : T <i>Coil</i> : S, C |
| Stride [146] | 29.1.26 | Helix: H, G, I Strang: E <i>Turn</i> : T <i>Coil</i> : B, b, C |
| P-SEA [151] | 0.99 LMCP | Helix: a Strang: b <i>Coil</i> : c |
| KAKSI [150] | 2.3rc1 | Helix: H Strang: E <i>Coil</i> : . |

6.6 Informatische Module

Um die verschiedenen Parameter zu berechnen und in Form von bioinformatischen Modulen zu nutzen wurden im Zuge dieser Arbeit viele verschiedene C++ [252] und Python [253] Module genutzt. Die untenstehende Tabelle fasst alle zur Implementierung genutzten Module zusammen. Des Weiteren wurden sämtliche Webseiten mit Hilfe von php, *javascript* und css erstellt.

Tabelle 55: Im Zuge der Arbeit genutzte Python und C++ Module

| Python | C++ |
|------------------|-----------|
| Numpy [254] | iostream |
| Pandas [255] | String |
| Biopandas [256] | Vector |
| Matplotlib [257] | Glob.h |
| Os | Algorithm |
| Glob | Fstream |
| Pickle | Map |
| Sys | Cstring |
| | Sstream |
| | Cmath |
| | Eigen |
| | Sys |
| | Ctime |

7. Literaturverzeichnis

- [1] Astbury, W.T.; Reed, R.; Spark, L.C. (1948): An X-ray and electron microscope study of tropomyosin. *Biochem J* 2, 282–287.
- [2] Pauling, L.; Corey, R.B. (1951): Configuration of polypeptide chains. *Nature* 4274, 550–551.
- [3] Pauling, L.; Corey, R.B. (1951): Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds. Two New Pleated Sheets. *Proc Natl Acad Sci U S A* 11, 729–740.
- [4] Pauling, L.; Corey, R.B. (1951): The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A* 5, 251–256.
- [5] Voet, D.; Voet, J.G.; Pratt, C.W. (2019): *Lehrbuch der Biochemie*
- [6] Nič, M., Jirát, J., Košata, B., Jenkins, A., McNaught, A. (eds.) (2009): *IUPAC Compendium of Chemical Terminology*. IUPAC, Research Triangle Park, NC.
- [7] Ilari, A.; Savino, C. (2008): Protein structure determination by x-ray crystallography. *Methods Mol Biol*, 63–87.
- [8] Papageorgiou, A.C.; Mattsson, J. (2014): Protein structure validation and analysis with X-ray crystallography. *Methods Mol Biol*, 397–421.
- [9] Pedrini, B.; Serrano, P.; Mohanty, B.; Geralt, M.; Wüthrich, K. (2013): NMR-profiles of protein solutions. *Biopolymers* 11, 825–831.
- [10] Hodgson, D.J.; Aubin, Y. (2017): Assessment of the structure of pegylated-recombinant protein therapeutics by the NMR fingerprint assay. *J Pharm Biomed Anal*, 351–356.
- [11] Faraggi, E.; Dunker, A.K.; Sussman, J.L.; Kloczkowski, A. (2018): Comparing NMR and X-ray protein structure. Lindemann-like parameters and NMR disorder. *J Biomol Struct Dyn* 9, 2331–2341.
- [12] Khoshouei, M.; Radjainia, M.; Baumeister, W.; Danev, R. (2017): Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat Commun*, 16099.
- [13] Sharif, H.; Li, Y.; Dong, Y.; Dong, L.; Wang, W.L.; Mao, Y.; Wu, H. (2017): Cryo-EM structure of the DNA-PK holoenzyme. *Proc Natl Acad Sci U S A* 28, 7367–7372.
- [14] Li, J.; Sun, J. (2017): Application of X-ray Diffraction and Electron Crystallography for Solving Complex Structure Problems. *Acc Chem Res* 11, 2737–2745.

- [15] Wang, H.-W.; Wang, J.-W. (2017): How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci* 1, 32–39.
- [16] Kühlbrandt, W. (2013): Introduction to electron crystallography. *Methods Mol Biol*, 1–16.
- [17] Liebschner, D.; Afonine, P.V.; Moriarty, N.W.; Langan, P.; Adams, P.D. (2018): Evaluation of models determined by neutron diffraction and proposed improvements to their validation and deposition. *Acta Crystallogr D Struct Biol Pt 8*, 800–813.
- [18] Marvin, D.A. (2017): Fibre diffraction studies of biological macromolecules. *Prog Biophys Mol Biol*, 43–87.
- [19] Kessel, A.; Ben-Tal, N. (2011): Introduction to proteins. Structure, function, and motion. Chapman & Hall/CRC mathematical and computational biology series. CRC Press, Boca Raton.
- [20] Engin, H.B.; Kreisberg, J.F.; Carter, H. (2016): Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS ONE* 4, e0152929.
- [21] Kumalo, H.M.; Bhakat, S.; Soliman, M.E. (2015): Heat-shock protein 90 (Hsp90) as anticancer target for drug discovery. An ample computational perspective. *Chem Biol Drug Des* 5, 1131–1160.
- [22] Wu, B.X.; Hong, F.; Zhang, Y.; Ansa-Addo, E.; Li, Z. (2016): GRP94/gp96 in Cancer. *Biology, Structure, Immunology, and Drug Development. Adv Cancer Res*, 165–190.
- [23] Niu, B.; Scott, A.D.; Sengupta, S.; Bailey, M.H.; Batra, P.; Ning, J.; Wyczalkowski, M.A.; Liang, W.-W.; Zhang, Q.; McLellan, M.D.; Sun, S.Q.; Tripathi, P.; Lou, C.; Ye, K.; Mashl, R.J.; Wallis, J.; Wendl, M.C.; Chen, F.; Ding, L. (2016): Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* 8, 827–837.
- [24] Lee, J.; Nyenhuis, D.A.; Nelson, E.A.; Cafiso, D.S.; White, J.M.; Tamm, L.K. (2017): Structure of the Ebola virus envelope protein MPER/TM domain and its interaction with the fusion loop explains their fusion activity. *Proc Natl Acad Sci U S A* 38, E7987-E7996.
- [25] Wan, W.; Kolesnikova, L.; Clarke, M.; Koehler, A.; Noda, T.; Becker, S.; Briggs, J.A.G. (2017): Structure and assembly of the Ebola virus nucleocapsid. *Nature* 7680, 394–397.
- [26] Lee, W.; Tonelli, M.; Wu, C.; Aceti, D.J.; Amarasinghe, G.K.; Markley, J.L. (2019): Backbone resonance assignments and secondary structure of Ebola nucleoprotein 600-739 construct. *Biomol NMR Assign* .

- [27] Huizing, A.P.; Mondal, M.; Hirsch, A.K.H. (2015): Fighting malaria. Structure-guided discovery of nonpeptidomimetic plasmepsin inhibitors. *J Med Chem* 13, 5151–5163.
- [28] Liu, X.; Huang, Y.; Liang, J.; Wang, J.; Shen, Y.; Li, Y.; Zhao, Y. (2016): Characterization of the malaria parasite protein PfTip, a novel invasion-related protein. *Mol Med Rep* 4, 3303–3310.
- [29] Freedman, H.; Logan, M.R.; Law, J.L.M.; Houghton, M. (2016): Structure and Function of the Hepatitis C Virus Envelope Glycoproteins E1 and E2. *Antiviral and Vaccine Targets. ACS Infect Dis* 11, 749–762.
- [30] McCauley, J.A.; Rudd, M.T. (2016): Hepatitis C virus NS3/4a protease inhibitors. *Curr Opin Pharmacol*, 84–92.
- [31] Arnold, E., Himmel, D.M. (eds.) (2011): *Crystallography of biological macromolecules*, 2nd edn. *International tables for crystallography*, / publ. for the International Union of Crystallography ; Vol. F. Wiley, Chichester.
- [32] Benjak, A.; Sala, C.; Hartkoorn, R.C. (2015): Whole-genome sequencing for comparative genomics and de novo genome assembly. *Methods Mol Biol*, 1–16.
- [33] Land, M.; Hauser, L.; Jun, S.-R.; Nookaew, I.; Leuze, M.R.; Ahn, T.-H.; Karpinets, T.; Lund, O.; Kora, G.; Wassenaar, T.; Poudel, S.; Ussery, D.W. (2015): Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 2, 141–161.
- [34] Nakagawa, H.; Wardell, C.P.; Furuta, M.; Taniguchi, H.; Fujimoto, A. (2015): Cancer whole-genome sequencing. Present and future. *Oncogene* 49, 5943–5950.
- [35] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. (2000): The Protein Data Bank. *Nucleic Acids Res* 1, 235–242.
- [36] Schwede, T. (2013): Protein Modelling. What Happened to the “Protein Structure Gap”? *Structure* 9, [EPub].
- [37] Bourne, P.E., Weissig, H. (eds.) (2005): *Structural bioinformatics. Methods of Biochemical Analysis*, vol. 44. Wiley, Hoboken, NJ.
- [38] Kukul, A. (ed.) (2008): *Molecular Modeling of Proteins. SpringerLink Bücher*, vol. 443. Humana Press, Totowa, NJ.
- [39] Godzik, A. (2005): Fold Recognition Methods. In: Bourne, P.E., Weissig, H. (eds.) *Structural bioinformatics. Methods of Biochemical Analysis*, vol. 44, pp. 525–546. Wiley, Hoboken, NJ.

- [40] Krieger, E.; Nabuurs, S.B.; Vriend, G. (2005): Homology Modeling. In: Bourne, P.E., Weissig, H. (eds.) Structural bioinformatics. Methods of Biochemical Analysis, vol. 44, pp. 509–523. Wiley, Hoboken, NJ.
- [41] Lushington, G.H. (2008): Comparative Modeling of Proteins. In: Kukol, A. (ed.) Molecular Modeling of Proteins. SpringerLink Bücher, vol. 443, pp. 199–212. Humana Press, Totowa, NJ.
- [42] Tramontano, A.; Cozzetto, D. (2011): Evaluation of Protein Structure Prediction Methods. Issues and Strategies. In: Kolinski, A. (ed.) Multiscale Approaches to Protein Modeling. Structure Prediction, Dynamics, Thermodynamics and Macromolecular Assemblies. Multiscale approaches to protein modeling, 1st edn., pp. 315–339. Springer Science+Business Media LLC, New York, NY.
- [43] Martí-Renom, M.A.; Stuart, A.C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. (2000): Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 291–325.
- [44] Zhang, Y.; Skolnick, J. (2005): The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci U S A* 4, 1029–1034.
- [45] Chothia, C.; Lesk, A.M. (1986): The relation between the divergence of sequence and structure in proteins. *EMBO J* 4, 823–826.
- [46] Orengo, C.A.; Flores, T.P.; Jones, D.T.; Taylor, W.R.; Thornton, J.M. (1993): Recurring structural motifs in proteins with different functions. *Curr Biol* 3, 131–139.
- [47] Torda, A.E. (1997): Perspectives in protein-fold recognition. *Curr Opin Struct Biol* 2, 200–205.
- [48] Deng, H.; Jia, Y.; Zhang, Y. (2018): Protein structure prediction. *Int J Mod Phys B* 18, [EPub].
- [49] Bordoli, L.; Schwede, T. (2012): Automated protein structure modeling with SWISS-MODEL Workspace and the Protein Model Portal. *Methods Mol Biol*, 107–136.
- [50] Yang, J.; Zhang, Y. (2015): Protein Structure and Function Prediction Using I-TASSER. *Curr Protoc Bioinformatics*, 5.8.1-15.
- [51] Roche, D.B.; McGuffin, L.J. (2016): Toolbox for Protein Structure Prediction. *Methods Mol Biol*, 363–377.
- [52] Park, H.; Kim, D.E.; Ovchinnikov, S.; Baker, D.; DiMaio, F. (2018): Automatic structure prediction of oligomeric assemblies using Robetta in CASP12. *Proteins*, 283–291.

- [53] Poupon, A.; Janin, J. (2010): Analysis and Prediction of Protein Quaternary Structure. In: Carugo, O., Eisenhaber, F. (eds.) *Data Mining Techniques for the Life Sciences*. SpringerLink Bücher, vol. 609, pp. 349–364. Humana Press, Totowa, NJ.
- [54] Huang, S.-Y. (2014): Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov Today* 8, 1081–1096.
- [55] Szilagyi, A.; Zhang, Y. (2014): Template-based structure modeling of protein-protein interactions. *Curr Opin Struct Biol*, 10–23.
- [56] Huang, S.-Y. (2015): Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov Today* 8, 969–977.
- [57] Bordogna, A.; Pandini, A.; Bonati, L. (2011): Predicting the accuracy of protein-ligand docking on homology models. *J Comput Chem* 1, 81–98.
- [58] Ehrt, C.; Brinkjost, T.; Koch, O. (2016): Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J Med Chem* 9, 4121–4151.
- [59] Mor, M.; Rivara, S.; Lodola, A.; Lorenzi, S.; Bordi, F.; Plazzi, P.V.; Spadoni, G.; Bedini, A.; Duranti, A.; Tontini, A.; Tarzia, G. (2005): Application of 3D-QSAR in the rational design of receptor ligands and enzyme inhibitors. *Chem Biodivers* 11, 1438–1451.
- [60] Rognan, D. (2010): Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol Inform* 3, 176–187.
- [61] Verma, J.; Khedkar, V.M.; Coutinho, E.C. (2010): 3D-QSAR in drug design—a review. *Curr Top Med Chem* 1, 95–115.
- [62] Kalyaanamoorthy, S.; Chen, Y.-P.P. (2011): Structure-based drug design to augment hit discovery. *Drug Discov Today* 17-18, 831–839.
- [63] Vedani, A.; Dobler, M. (2012): Multi-dimensional QSAR in drug research. In: Jucker, E. (ed.) *Progress in Drug Research*, pp. 105–135. Birkhauser, Basel.
- [64] Bienert, S.; Waterhouse, A.; Beer, T.A.P. de; Tauriello, G.; Studer, G.; Bordoli, L.; Schwede, T. (2017): The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res D1*, D313-D319.
- [65] Haas, J.; Roth, S.; Arnold, K.; Kiefer, F.; Schmidt, T.; Bordoli, L.; Schwede, T. (2013): The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database (Oxford)*, bat031.

- [66] Abriata, L.A.; Tamò, G.E.; Monastyrskyy, B.; Kryshtafovych, A.; Dal Peraro, M. (2018): Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, 97–112.
- [67] Kinch, L.N.; Li, W.; Monastyrskyy, B.; Kryshtafovych, A.; Grishin, N.V. (2016): Evaluation of free modeling targets in CASP11 and ROLL. *Proteins*, 51–66.
- [68] Kryshtafovych, A.; Monastyrskyy, B.; Fidelis, K.; Schwede, T.; Tramontano, A. (2018): Assessment of model accuracy estimations in CASP12. *Proteins*, 345–360.
- [69] Moulton, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. (2018): Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins*, 7–15.
- [70] Zemla, A. (2003): LGA. A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31, 3370–3374.
- [71] Zhang, Y.; Skolnick, J. (2005): TM-align. A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33, 2302–2309.
- [72] Laskowski, R.A.; MacArthur, M.W.; Moss, D.S.; Thornton, J.M. (1993): PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26, 283–291.
- [73] Vriend, G. (1991): What IF: a molecular modeling and drug design program. *Journal of Molecular Graphics* 1, 52–56.
- [74] Melo, F.; Devos, D.; Depiereux, E.; Feytmans, E. (1997): ANOLEA: A www Server to Assess Protein Structures. *Proc Int Conf Intell Syst Mol Biol* 5, 187–190.
- [75] Wiederstein, M.; Sippl, M.J. (2007): ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res Web Server issue*, 407–410.
- [76] Benkert, P.; Künzli, M.; Schwede, T. (2009): QMEAN server for protein model quality estimation. *Nucleic Acids Res Web Server issue*, 510–514.
- [77] Davis, I.W.; Leaver-Fay, A.; Chen, V.B.; Block, J.N.; Kapral, G.J.; Wang, X.; Murray, L.W.; Arendall, W.B.; Snoeyink, J.; Richardson, J.S.; Richardson, D.C. (2007): MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res Web Server issue*, 375–383.
- [78] Singh, A.; Kaushik, R.; Mishra, A.; Shanker, A.; Jayaram, B. (2016): ProTSAV. A protein tertiary structure analysis and validation server. *Biochim Biophys Acta* 1864, 11–19.

- [79] Jamroz, M.; Kolinski, A.; Kihara, D. (2016): Ensemble-based evaluation for protein structure models. *Bioinformatics* 12, i314-i321.
- [80] Jing, X.; Dong, Q. (2017): MQAPRank. Improved global protein model quality assessment by learning-to-rank. *BMC Bioinformatics* 1, 275.
- [81] Olechnovič, K.; Venclovas, Č. (2017): VoromQA. Assessment of protein structure quality using interatomic contact areas. *Proteins* 6, 1131–1145.
- [82] Shin, W.-H.; Kang, X.; Zhang, J.; Kihara, D. (2017): Prediction of Local Quality of Protein Structure Models Considering Spatial Neighbors in Graphical Models. *Sci Rep*, 40629.
- [83] Adiyaman, R.; McGuffin, L.J. (2019): Methods for the Refinement of Protein Structure 3D Models. *Int J Mol Sci* 9, [EPub].
- [84] The UniProt Consortium (2019): UniProt. A worldwide hub of protein knowledge. *Nucleic Acids Res* D1, D506-D515.
- [85] Burley, S.K.; Petsko, G.A. (1985): Aromatic-aromatic interaction. A mechanism of protein structure stabilization. *Science* 4708, 23–28.
- [86] Hinnebusch, A.G. (2014): The scanning mechanism of eukaryotic translation initiation. *Annu Rev Biochem*, 779–812.
- [87] Wingfield, P.T. (2017): N-Terminal Methionine Processing. *Curr Protoc Protein Sci*, 6.14.1-6.14.3.
- [88] Strader, M.B.; Hicks, W.A.; Kassa, T.; Singleton, E.; Soman, J.; Olson, J.S.; Weiss, M.J.; Mollan, T.L.; Wilson, M.T.; Alayash, A.I. (2014): Post-translational transformation of methionine to aspartate is catalyzed by heme iron and driven by peroxide. A novel subunit-specific mechanism in hemoglobin. *J Biol Chem* 32, 22342–22357.
- [89] The UniProt Consortium (2019): UniProtKB/SwissProt Statistics release 2019_03. <https://www.uniprot.org/statistics/Swiss-Prot> (Zuletzt geprüft: 02.05.2019).
- [90] Falsafi-Zadeh, S.; Karimi, Z.; Galehdari, H. (2012): VMD DisRg. New User-Friendly Implement for calculation distance and radius of gyration in VMD program. *Bioinformation* 7, 341–343.
- [91] Dong, G.; Hutagalung, A.H.; Fu, C.; Novick, P.; Reinisch, K.M. (2005): The structures of exocyst subunit Exo70p and the Exo84p C-terminal domains reveal a common motif
- [92] Wilding, M.; Peat, T.S.; Newman, J.; Scott, C. (2016): Characterization of a Novel Transaminase from *Pseudomonas* sp. *Strain AAC*

- [93] Flory, P.J. (1949): The Configuration of Real Polymer Chains. *The Journal of Chemical Physics* 3, 303–310.
- [94] Kohn, J.E.; Millett, I.S.; Jacob, J.; Zagrovic, B.; Dillon, T.M.; Cingel, N.; Dothager, R.S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T.R.; Hasan, M.Z.; Pande, V.S.; Ruczinski, I.; Doniach, S.; Plaxco, K.W. (2004): Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci U S A* 34, 12491–12496.
- [95] Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. (2012): Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc Natl Acad Sci U S A* 40, 16155–16160.
- [96] Gong, H.; Fleming, P.J.; Rose, G.D. (2005): Building native protein conformation from highly approximate backbone torsion angles. *Proc Natl Acad Sci U S A* 45, 16227–16232.
- [97] Hong, L.; Lei, J. (2009): Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity. *J. Polym. Sci. B Polym. Phys.* 2, 207–214.
- [98] Tanner, J.J. (2016): Empirical power laws for the radii of gyration of protein oligomers. *Acta Crystallogr D Struct Biol Pt* 10, 1119–1129.
- [99] Wilkins, D.K.; Grimshaw, S.B.; Receveur, V.; Dobson, C.M.; Jones, J.A.; Smith, L.J. (1999): Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry* 50, 16424–16431.
- [100] Morris, A.L.; MacArthur, M.W.; Hutchinson, E.G.; Thornton, J.M. (1992): Stereochemical quality of protein structure coordinates. *Proteins* 4, 345–364.
- [101] Schreiner, E.; Trabuco, L.G.; Freddolino, P.L.; Schulten, K. (2011): Stereochemical errors and their implications for molecular dynamics simulations. *BMC Bioinformatics*, 190.
- [102] Stewart, D.E.; Sarkar, A.; Wampler, J.E. (1990): Occurrence and role of cis peptide bonds in protein structures. *J Mol Biol* 1, 253–260.
- [103] MacArthur, M.W.; Thornton, J.M. (1991): Influence of proline residues on protein conformation. *J Mol Biol* 2, 397–412.
- [104] Jabs, A.; Weiss, M.S.; Hilgenfeld, R. (1999): Non-proline cis peptide bonds in proteins. *J Mol Biol* 1, 291–304.
- [105] Croll, T.I. (2015): The rate of cis-trans conformation errors is increasing in low-resolution crystal structures. *Acta Crystallogr D Biol Crystallogr Pt* 3, 706–709.

- [106] Chen, V.B.; Arendall, W.B.; Headd, J.J.; Keedy, D.A.; Immormino, R.M.; Kapral, G.J.; Murray, L.W.; Richardson, J.S.; Richardson, D.C. (2010): MolProbity. All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr. Pt 1*, 12–21.
- [107] Williams, C.J.; Richardson, J.S. (2015): Avoiding excess cis peptides at low resolution or high B. *Computational Crystallography Newsletter* 6, 2–6.
- [108] Fisher, S.J.; Blakeley, M.P.; Cianci, M.; McSweeney, S.; Helliwell, J.R. (2012): Protonation-state determination in proteins using high-resolution X-ray crystallography. Effects of resolution and completeness. *Acta Crystallogr. D Biol. Crystallogr. Pt 7*, 800–809.
- [109] Ahmed, H.U.; Blakeley, M.P.; Cianci, M.; Cruickshank, D.W.J.; Hubbard, J.A.; Helliwell, J.R. (2007): The determination of protonation states in proteins. *Acta Crystallogr D Biol Crystallogr Pt 8*, 906–922.
- [110] Eyring, H. (1932): Steric Hinderance and Collision Diameters 1. *J Am Chem Soc* 8, 3191–3203.
- [111] Janin, J.; Wodak, S.; Levitt, M.; Maigret, B. (1978): Conformation of amino acid side-chains in proteins. *J Mol Biol* 3, 357–386.
- [112] Bhat, T.N.; Sasisekharan, V.; Vijayan, M. (1979): An Analysis of Side-chain Conformation in Proteins*. *Int J Pept Protein Res* 2, 170–184.
- [113] McGregor, M.J.; Islam, S.A.; Sternberg, M.J.E. (1987): Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol* 2, 295–310.
- [114] Ramachandran, G.N.; Lakshminarayanan, A.V.; Balasubramanian, R.; Tegoni, G. (1970): Studies on the conformation of amino acids. XII. Energy calculations on prolyl residue. *Biochim Biophys Acta* 2, 165–181.
- [115] Ho, B.K.; Coutsiias, E.A.; Seok, C.; Dill, K.A. (2005): The flexibility in the proline ring couples to the protein backbone. *Protein Sci* 4, 1011–1018.
- [116] Willard, L.; Ranjan, A.; Zhang, H.; Monzavi, H.; Boyko, R.F.; Sykes, B.D.; Wishart, D.S. (2003): VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res* 13, 3316–3319.
- [117] Barlow, D.J.; Thornton, J.M. (1983): Ion-pairs in proteins. *J Mol Biol* 4, 867–885.

- [118] Lee, C.-W.; Wang, H.-J.; Hwang, J.-K.; Tseng, C.-P. (2014): Protein thermal stability enhancement by designing salt bridges. A combined computational and experimental study. *PLoS ONE* 11, e112751.
- [119] Sarakatsannis, J.N.; Duan, Y. (2005): Statistical characterization of salt bridges in proteins. *Proteins* 4, 732–739.
- [120] Wigley, D.B.; Lyall, A.; Hart, K.W.; Holbrook, J.J. (1987): The greater strength of arginine. Carboxylate over lysine carboxylate ion pairs implications for the design of novel enzymes and drugs. *Biochem Biophys Res Commun* 3, 927–929.
- [121] Burley, S.K.; Petsko, G.A. (1986): Amino-aromatic interactions in proteins. *FEBS Lett* 2, 139–143.
- [122] Flocco, M.M.; Mowbray, S.L. (1994): Planar stacking interactions of arginine and aromatic side-chains in proteins. *J Mol Biol* 2, 709–717.
- [123] Gallivan, J.P.; Dougherty, D.A. (1999): Cation- π interactions in structural biology. *Proc Natl Acad Sci U S A* 17, 9459–9464.
- [124] Crowley, P.B.; Golovin, A. (2005): Cation- π interactions in protein-protein interfaces. *Proteins* 2, 231–239.
- [125] Dougherty, D.A. (1996): Cation- π interactions in chemistry and biology. A new view of benzene, Phe, Tyr, and Trp. *Science* 5246, 163–168.
- [126] McGaughey, G.B.; Gagné, M.; Rappé, A.K. (1998): π -Stacking interactions. Alive and well in proteins. *J Biol Chem* 25, 15458–15463.
- [127] Jorgensen, W.L.; Severance, D.L. (1990): Aromatic-aromatic interactions. Free energy profiles for the benzene dimer in water, chloroform, and liquid benzene. *J Am Chem Soc* 12, 4768–4774.
- [128] Chourasia, M.; Sastry, G.M.; Sastry, G.N. (2011): Aromatic-Aromatic Interactions Database, A(2)ID. An analysis of aromatic π -networks in proteins. *Int J Biol Macromol* 4, 540–552.
- [129] Jaffe, R.L.; Smith, G.D. (1996): A quantum chemistry study of benzene dimer. *The Journal of Chemical Physics* 7, 2780–2788.
- [130] Anjana, R.; Vaishnavi, M.K.; Sherlin, D.; Kumar, S.P.; Naveen, K.; Kanth, P.S.; Sekar, K. (2012): Aromatic-aromatic interactions in structures of proteins and protein-DNA complexes. A study based on orientation and distance. *Bioinformation* 24, 1220–1224.

- [131] Li, J.; Wang, Y.; An, L.; Chen, J.; Yao, L. (2018): Direct Observation of CH/CH van der Waals Interactions in Proteins by NMR. *J Am Chem Soc* 9, 3194–3197.
- [132] Scuseria, G.E.; Janssen, C.L.; Schaefer, H.F. (1988): An efficient reformulation of the closed-shell coupled cluster single and double excitation (CCSD) equations. *The Journal of Chemical Physics* 12, 7382–7387.
- [133] Bagno, A.; Saielli, G.; Scorrano, G. (2002): Through-Space Spin–Spin Coupling in van der Waals Dimers and CH/ π Interacting Systems. An Ab Initio and DFT Study. *Chemistry* 9, 2047.
- [134] Bentley, J.L. (1975): Multidimensional binary search trees used for associative searching. *Commun. ACM* 9, 509–517.
- [135] Lifson, S.; Sander, C. (1980): Specific recognition in the tertiary structure of beta-sheets of proteins. *J Mol Biol* 4, 627–639.
- [136] Devlin, T.M. (ed.) (2011): *Textbook of biochemistry. With clinical correlations*, 7th edn. Wiley, Hoboken, NJ.
- [137] Kalinowska, B.; Banach, M.; Wiśniowski, Z.; Konieczny, L.; Roterman, I. (2017): Is the hydrophobic core a universal structural element in proteins? *J Mol Model* 7, 205.
- [138] Konieczny, L.; Brylinski, M.; Roterman, I. (2006): Gauss-function-Based model of hydrophobicity density in proteins. *In Silico Biol (Gedruckt)* 1-2, 15–22.
- [139] Kalinowska, B.; Banach, M.; Konieczny, L.; Roterman, I. (2015): Application of Divergence Entropy to Characterize the Structure of the Hydrophobic Core in DNA Interacting Proteins. *Entropy* 3, 1477–1507.
- [140] Kauzmann, W. (1959): Some factors in the interpretation of protein denaturation. *Adv Protein Chem*, 1–63.
- [141] Alberts, B.; Jaenicke, L. (2004): *Molekularbiologie der Zelle*. [mit Cell biology interactive], 4th edn. Wiley-VCH, Weinheim.
- [142] Blakeley, M.P.; Hasnain, S.S.; Antonyuk, S.V. (2015): Sub-atomic resolution X-ray crystallography and neutron crystallography. Promise, challenges and potential. *IUCrJ Pt* 4, 464–474.
- [143] Tosatto, S.C.E. (2005): The victor/FRST function for model quality estimation. *J Comput Biol* 10, 1316–1327.
- [144] Jeffrey, G.A. (1997): *An introduction to hydrogen bonding. Topics in physical chemistry*. Oxford Univ. Press, New York.

- [145] Kabsch, W.; Sander, C. (1983): Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.
- [146] Richards, F.M.; Kundrot, C.E. (1988): Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins* 2, 71–84.
- [147] Frishman, D.; Argos, P. (1995): Knowledge-based protein secondary structure assignment. *Proteins* 4, 566–579.
- [148] Frishman, D.; Argos, P. (1997): Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 3, 329–335.
- [149] Majumdar, I.; Krishna, S.S.; Grishin, N.V. (2005): PALSSE. A program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics*, 202.
- [150] Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; Brevern, A.G. de; Gibrat, J.-F. (2005): Protein secondary structure assignment revisited. A detailed analysis of different assignment methods. *BMC Struct. Biol.*, 17.
- [151] Labesse, G.; Colloc'h, N.; Pothier, J.; Mornon, J.P. (1997): P-SEA. A new efficient assignment of secondary structure from C alpha trace of proteins. *Comput Appl Biosci* 3, 291–295.
- [152] Zhang, Y.; Sagui, C. (2015): Secondary structure assignment for conformationally irregular peptides. Comparison between DSSP, STRIDE and KAKSI. *J Mol Graph Model*, 72–84.
- [153] Andersen, C.A.F.; Rost, B. (2003): Secondary structure assignment. *Methods Biochem Anal*, 341–363.
- [154] Lifson, S.; Hagler, A.T.; Dauber, P. (1979): Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. Carboxylic acids, amides, and the C. O.cntdot.cntdot.cntdot.H- hydrogen bonds. *J Am Chem Soc* 18, 5111–5121.
- [155] Creighton, T.E. (1994): *Proteins. Structures and molecular properties*, 2nd edn. Freeman, New York.
- [156] Boobbyer, D.N.A.; Goodford, P.J.; McWhinnie, P.M.; Wade, R.C. (1989): New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J Med Chem* 5, 1083–1094.

- [157] Wade, R.C.; Clark, K.J.; Goodford, P.J. (1993): Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J Med Chem* 1, 140–147.
- [158] Richardson, J.S. (1981): The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 167–339.
- [159] Wilmot, C.M.; Thornton, J.M. (1988): Analysis and prediction of the different types of β -turn in proteins. *J Mol Biol* 1, 221–232.
- [160] Cuff, J.A.; Barton, G.J. (1999): Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 4, 508–519.
- [161] Zhang, W.; Dunker, A.K.; Zhou, Y. (2008): Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* 1, 61–67.
- [162] Shindyalov, I.N.; Bourne, P.E. (1998): Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 9, 739–747.
- [163] Holm, L.; Sander, C. (1994): The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res* 17, 3600–3609.
- [164] Cao, C.; Wang, G.; Liu, A.; Xu, S.; Wang, L.; Zou, S. (2016): A New Secondary Structure Assignment Algorithm Using $C\alpha$ Backbone Fragments. *Int J Mol Sci* 3, 333.
- [165] Rost, B.; Sander, C.; Schneider, R. (1994): Redefining the goals of protein secondary structure prediction. *J Mol Biol* 1, 13–26.
- [166] Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. (1999): A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* 2, 220–223.
- [167] Rost, B.; Sander, C. (1993): Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 2, 584–599.
- [168] Salamov, A.A.; Solovyev, V.V. (1995): Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol* 1, 11–15.
- [169] Berg, J.M.; Tymoczko, J.L.; Stryer, L.; Häcker, B. (2011): *Biochemie*, 6th edn. Spektrum Akad. Verl., Heidelberg.

- [170] Rotondi, K.S.; Gierasch, L.M. (2006): Natural polypeptide scaffolds: beta-sheets, beta-turns, and beta-hairpins. *Biopolymers* 1, 13–22.
- [171] Crisma, M.; Formaggio, F.; Moretto, A.; Toniolo, C. (2006): Peptide helices based on alpha-amino acids. *Biopolymers* 1, 3–12.
- [172] Adzhubei, A.A.; Sternberg, M.J.E.; Makarov, A.A. (2013): Polyproline-II helix in proteins. Structure and function. *J Mol Biol* 12, 2100–2132.
- [173] Fodje, M.N.; Al-Karadaghi, S. (2002): Occurrence, conformational features and amino acid propensities for the pi-helix. *Protein Eng* 5, 353–358.
- [174] Pal, L.; Basu, G.; Chakrabarti, P. (2002): Variants of 3(10)-helices in proteins. *Proteins* 3, 571–579.
- [175] Levitt, M.; Greer, J. (1977): Automatic identification of secondary structure in globular proteins. *J Mol Biol* 2, 181–239.
- [176] Chou, P.Y.; Fasman, G.D. (1974): Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* 2, 211–222.
- [177] Chou, P.Y.; Fasman, G.D. (1974): Prediction of protein conformation. *Biochemistry* 2, 222–245.
- [178] Chou, P.Y.; Fasman, G.D. (1977): Beta-turns in proteins. *J Mol Biol* 2, 135–175.
- [179] Chou, P.Y.; Fasman, G.D. (1978): Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*, 45–148.
- [180] Costantini, S.; Colonna, G.; Facchiano, A.M. (2006): Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem Biophys Res Commun* 2, 441–451.
- [181] Nakashima, H.; Nishikawa, K.; Ooi, T. (1986): The folding type of a protein is relevant to the amino acid composition. *J Biochem* 1, 153–162.
- [182] Chou, K.C. (1995): A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins* 4, 319–344.
- [183] Fujiwara, K.; Toda, H.; Ikeguchi, M. (2012): Dependence of α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.*, 18.
- [184] Fujiwara, K.; Ikeguchi, M. (2008): OLIGAMI. OLIGomer Architecture and Molecular Interface. *TOBIOIJ* 1, 50–53.

- [185] Hubbard, T.J.; Ailey, B.; Brenner, S.E.; Murzin, A.G.; Chothia, C. (1999): SCOP. A Structural Classification of Proteins database. *Nucleic Acids Res* 1, 254–256.
- [186] Malkov, S.N.; Zivković, M.V.; Beljanski, M.V.; Hall, M.B.; Zarić, S.D. (2008): A reexamination of the propensities of amino acids towards a particular secondary structure. Classification of amino acids based on their chemical structure. *J Mol Model* 8, 769–775.
- [187] Richardson, J.S.; Richardson, D.C. (1988): Amino acid preferences for specific locations at the ends of alpha helices. *Science* 4859, 1648–1652.
- [188] Aurora, R.; Rose, G.D. (1998): Helix capping. *Protein Sci* 1, 21–38.
- [189] Vasudev, P.G.; Banerjee, M.; Ramakrishnan, C.; Balaram, P. (2012): Asparagine and glutamine differ in their propensities to form specific side chain-backbone hydrogen bonded motifs in proteins. *Proteins* 4, 991–1002.
- [190] Presta, L.G.; Rose, G.D. (1988): Helix signals in proteins. *Science* 4859, 1632–1641.
- [191] Shoemaker, K.R.; Kim, P.S.; York, E.J.; Stewart, J.M.; Baldwin, R.L. (1987): Tests of the helix dipole model for stabilization of alpha-helices. *Nature* 6113, 563–567.
- [192] Hovmöller, S.; Zhou, T.; Ohlson, T. (2002): Conformations of amino acids in proteins. *Acta Crystallogr. D Biol. Crystallogr. Pt 5*, 768–776.
- [193] Street, A.G.; Mayo, S.L. (1999): Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc Natl Acad Sci U S A* 16, 9074–9076.
- [194] Muñoz, V.; Serrano, L. (1994): Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices. Comparison with experimental scales. *Proteins* 4, 301–311.
- [195] Gasteiger, J.; Marsili, M. (1980): Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 22, 3219–3228.
- [196] Rappe, A.K.; Goddard, W.A. (1991): Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* 8, 3358–3363.
- [197] West, M.W.; Hecht, M.H. (1995): Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci* 10, 2032–2039.
- [198] Mandel-Gutfreund, Y.; Gregoret, L.M. (2002): On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J Mol Biol* 3, 453–461.

- [199] Venkatachalam, C.M. (1968): Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* 10, 1425–1436.
- [200] Chou, K.-C. (2000): Prediction of tight turns and their types in proteins. *Anal Biochem* 1, 1–16.
- [201] Némethy, G.; Printz, M.P. (1972): The γ Turn, a Possible Folded Conformation of the Polypeptide Chain. Comparison with the β Turn. *Macromolecules* 6, 755–758.
- [202] Chou, K.-C. (1997): Prediction and classification of α -turn types. *Biopolymers* 7, 837–853.
- [203] Rajashankar, K.R.; Ramakumar, S. (1996): Pi-Turns in proteins and peptides: Classification, conformation, occurrence, hydration and sequence. *Protein Sci* 5, 932–946.
- [204] Dhar, J.; Chakrabarti, P.; Saini, H.; Raghava, G.P.S.; Kishore, R. (2015): ω -Turn. A novel β -turn mimic in globular proteins stabilized by main-chain to side-chain C–H \cdots O interaction. *Proteins* 2, 203–214.
- [205] Toniolo, C.; Crisma, M.; Formaggio, F.; Alemán, C.; Ramakrishnan, C.; Kalmankar, N.; Balaram, P. (2017): Intramolecular backbone \cdots backbone hydrogen bonds in polypeptide conformations. The other way around. ϵ -turn. *Biopolymers* 1, [EPub].
- [206] Thornton, J.M.; Sibanda, B.L.; Edwards, M.S.; Barlow, D.J. (1988): Analysis, design and modification of loop regions in proteins. *Bioessays* 2, 63–69.
- [207] Smith, J.A.; Pease, L.G.; Kopple, K.D. (2008): Reverse Turns in Peptides and Protein. *Critical Reviews in Biochemistry* 4, 315–399.
- [208] Engh, R.A.; Huber, R. (1991): Accurate bond and angle parameters for X-ray protein structure refinement. *Acta crystallographica* 47, 392–400.
- [209] Hooft, R.W.W.; Sander, C.; Vriend, G.: Verification of Protein Structures: Side-Chain Planarity .
- [210] Ponder, J.W.; Case, D.A. (2003): Force fields for protein simulations. *Adv Protein Chem*, 27–85.
- [211] Wang, J.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. (2004): Development and testing of a general amber force field. *J Comput Chem* 9, 1157–1174.
- [212] Wang, J.; Wang, W.; Kollman, P.A.; Case, D.A. (2006): Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 2, 247–260.

- [213] Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. (2015): ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 8, 3696–3713.
- [214] Bagaria, A.; Jaravine, V.; Huang, Y.J.; Montelione, G.T.; Güntert, P. (2012): Protein structure validation by generalized linear model root-mean-square deviation prediction. *Protein Sci* 2, 229–238.
- [215] Walker, J.M.; Kukul, A. (2008): *Molecular Modeling of Proteins*, vol. 443. Humana Press, Totowa, NJ.
- [216] Berkholz, D.S.; Driggers, C.M.; Shapovalov, M.V.; Dunbrack, R.L.; Karplus, P.A. (2012): Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proc Natl Acad Sci U S A* 2, 449–453.
- [217] Ramachandran, G.N.; Ramakrishnan, C.; Sasisekharan, V. (1963). *J Mol Biol* 7, 95–99.
- [218] Ramakrishnan, C.; Ramachandran, G.N. (1965): Stereochemical Criteria for Polypeptide and Protein Chain Conformations. II Allowed Conformations for a Pair of Peptide Units. *Biophysical Journal*, 909–933.
- [219] Ramachandran, G.N.; Venkatachalam, C.M.; Krimm, S. (1966): Stereochemical Criteria for Polypeptide and Protein Chain Conformations. III Helical and Hydrogen-bonded. *Biophysical Journal*, 849–872.
- [220] Lovell, S.C.; Davis, I.W.; Arendall, W.B.; de Bakker, Paul I. W.; Word, J.M.; Prisant, M.G.; Richardson, J.S.; Richardson, D.C. (2003): Structure validation by C[alpha] geometry: [phis],[psi] and C[beta] deviation. *Proteins* 50, 437–450.
- [221] Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.J.; Chen, V.B.; Jain, S.; Lewis, S.M.; Arendall, W.B.; Snoeyink, J.; Adams, P.D.; Lovell, S.C.; Richardson, J.S.; Richardson, D.C. (2018): MolProbity. More and better reference data for improved all-atom structure validation. *Protein Sci* 1, 293–315.
- [222] Eliel, E.L.; Wilen, S.H.; Mander, L.N. (1994): *Stereochemistry of organic compounds*. Wiley, New York.
- [223] Voronoi, G. (1908): Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les paralléloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelle's Journal)* 134, [EPub].

- [224] Richards, F.M. (1974): The interpretation of protein structures: Total volume, group volume distributions and packing density. *J Mol Biol* 1, 1–14.
- [225] Finney, J.L. (1975): Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol* 4, 721–732.
- [226] Gellatly, B.J.; Finney, J.L. (1982): Calculation of protein volumes: An alternative to the Voronoi procedure. *J Mol Biol* 2, 305–322.
- [227] Pontius, J.; Richelle, J.; Wodak, S.J. (1996): Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *J Mol Biol* 264, 121–136.
- [228] Richards, F.M. (1977): Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng*, 151–176.
- [229] Connolly, M.L. (1985): Computation of molecular volume. *J Am Chem Soc* 5, 1118–1124.
- [230] Pyykko, P.; Atsumi, M. (2009): Molecular double-bond covalent radii for elements Li–E112. *Chemistry* 46, 12770–12779.
- [231] Cordero, B.; Gomez, V.; Platero-Prats, A.E.; Reves, M.; Echeverria, J.; Cremades, E.; Barragan, F.; Alvarez, S. (2008): Covalent radii revisited. *Dalton Trans* 21, 2832–2838.
- [232] Bondi, A. (1964): van der Waals Volumes and Radii. *Journal of physical Chemistry* 3, 441–451.
- [233] Richmond, T.J. (1984): Solvent accessible surface area and excluded volume in proteins. *J Mol Biol* 1, 63–89.
- [234] Gibson, K.D.; Scheraga, H.A. (1987): Volume of the intersection of three spheres of unequal size. A simplified formula. *J. Phys. Chem.* 15, 4121–4122.
- [235] Kang, Y.K.; Némethy, G.; Scheraga, H.A. (1987): Free energies of hydration of solute molecules. 1. Improvement of the hydration shell model by exact computations of overlapping volumes. *J. Phys. Chem.* 15, 4105–4109.
- [236] Carrasco, B.; La García de Torre, J.; Zipper, P. (1999): Calculation of hydrodynamic properties of macromolecular bead models with overlapping spheres. *European Biophysics Journal* 6, 510–515.
- [237] Walton, J.R.; Rivera-Rivera, L.A.; Lucchese, R.R.; Bevan, J.W. (2017): Is there any fundamental difference between ionic, covalent, and others types of bond? A canonical perspective on the question. *Phys Chem Chem Phys* 24, 15864–15869.

- [238] Harpaz, Y.; Gerstein, M.; Chothia, C. (1994): Volume changes on protein folding. *Structure* 7, 641–649.
- [239] Chothia, C. (1975): Structural invariants in protein folding. *Nature* 5498, 304–308.
- [240] Gerstein, M.; Tsai, J.W.; Levitt, M. (1995): The volume of atoms on the protein surface. Calculated from simulation, using Voronoi polyhedra. *J Mol Biol* 5, 955–966.
- [241] Word, J.M.; Lovell, S.C.; LaBean, T.H.; Taylor, H.C.; Zalis, M.E.; Presley, B.K.; Richardson, J.S.; Richardson, D.C. (1999): Visualizing and quantifying molecular goodness-of-fit. Small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 4, 1711–1733.
- [242] Melo, F.; Feytmans, E. (1997): Novel Knowledge-based Mean Force Potential at Atomic Level. *J Mol Biol* 267, 207–222.
- [243] Jones, D.T. (1999): GenTHREADER. An efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 4, 797–815.
- [244] Li, J.; Fang, H. (2016): A comparison of different functions for predicted protein model quality assessment. *J Comput Aided Mol Des* 7, 553–558.
- [245] Benkert, P.; Tosatto, S.C.E.; Schomburg, D. (2008): QMEAN. A comprehensive scoring function for model quality assessment. *Proteins* 1, 261–277.
- [246] Russell, R.B.; Barton, G.J. (1993): The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J Mol Biol* 4, 951–957.
- [247] McGuffin, L.J.; Jones, D.T. (2003): Benchmarking secondary structure prediction for fold recognition. *Proteins* 2, 166–175.
- [248] Swanson, R.; Kagiampakis, I.; Tsai, J.W. (2008): An information measure of the quality of protein secondary structure prediction. *J Comput Biol* 1, 65–79.
- [249] Kim, T.-R.; Yang, J.S.; Shin, S.; Lee, J. (2013): Statistical torsion angle potential energy functions for protein structure modeling. A bicubic interpolation approach. *Proteins* 7, 1156–1165.
- [250] Li, Y.; Liu, H.; Rata, I.; Jakobsson, E. (2013): Building a knowledge-based statistical potential by capturing high-order inter-residue interactions and its applications in protein secondary structure assessment. *J Chem Inf Model* 2, 500–508.
- [251] Lee, B.; Richards, F.M. (1971): The interpretation of protein structures. Estimation of static accessibility. *J Mol Biol* 3, 379–400.
- [252] ISO: ISO/IEC 14882:2017 .

Literaturverzeichnis

- [253] van Rossum, G.; Drake, F.L. (2001): Python Reference Manual (Zuletzt geprüft: 07.06.2016).
- [254] Dubois, P.F.; Hinsen, K.; Hugunin, J. (1996): Numerical Python. Computers in Physics 3, [EPub].
- [255] McKinney, W.: pandas: a Foundational Python Library for Data Analysis and Statistics.
- [256] Raschka, S. (2015): Biopandas. <https://pypi.python.org/pypi/BioPandas/0.1.4> (Zuletzt geprüft: 02.08.2016).
- [257] Hunter, J.D. (2007): Matplotlib: A 2D graphics environment. Computing In Science & Engineering 3, 90–95.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die eingereichte Dissertation selbständig und ohne unzulässige fremde Hilfe verfasst, andere als die in ihr angegebene Literatur nicht benutzt und dass ich alle ganz oder annähernd übernommenen Textstellen sowie verwendete Grafiken, Tabellen und Auswertungsprogramme kenntlich gemacht habe. Außerdem versichere ich, dass die vorgelegte elektronische mit der schriftlichen Version der Dissertation übereinstimmt und die Abhandlung in dieser oder ähnlicher Form noch nicht anderweitig als Promotionsleistung vorgelegt und bewertet wurde.

[Unterschrift]

ORT, DATUM

Danksagung

Zu aller Erst möchte ich mich bei PD. Dr. Iris Thondorf für die Möglichkeit die vorliegende Arbeit in ihrer Arbeitsgruppe zu schreiben bedanken. Durch die vielen Freiheiten die sie mir während der Arbeit gewährte konnte ich weitreichende Ideen entwickeln. Die Besprechungen mit ihr lieferten mir immer wieder neue Ideen zur Erweiterung bzw. Verbesserung bereits entwickelter Programmteile. Außerdem halfen die vielen Diskussionen sowohl während der Entwicklung der Programme als auch beim Schreiben der vorliegenden Dissertation bei der genauen Einschätzung der erhaltenen Ergebnisse.

Des Weiteren bedanke ich mich herzlichst bei Michael Wenk und Lucas Roschig die mir bei der Entwicklung der *k-d* Bäume bzw. der Erstellung der Webseite geholfen haben. Vielen Dank auch an die vielen Studentinnen und Studenten die ich während meiner Zeit in der Arbeitsgruppe betreuen durfte. Durch sie habe ich viele Anregungen sowohl im fachlichen wie auch nicht-fachlichen Bereich erhalten.

Zu guter Letzt möchte ich mich bei meiner Mutter Kathrin Knorr und meinem Onkel Gundolf Stoll bedanken die mich während meines Studiums und meiner Promotion sowohl finanziell als auch psychologisch unterstützt haben. Sie halfen mir mit ihren immer offenen Ohren durch einige schwere Phase während dieser Zeit.

Anhang A1 - Tabellen

Tabelle A.1: Vorkommen aller kanonischen Aminosäuren in den Datensätzen

| Aminosäure | XRD | TRD | TED | MBD | SMD |
|--------------|------------------|-----------------|----------------|------------------|------------------|
| Alanin | 115340 (8,0%) | 20167 (8,3%) | 2258 (7,9%) | 88684 (13,3%) | 51341 (13,2%) |
| Arginin | 73523 (5,1%) | 11365 (4,7%) | 1421 (5,0%) | 49993 (7,5%) | 30196 (7,8%) |
| Asparagin | 64995 (4,5%) | 11912 (4,9%) | 1376 (4,8%) | 15568 (2,3%) | 7950 (2,1%) |
| Aspartat | 86363 (6,0%) | 14989 (6,2%) | 1716 (6,0%) | 40478 (6,1%) | 24768 (6,4%) |
| Cystein | 18441 (1,3%) | 2838 (1,2%) | 324 (1,1%) | 6188 (0,9%) | 3424 (0,9%) |
| Glutamin | 56688 (3,9%) | 9338 (3,8%) | 1259 (4,4%) | 20022 (3,0%) | 12059 (3,1%) |
| Glutamat | 94982 (6,6%) | 14719 (6,1%) | 1725 (6,0%) | 32609 (4,9%) | 20458 (5,3%) |
| Glycin | 104424 (7,2%) | 18827 (7,7%) | 2100 (7,4%) | 62246 (9,3%) | 34690 (9,0%) |
| Histidin | 33172 (2,3%) | 5490 (2,3%) | 628 (2,2%) | 15648 (2,3%) | 9383 (2,4%) |
| Isoleucin | 80402 (5,6%) | 12891 (5,3%) | 1468 (5,1%) | 29210 (4,4%) | 16292 (4,2%) |
| Leucin | 132256 (9,2%) | 21403 (8,8%) | 2525 (8,8%) | 68788 (10,3%) | 39981 (10,3%) |
| Lysin | 86148 (6,0%) | 13505 (5,6%) | 1520 (5,3%) | 13154 (2,0%) | 7675 (2,0%) |
| Methionin | 24520 (1,7%) | 4356 (1,8%) | 525 (1,8%) | 11690 (1,7%) | 6723 (1,7%) |
| Phenylalanin | 59708 (4,1%) | 9930 (4,1%) | 1139 (4,0%) | 19623 (2,9%) | 10736 (2,8%) |
| Prolin | 68004 (4,7%) | 11663 (4,8%) | 1324 (4,6%) | 36110 (5,4%) | 20812 (5,4%) |
| Serin | 88543 (6,1%) | 15257 (6,3%) | 1838 (6,4%) | 35312 (5,3%) | 20234 (5,2%) |
| Threonin | 79677 (5,5%) | 14407 (5,9%) | 1787 (6,3%) | 38573 (5,8%) | 21873 (5,6%) |
| Tryptophan | 22441 (1,6%) | 3997 (1,6%) | 479 (1,7%) | 9507 (1,4%) | 5050 (1,3%) |
| Tyrosin | 54240 (3,6%) | 9547 (3,9%) | 1128 (4,0%) | 14084 (2,1%) | 8012 (2,1%) |
| Valin | 99125 (6,9%) | 16416 (6,8%) | 2010 (7,0%) | 60709 (9,1%) | 35828 (9,2%) |

Anhang

Tabelle A.2: Wasserstoffbrücken-Cluster im Datensatz XRD

In der Tabelle sind gefundene Wasserstoffbrücken mit Sequenzseparationen von mehr als 50 Aminosäuren zusammengefasst. Zusätzlich sind noch die Anzahlen an Strukturen gezeigt die eine gewisse Anzahl von Aminosäuren enthalten angeben. Es ist erkennbar, dass es nur sehr wenige sehr lange Strukturen gibt und dass ein Großteil der gefundenen größeren Sequenzseparationen mit den Längen der Proteinstrukturen übereinstimmt.

| Cluster | Wasserstoffbrücken (Sequenzseparation) | | Strukturen (Kettenlänge) | |
|-------------|--|--------|--------------------------|--------|
| | # | % | # | % |
| 50 - 100 | 57771 | 54,54% | 490 | 8,92% |
| 100 - 150 | 21739 | 20,52% | 894 | 16,27% |
| 150 - 200 | 10048 | 9,49% | 810 | 14,74% |
| 200 - 250 | 6710 | 6,34% | 724 | 13,18% |
| 250 - 300 | 3762 | 3,55% | 662 | 12,05% |
| 300 - 350 | 2213 | 2,09% | 569 | 10,35% |
| 350 - 400 | 1178 | 1,11% | 388 | 7,06% |
| 400 - 450 | 618 | 0,58% | 300 | 5,46% |
| 450 - 500 | 371 | 0,35% | 191 | 3,48% |
| 500 - 550 | 216 | 0,20% | 131 | 2,38% |
| 550 - 600 | 101 | 0,10% | 114 | 2,07% |
| 600 - 650 | 80 | 0,08% | 45 | 0,82% |
| 650 - 700 | 97 | 0,09% | 37 | 0,67% |
| 700 - 750 | 233 | 0,22% | 27 | 0,49% |
| 750 - 800 | 357 | 0,34% | 22 | 0,40% |
| 800 - 850 | 162 | 0,15% | 15 | 0,27% |
| 850 - 900 | 38 | 0,04% | 9 | 0,16% |
| 900 - 950 | 9 | 0,01% | 43 | 0,78% |
| 950 - 1000 | 181 | 0,17% | 4 | 0,07% |
| 1000 - 1050 | 0 | 0,00% | 3 | 0,05% |
| 1050 - 1100 | 0 | 0,00% | 8 | 0,15% |
| 1100 - 1150 | 1 | 0,00% | 1 | 0,02% |
| 1150 - 1200 | 31 | 0,03% | 1 | 0,02% |
| 1200 - 1250 | 3 | 0,00% | 1 | 0,02% |
| 1250 - 1300 | 0 | 0,00% | 2 | 0,04% |
| 1300 - 1350 | 0 | 0,00% | 0 | 0,00% |
| 1350 - 1400 | 0 | 0,00% | 0 | 0,00% |
| 1400 - 1450 | 0 | 0,00% | 0 | 0,00% |
| 1450 - 1500 | 0 | 0,00% | 4 | 0,07% |
| 1500 - 1550 | 0 | 0,00% | 0 | 0,00% |
| 1550 - 1600 | 0 | 0,00% | 0 | 0,00% |

Anhang

Tabelle A.3: Aminosäurezusammensetzungen des Datensatzes XRD im Vergleich zu Datensätzen verschiedener älterer Studien

Die Prozentzahlen geben den Anteil der jeweiligen Aminosäure am gesamten Datensatz an.

| Aminosäure | XRD | Chou [176] | Costantini [180] |
|--------------|---------------|------------|------------------|
| Alanin | 115340 (8,0%) | 228 (9,2%) | 7,7% |
| Arginin | 73523 (5,1%) | 78 (3,2%) | 5,0% |
| Asparagin | 64995 (4,5%) | 133 (5,4%) | 4,5% |
| Aspartat | 86363 (6,0%) | 111 (4,5%) | 5,8% |
| Cystein | 18441 (1,3%) | 54 (2,2%) | 1,8% |
| Glutamin | 56688 (3,9%) | 95 (3,8%) | 3,9% |
| Glutamat | 94982 (6,6%) | 113 (4,6%) | 6,6% |
| Glycin | 104424 (7,2%) | 232 (9,4%) | 7,1% |
| Histidin | 33172 (2,3%) | 74 (3,0%) | 2,4% |
| Isoleucin | 80402 (5,6%) | 106 (4,3%) | 5,7% |
| Leucin | 132256 (9,2%) | 196 (7,9%) | 8,8% |
| Lysin | 86148 (6,0%) | 175 (7,1%) | 6,3% |
| Methionin | 24520 (1,7%) | 28 (1,1%) | 2,1% |
| Phenylalanin | 59708 (4,1%) | 85 (3,4%) | 4,1% |
| Prolin | 68004 (4,7%) | 82 (3,3%) | 4,5% |
| Serin | 88543 (6,1%) | 202 (8,2%) | 6,1% |
| Threonin | 79677 (5,5%) | 156 (6,3%) | 5,5% |
| Tryptophan | 22441 (1,6%) | 44 (1,8%) | 1,5% |
| Tyrosin | 54240 (3,6%) | 100 (4,0%) | 3,5% |
| Valin | 99125 (6,9%) | 181 (7,3%) | 6,9% |

Anhang

Tabelle A.4: Statistische Präferenzen von β -Strängen mit weniger als neun Aminosäuren

Die Tabelle zeigt die positionsspezifischen Aminosäure-Präferenzen von β -Strängen mit weniger als neun Aminosäuren. Die Daten wurden nochmals unterteilt in Werte für verschiedene Strang-Längen (maximal vier, sechs oder acht Aminosäuren). Die Ergebnisse zeigen keine gravierenden Unterschiede zu den Ergebnissen der längeren Stränge (Tabelle 37).

| | 3 oder 4 Aminosäuren | | | | | 5 oder 6 Aminosäuren | | | | | | | 7 oder 8 Aminosäuren | | | | | | | | |
|-----|----------------------|-------------|-------------|-------------|-------------|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | -1 | 1 | M | n | n+1 | -1 | 1 | 2 | M | n-1 | n | n+1 | -1 | 1 | 2 | 3 | M | n-2 | n-1 | n | n+1 |
| ALA | 0,53 | 0,98 | 0,76 | 0,75 | 0,76 | 0,53 | 0,97 | 0,78 | 0,98 | 0,73 | 0,87 | 0,79 | 0,48 | 0,85 | 0,69 | 0,81 | 0,81 | 0,94 | 0,85 | 0,77 | 0,79 |
| ARG | 0,77 | 1,14 | 1,05 | 0,74 | 0,69 | 1,03 | 1,20 | 0,84 | 0,74 | 0,86 | 0,71 | 0,66 | 1,02 | 1,16 | 1,13 | 0,98 | 0,78 | 0,84 | 0,98 | 0,85 | 0,82 |
| ASN | 1,42 | 0,62 | 0,41 | 0,97 | 1,50 | 1,70 | 0,71 | 0,36 | 0,33 | 0,43 | 1,22 | 1,61 | 1,64 | 0,72 | 0,48 | 0,40 | 0,44 | 0,38 | 0,44 | 1,20 | 1,65 |
| ASP | 1,51 | 0,50 | 0,27 | 1,06 | 1,47 | 1,57 | 0,69 | 0,28 | 0,24 | 0,34 | 1,52 | 1,60 | 1,52 | 0,70 | 0,47 | 0,42 | 0,34 | 0,37 | 0,62 | 1,49 | 1,62 |
| CYS | 0,70 | 1,13 | 1,21 | 1,62 | 1,05 | 0,58 | 2,03 | 1,21 | 1,49 | 1,81 | 1,99 | 0,85 | 0,46 | 1,00 | 1,09 | 1,43 | 1,64 | 1,99 | 1,81 | 1,23 | 0,75 |
| GLN | 0,74 | 0,83 | 0,74 | 0,64 | 0,65 | 1,02 | 0,90 | 0,70 | 0,58 | 0,73 | 0,75 | 0,58 | 0,82 | 0,96 | 0,79 | 0,59 | 0,59 | 0,56 | 0,81 | 0,79 | 0,72 |
| GLU | 0,67 | 0,82 | 0,70 | 0,62 | 0,93 | 0,74 | 0,85 | 0,56 | 0,60 | 0,68 | 0,60 | 0,86 | 0,78 | 0,89 | 0,68 | 0,59 | 0,56 | 0,52 | 0,81 | 0,65 | 0,91 |
| GLY | 2,98 | 0,74 | 0,40 | 0,54 | 1,63 | 2,85 | 0,80 | 0,44 | 0,51 | 0,53 | 0,68 | 2,03 | 3,58 | 0,83 | 0,38 | 0,52 | 0,58 | 0,56 | 0,67 | 0,84 | 1,74 |
| HIS | 1,14 | 0,97 | 0,87 | 1,08 | 0,92 | 1,01 | 1,13 | 0,72 | 0,91 | 0,91 | 1,33 | 1,19 | 1,10 | 1,17 | 0,99 | 1,03 | 0,77 | 1,07 | 1,13 | 1,42 | 0,94 |
| ILE | 0,67 | 1,26 | 1,84 | 1,38 | 0,79 | 0,60 | 1,02 | 2,31 | 2,05 | 1,92 | 1,08 | 0,54 | 0,35 | 0,89 | 1,71 | 1,93 | 1,99 | 1,74 | 1,47 | 0,81 | 0,50 |
| LEU | 0,68 | 0,76 | 1,11 | 1,00 | 0,97 | 0,52 | 0,79 | 1,39 | 1,30 | 1,19 | 0,87 | 0,71 | 0,49 | 0,67 | 1,15 | 1,25 | 1,57 | 1,30 | 0,96 | 0,76 | 0,63 |
| LYS | 1,12 | 1,00 | 0,86 | 0,62 | 0,68 | 1,26 | 1,08 | 0,70 | 0,59 | 0,70 | 0,59 | 0,73 | 1,21 | 1,47 | 1,00 | 0,60 | 0,54 | 0,57 | 0,84 | 0,78 | 1,12 |
| MET | 0,59 | 1,18 | 1,07 | 1,08 | 0,60 | 0,55 | 0,91 | 0,93 | 1,21 | 0,96 | 0,79 | 0,61 | 0,52 | 1,51 | 0,91 | 1,37 | 1,06 | 1,09 | 1,06 | 1,08 | 0,53 |
| PRO | 0,93 | 0,75 | 0,39 | 0,96 | 1,56 | 1,10 | 0,93 | 0,47 | 0,22 | 0,45 | 1,01 | 1,95 | 0,95 | 1,11 | 0,72 | 0,19 | 0,32 | 0,29 | 0,52 | 1,10 | 1,98 |
| PHE | 0,65 | 1,32 | 1,69 | 1,23 | 0,72 | 0,47 | 1,12 | 1,36 | 1,63 | 1,67 | 1,12 | 0,61 | 0,37 | 0,84 | 1,44 | 1,57 | 1,75 | 1,73 | 1,40 | 0,97 | 0,51 |
| SER | 0,75 | 1,13 | 0,80 | 1,25 | 1,09 | 0,88 | 1,05 | 0,74 | 0,66 | 0,80 | 1,40 | 1,23 | 0,85 | 1,21 | 0,76 | 0,76 | 0,72 | 0,77 | 0,81 | 1,53 | 1,40 |
| THR | 0,94 | 1,22 | 1,24 | 1,36 | 1,25 | 0,87 | 1,27 | 1,21 | 1,13 | 1,21 | 1,20 | 0,92 | 1,04 | 1,31 | 1,29 | 1,15 | 1,14 | 1,07 | 1,45 | 1,38 | 0,94 |
| TRP | 0,57 | 1,50 | 1,50 | 1,28 | 0,72 | 0,46 | 1,18 | 1,51 | 1,49 | 1,63 | 1,09 | 0,63 | 0,34 | 1,35 | 1,24 | 1,10 | 1,16 | 1,29 | 1,23 | 0,97 | 0,57 |
| TYR | 0,71 | 1,53 | 1,75 | 1,26 | 0,54 | 0,63 | 1,27 | 1,37 | 1,63 | 1,53 | 0,99 | 0,70 | 0,53 | 1,21 | 1,36 | 1,64 | 1,44 | 1,58 | 1,52 | 1,12 | 0,49 |
| VAL | 0,87 | 1,47 | 2,08 | 1,54 | 0,81 | 0,58 | 1,28 | 2,25 | 2,36 | 2,02 | 1,26 | 0,64 | 0,46 | 1,13 | 2,04 | 2,29 | 2,06 | 2,27 | 1,68 | 1,02 | 0,58 |

Anhang

Tabelle A.5: Seitenketten-Torsionswinkel, die zur Erstellung von Diagrammen genutzt wurden

Die Tabelle fasst alle genutzten Seitenketten-Torsionswinkel zusammen, für die im Zuge der Analyse der Torsionswinkel (Kapitel 2.4.2 Analyse von Torsionswinkeln) neue Diagramme erstellt wurden. Dabei wurde für zwei aufeinanderfolgende Winkel jeweils ein *Plot* erstellt, d.h. für Glutamin existiert ein Chi1/Chi2- sowie ein Chi2/Chi3-Diagramm. Torsionswinkel innerhalb von aromatischen Ringsystemen wurden genauso nicht beachtet wie Winkel, bei denen keine Unterscheidung zwischen verschiedenen Endpunkten, z.B. Chi3 von Glutamat, getroffen werden konnte.

| Aminosäure | Chi1 | Chi2 | Chi3 | Chi4 |
|---------------------|---------------------------------------|---|---|---|
| Arginin | N-C α -C β -C γ | C α -C β -C γ -C δ | C β -C γ -C δ -N ϵ | C γ -C δ -N ϵ -C ζ |
| Asparagin | N-C α -C β -C γ | C α -C β -C γ -N δ 2 | | |
| Glutamat | N-C α -C β -C γ | C α -C β -C γ -C δ | | |
| Glutamin | N-C α -C β -C γ | C α -C β -C γ -C δ | C β -C γ -C δ -N ϵ 2 | |
| Histidin | N-C α -C β -C γ | C α -C β -C γ -N δ 1 | | |
| Isoleucin | N-C α -C β -C γ 1 | C α -C β -C γ -C δ 1 | | |
| Leucin | N-C α -C β -C γ | C α -C β -C γ -C δ 1 | | |
| Lysin | N-C α -C β -C γ | C α -C β -C γ -C δ | C β -C γ -C δ -C ϵ | |
| Methionin | N-C α -C β -C γ | C α -C β -C γ -S δ | C β -C γ -S δ -C ϵ | C γ -C δ -C ϵ -N ζ |
| Phenylalanin | N-C α -C β -C γ | C α -C β -C γ -C δ 1 | | |
| Tryptophan | N-C α -C β -C γ | C α -C β -C γ -C δ 1 | | |
| Tyrosin | N-C α -C β -C γ | C α -C β -C γ -C δ 1 | | |

Anhang

Tabelle A.6: Mögliche Server des zu entwickelnden Netzwerks

| Serverkürzel | Englischer Name | Beschreibung |
|---------------------|---|--|
| VAProTS | <i>Validation and Analysis of Protein Tertiary Structures</i> | Server zur Analyse und Validierung von Protein Strukturen |
| VAProIN | <i>Validation and Analysis of Protein Interaction Networks</i> | Server zur Erstellung von Interaktionskarten für Proteinstrukturen |
| VAData | <i>Validation and Analysis Database</i> | Datenbank zur Sammlung von Validierungsergebnissen von Kristall-Strukturen |
| VADoc | <i>Validation and Analysis Documentation</i> | Dokumentation aller Server/Programme des Netzwerks |
| VASat | <i>Validation and Analysis Statistics</i> | Sammlung von Statistiken aller genutzten Datensätze |
| VAPara | <i>Validation and Analysis Parameters</i> | Sammlung der zugrundeliegenden Standard-Parameter der einzelnen Server |
| VANuAcS | <i>Validation and Analysis Of Nucleic Acid Structures</i> | Server zur Analyse und Validierung von Nukleinsäure-Strukturen |
| VAProC | <i>Validation and Analysis of Protein Clusters</i> | Server zur Erstellung einer vergleichenden Bewertung von Proteinstrukturen mit ähnlichen Sequenzen |
| VASOM | <i>Validation and Analysis of Small Organic Molecules</i> | Server zur Analyse und Validierung von Liganden in Proteinstrukturen |
| VATraRP | <i>Validation and Analysis of Transmembrane Regions of Proteins</i> | Server zur Identifikation, Analyse und Validierung von Transmembran-Regionen in Proteinstrukturen |
| VAProQS | <i>Validation and Analysis of Protein Quarternary Structures</i> | Server zur Analyse und Validierung von Protein-Protein-Interaktionen |
| VAProLI | <i>Validation and Analysis of Protein Ligand Interactions</i> | Server zur Analyse und Validierung von Protein-Ligand-Wechselwirkungen |
| VApomPS | <i>Validation and Analysis Of posttranslational Modifications in Protein Structures</i> | Server zur Analyse und Validierung von post-translationalen Modifikationen |
| VAProNAI | <i>Validation and Analysis of Protein and Nucleic Acid Interactions</i> | Server zur Analyse und Validierung von Protein-DNA/RNA-Interaktionen |

Anhang A2 - Abbildungen

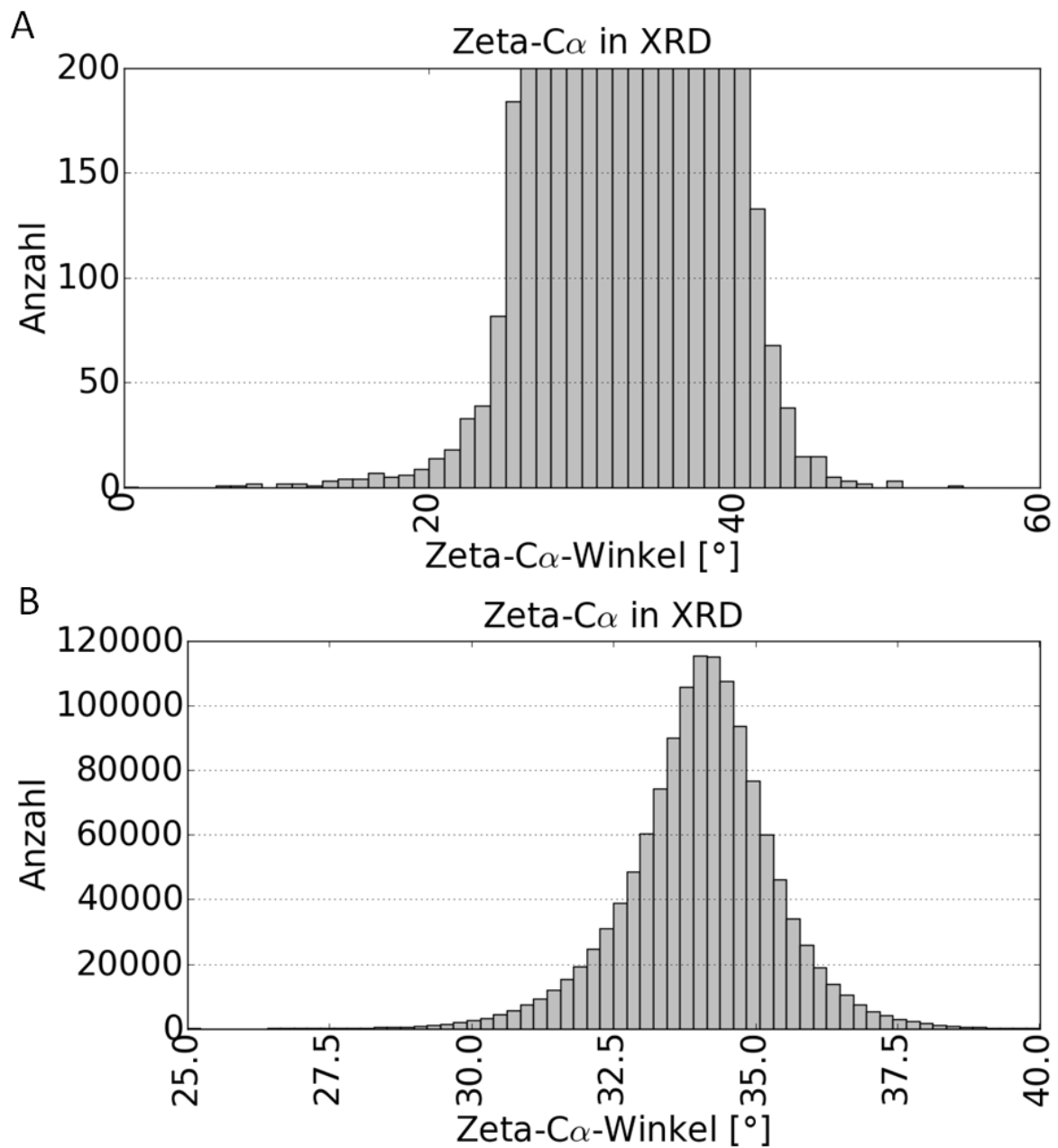


Abb. A. 1: Vergrößerung der Verteilungen der Zeta-C α -Winkel im Datensatz XRD

Die beiden Histogramme zeigen Vergrößerungen von Abb. 2 der Verteilung der Zeta-C α -Winkel der Aminosäuren im Datensatz XRD. Die Histogramme wurden entlang der y-Achse (A) oder der x-Achse (B) vergrößert, um einen genaueren Blick auf die Daten zu ermöglichen. In beiden Fällen ist erkennbar, dass sich die Verteilung zu großen Teilen im Bereich von 20° bis 50° bewegt.

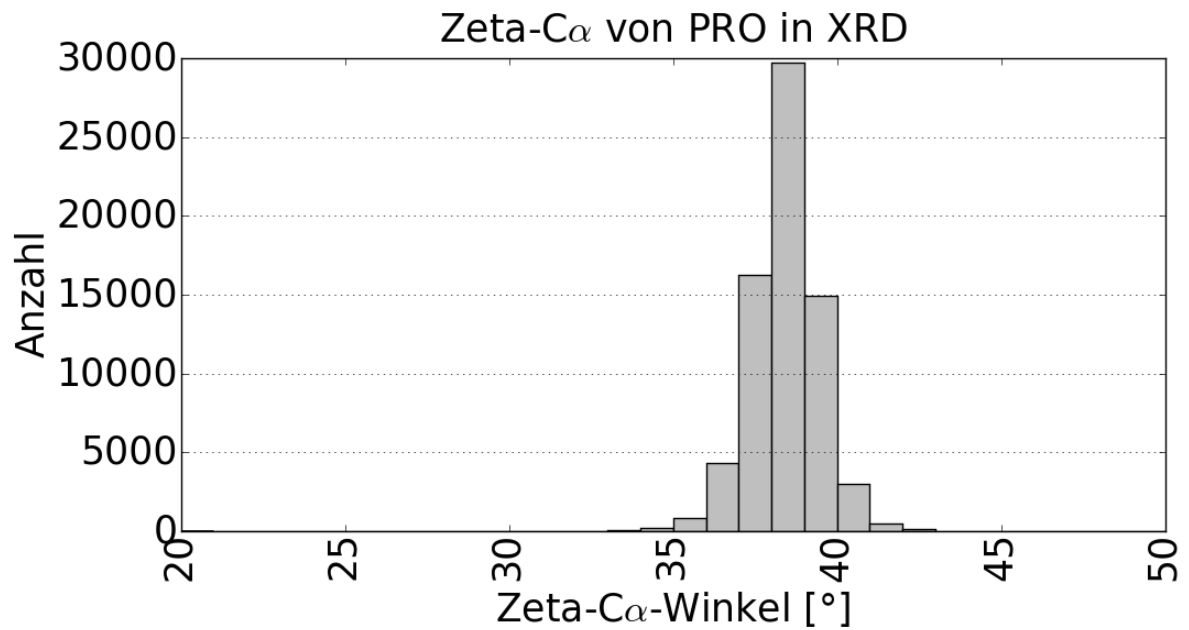


Abb. A. 2: Verteilung der Zeta-C α -Winkel von Prolin im Datensatz XRD

Das Histogramm zeigt die Verteilung der berechneten Zeta-C α -Winkel aller Proline im Datensatz XRD. Es ist gut das Maximum im Bereich von 38° zu erkennen, das sich deutlich vom Standard-Wert von 34° der restlichen Aminosäuren unterscheidet.

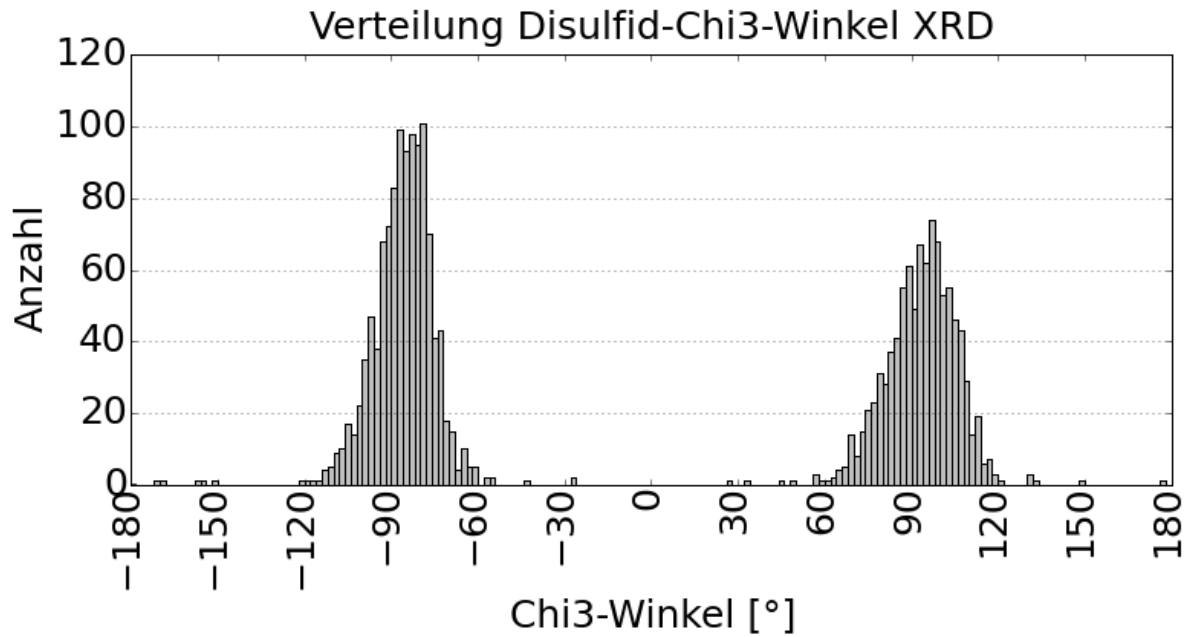


Abb. A. 3: Verteilung der berechneten Chi3-Torsionswinkel von möglichen Disulfidbindungen

Das Histogramm zeigt die berechneten Chi3-Torsionswinkel von allen möglichen Disulfidbindungen des Datensatzes XRD. Es sind zwei Maxima zu erkennen, die jeweils der linksgängigen Konformation ($\sim 85^\circ$) und der rechtsgängigen Konformation ($\sim 95^\circ$) zuordnen lassen. Insgesamt sind die linksgängigen Konformationen häufiger und zeigen eine engere Verteilung als die rechtsgängigen Konformationen. Für beide Konformationen existieren Bindungen mit Winkeln die stark vom Optimum abweichen ($\pm 180^\circ$) und bei einer Bestimmung von Disulfidbindungen ausgeschlossen werden sollten [100].

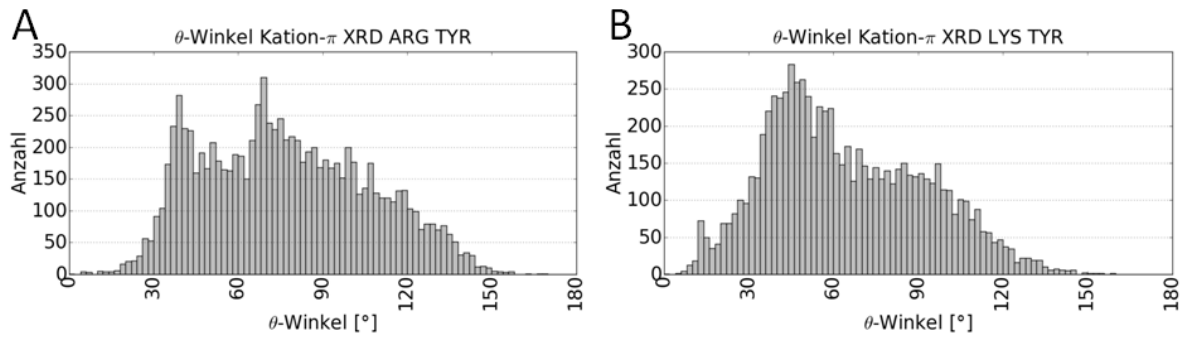


Abb. A. 4: θ -Winkel von Kation- π -Wechselwirkungen mit Tyrosin

Die beiden Histogramme zeigen die berechneten θ -Winkel für Kation- π -Wechselwirkungen von Arginin (A) und Lysin (B) mit Tyrosin. Für beide Verteilungen ist eine starke Tendenz zu Winkeln im Bereich von 0° bis 90° zu erkennen.

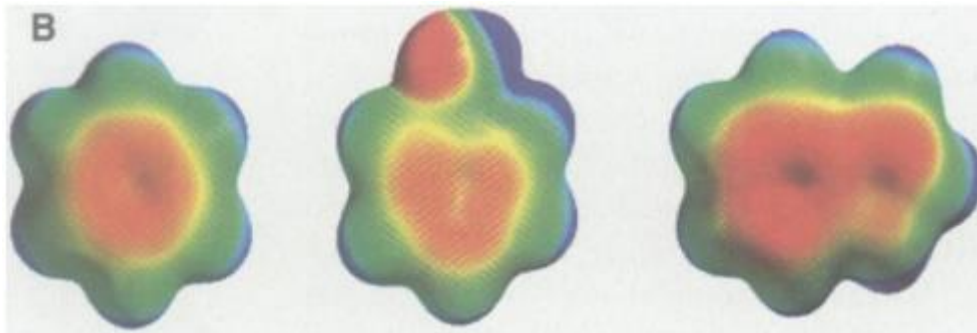


Abb. A. 5: Elektrostatistische Oberflächenpotentiale aromatischer Systeme [125]

Die Abbildung wurde der Arbeit von Dougherty [125] entnommen und zeigt die von ihm berechneten elektrostatischen Oberflächenpotentiale von aromatischen Aminosäuren. Die Oberflächen sind farbkodiert von Rot (-25 kcal/mol) zu Lila (25 kcal/mol). Beim Phenol-Ring (Mitte) ist sehr gut das zweite negative elektrostatische Potential im Bereich der Hydroxyl-Gruppe zu erkennen.

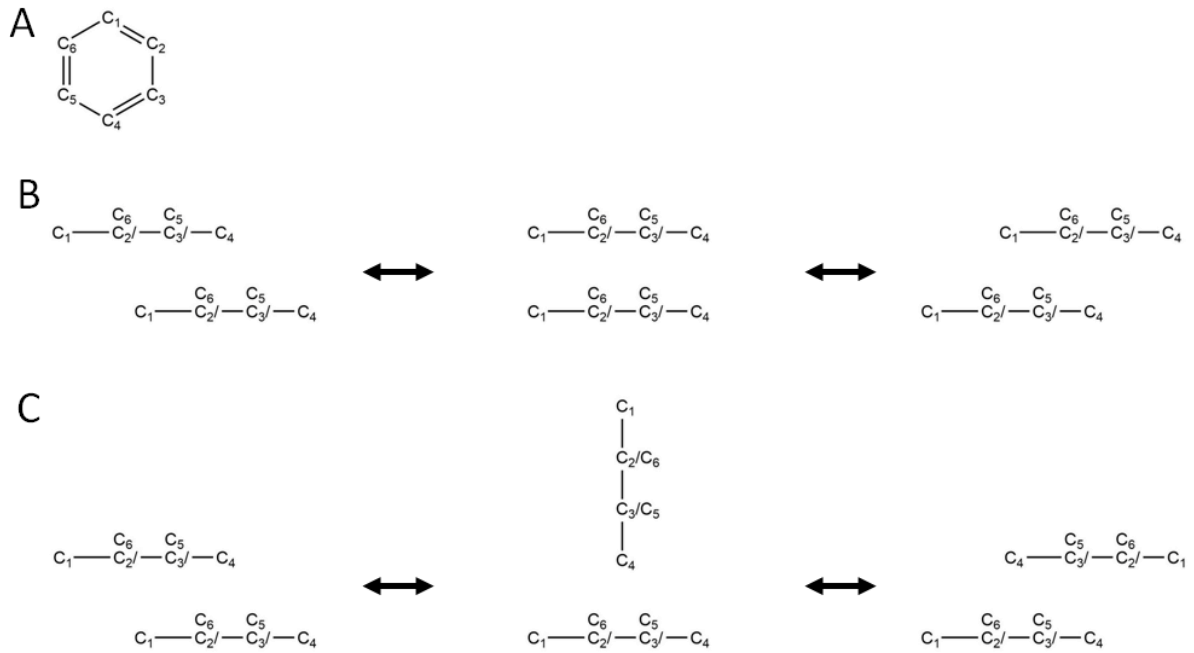


Abb. A. 6: Niedrig-Energieübergänge von π - π -Wechselwirkungen des Benzen-Dimers in der Gasphase

Die Abbildung zeigt schematisch die Niedrig-Energieübergänge von *parallel displaced* Anordnungen Benzen-Dimeren in der Gasphase. Ausgehend von der Nomenklatur in (A) gibt es zwei Möglichkeiten wie sich die Orientierung der beiden Dimere zueinander ändern kann. Dies erfolgt entweder über eine um 30° verdrehte *sandwich* Anordnung (B) oder über eine *T-stacked* Anordnung (C).

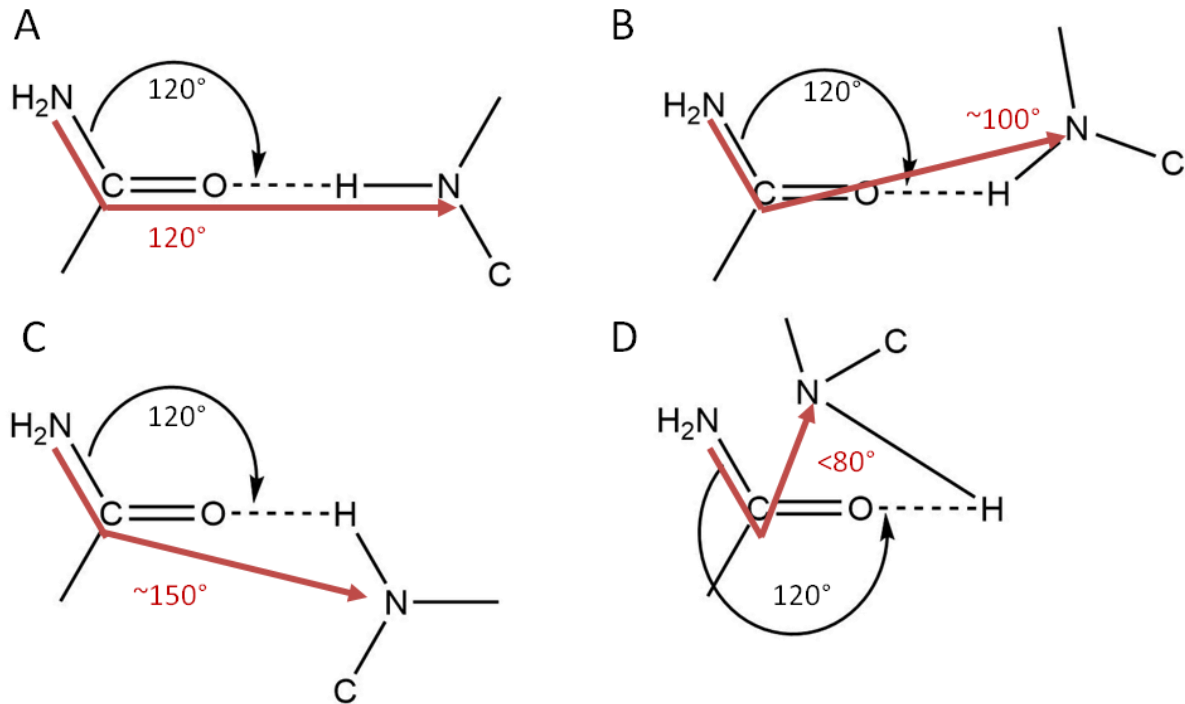


Abb. A. 7: Schematische Darstellung verschiedener C-O-N-Winkel von Wasserstoffbrücken des Proteinrückgrates

Die Abbildung zeigt schematisch vier verschiedene C-O-N-Winkel in Abhängigkeit der Orientierung des Amid-Stickstoffes zur Carboxyl-Gruppe. In allen vier Fällen beträgt der Winkel der Wasserstoffbrücke (N-C-H-Winkel, schwarzer Bogen und Schrift) 120°, der berechnete C-O-N-Winkel (rote Pfeile und Schrift) weicht jedoch teilweise deutlich davon ab. In (D) ist ein Extremfall gezeigt, der verdeutlicht, dass der C-O-N-Winkel mindestens 80° annehmen muss, da ansonsten der Abstand zwischen Carboxyl-Kohlenstoff und dem Stickstoff-Atom ($d(C,N)$) kleiner wird als der Abstand zwischen Carboxyl-Sauerstoff und dem Stickstoff-Atom ($d(O,N)$).

Anhang

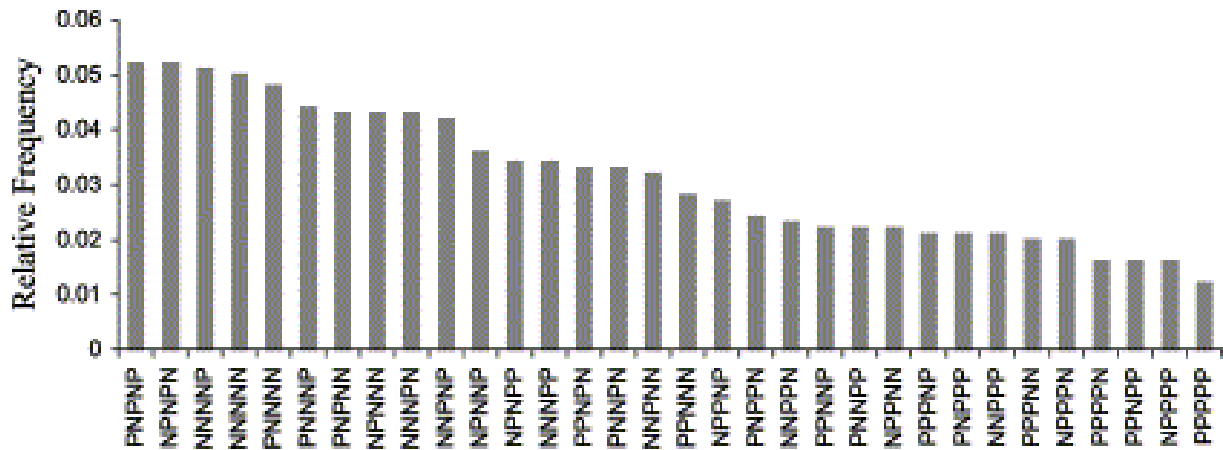


Abb. A. 8: Pentapeptidmuster in β -Strängen [198]

Die Abbildung wurde der Arbeit von Mandel-Gutfreund und Gregoret [198] entnommen und zeigt die von ihnen gefundenen Pentapeptidmuster von β -Strängen ihres Datensatzes. Man kann erkennen, dass die periodischen Pentapeptide am häufigsten vorkommen, jedoch sind Pentapeptide ohne polare Aminosäuren oder mit nur polaren Aminosäuren an den Enden insgesamt deutlich häufiger.

Anhang

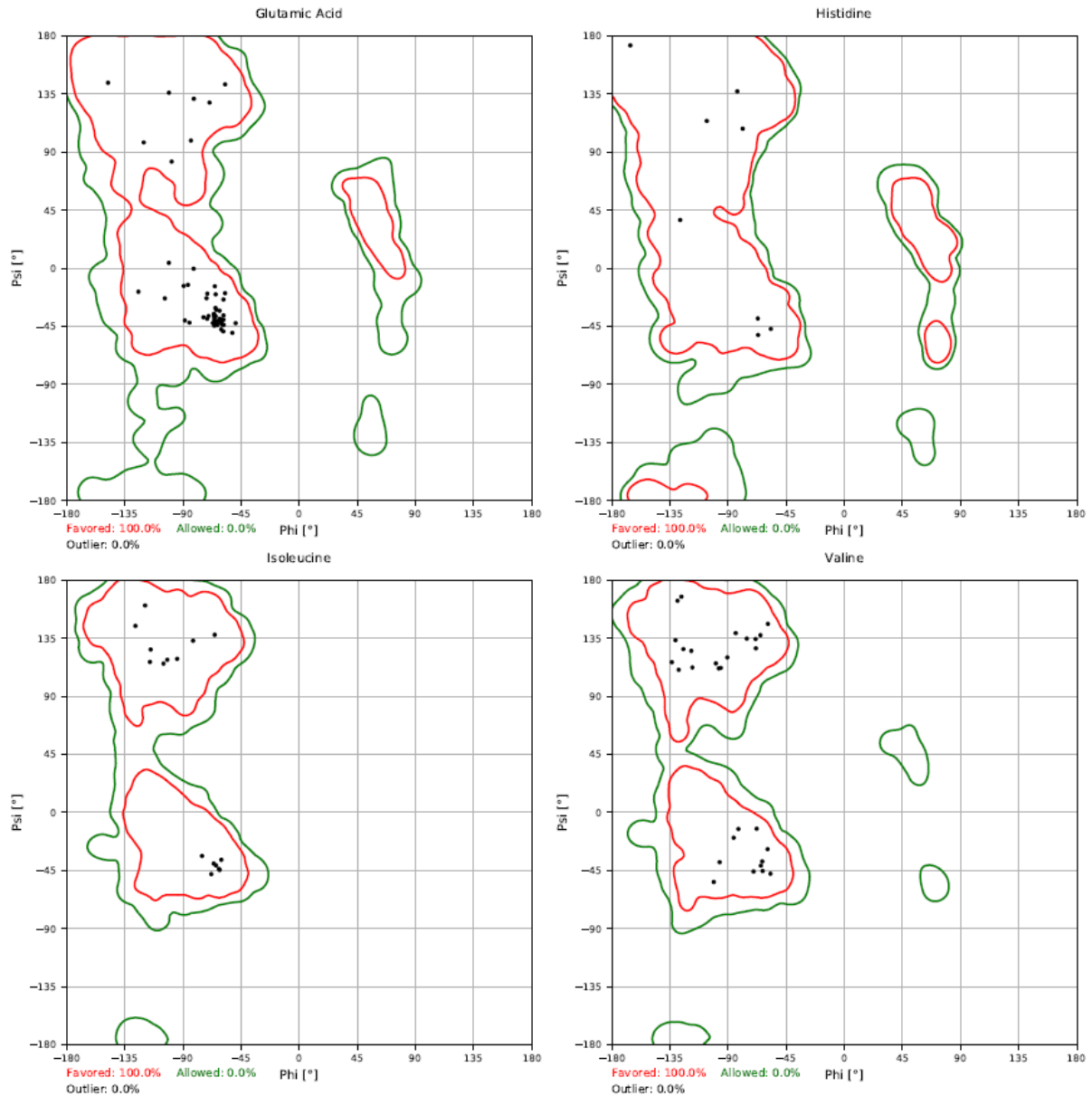


Abb. A.9: Phi/Psi-Plots für Glutamat, Histidin, Isoleucin und Valin

Die vier Diagramme zeigen die neu-entwickelten Phi/Psi-Diagramme für die Aminosäuren Glutamat (links oben), Histidin (rechts oben), Isoleucin (links unten) sowie Valin (rechts unten). Jeder Plot weist individuelle Unterschiede im Vergleich zum klassischen Diagramm (Abb. 26) auf. Bei Glutamat fehlt der favorisierte Bereich im dritten Quadranten, während bei Histidin der favorisierte Bereich im zweiten Quadranten deutlich vergrößert ist und ein kleiner zusätzlicher favorisierter im vierten Quadranten entsteht. Dieser entspricht einem *Cluster* der neu-definierten *Turns* (siehe Verfeinerung von *tight Turns*). Bei Isoleucin und Valin fehlt ebenfalls der favorisierte Bereich im dritten Quadranten und die erlaubten Regionen im ersten Quadranten fallen ganz (Isoleucin) oder nahezu komplett (Valin) weg. Zusätzlich ist der favorisierte Bereich im zweiten Quadranten kleiner und es fehlt die Ausbuchtung im Bereich von $45^\circ < \text{Psi} < 80^\circ$ und $-90^\circ < \text{Phi} < -110^\circ$.

Anhang A3 - Elektronischer Anhang

Lebenslauf

| | |
|----------------|--|
| Name | Christoph Knorr |
| E-Mail-Adresse | christoph.knorr@biochemtech.uni-halle.de |
| Geburtsdatum | 20.11.1987 |
| Geburtsort | Magdeburg |

Ausbildung

| | |
|-------------------|--|
| 10/2015 – 07/2020 | Promotion zum Doctor rerum naturalis Martin-Luther-Universität Halle (Saale) Thema des Promotionsvorhabens: "Entwicklung eines Servers zur Analyse, Bewertung und Verfeinerung von Proteinstrukturen" |
| 10/2013 - 09/2015 | Master of Science in Biochemie Martin-Luther-Universität Halle (Saale) Thema der Abschlussarbeit: "Entwicklung eines Programms zur Bewertung von Proteinstrukturen" |
| 10/2010 – 09/2013 | Bachelor of Science in Biochemie Martin-Luther-Universität Halle (Saale) Thema der Abschlussarbeit: "Charakterisierung von mitochondrialen Rezeptormechanismen bei der Kalziumregulation in neuronalen Zellen am genetischen Mausmodell" |
| 10/2007 – 05/2010 | Bachelor of Science in Biosystemtechnik Otto-von-Guericke Universität Magdeburg Nicht abgeschlossen |

Lebenslauf

07/1998 – 04/2007

Werner-von Siemens-Gymnasium Magdeburg

Abschluss: Abitur

Sprachkenntnisse

| | |
|-------------|--------------------|
| Deutsch | Muttersprache |
| Englisch | Verhandlungssicher |
| Französisch | Grundkenntnisse |
| Latein | Grundkenntnisse |

IT-Kenntnisse

| | | |
|---------------------|---|-----------------|
| Software | MS Office (Word, Excel, Power Point) | Fortgeschritten |
| | Amber | Grundkenntnisse |
| Programmiersprachen | Python | Fortgeschritten |
| | Matlab | Fortgeschritten |
| | C++ | Fortgeschritten |
| | php/ javascript | Fortgeschritten |
| | R | Grundkenntnisse |
| | Bash-Programmierung | Grundkenntnisse |

Unterschrift

Datum