Analyse von Form, Eigenschaften und „Druggability“ von Proteinbindetaschen

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Biowissenschaften
der Johann Wolfgang Goethe-Universität
Frankfurt am Main

von
Martin Weisel
aus Gießen

Frankfurt 2009
(D30)

Vom Fachbereich Biowissenschaften der
Johann Wolfgang Goethe-Universität Frankfurt als Dissertation angenommen.

Dekan:	Prof. Dr. Volker Müller
Erster Gutachter:	Prof. Dr. Gisbert Schneider
Zweiter Gutachter:	Prof. Dr. Joachim Engels
Datum der Disputation:	……

Inhaltsverzeichnis

Abkürzungsverzeichnis

Å	Ångström
AIDS	Erworbenes Immundefektsyndrom (Aquired Immune Deficiency Syndrome)
ANN	Künstliches neuronales Netzwerk (Artificial Neural Network)
Da	Dalton
GDH	Glutamat Dehydrogenase
GNG	Wachsendes neuronales Gas (Growing Neural Gas)
HIV	Humanes Immundefizienz-Virus
IC_{50}	Konzentration eines Wirkstoffes, welche für eine 50%ige Inhibition benötigt wird.
MD	Moleküldynamiksimulation
MSQE	Mittlerer quadratischer Quantisierungsfehler
NMR	Kernmagnetresonanz (Nuclear Magnetic Resonance)
PDB	Protein Datenbank
PPP	Potentieller Pharmakophorpunkt
QF	Quantisierungsfehler
RNA	Ribonukleinsäure
S	Svedberg
s	Sekunde(n)
SDF	Dateiformat für Molekülstrukturen (Structure Data File)
SOM	Selbstorganisierende Karte (Self Organizing Map)
SSSR	Kleinster Satz kleinster Ringe (Smallest Set of Smallest Rings)

Zusammenfassung

Kenntnisse über die dreidimensionale Struktur therapeutisch relevanter Zielproteine bieten wertvolle Informationen für den rationalen Wirkstoffentwurf. Die stetig wachsende Zahl aufgeklärter Kristallstrukturen von Proteinen ermöglicht eine qualitative und quantitative rechnergestützte Untersuchung von spezifischen Protein-Liganden Wechselwirkungen.

Im Rahmen dieser Arbeit wurden neue Algorithmen für die Identifikation und den Ähnlichkeitsvergleich von Proteinbindetaschen und ihren Eigenschaften entwickelt und in dem Programm *PocketomePicker* implementiert. Die Software gliedert sich in die Module *PocketPicker*, *PocketShapelets* und *PocketGraph*. Ferner wurde in dieser Arbeit die Methode *ReverseLIQUID* reimplementiert und für das strukturbasierte Virtuelle Screening angewendet.

Die Methode ***PocketPicker*** ermöglicht die Vorhersage potentieller Bindetaschen auf Proteinoberflächen. Diese Technik implementiert einen geometrischen Ansatz auf Basis „künstlicher Gitter“ zur Identifikation zusammenhängender vergrabener Bereiche der Proteinoberfläche als Orte möglicher Ligandenbindestellen. Die Methode erreicht eine korrekte Vorhersage der tatsächlichen Bindetasche für 73% der Einträge eines repräsentativen Datensatzes von Proteinstrukturen. Für 90% der Proteinstrukturen wird die tatsächlich Ligandenbindestelle unter den drei wahrscheinlichsten vorhergesagten Taschen gefunden. *PocketPicker* übertrifft die Vorhersagequalität anderer Algorithmen und ermöglichte Taschenidentifikationen auf *apo*-Strukturen ohne signifikante Einbußen des Vorhersageerfolges.
PocketPicker erlaubt den alignmentfreien Ähnlichkeitsvergleich von Bindetaschenformen durch die Kodierung berechneter Bindevolumen als Korrelationsdeskriptoren. Dieser Ansatz wurde erfolgreich für Funktionsvorhersage von Bindetaschen aus Homologiemodellen von APOBEC3C (engl. *apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3*) und der Glutamat Dehydrogenase des Malariaerregers *Plasmodium falciparum* angewendet. Diese beiden Projekte wurden in Zusammenarbeit mit Kollaborationspartnern durchgeführt. Zudem wurden *PocketPicker* Korrelations-

deskriptoren erfolgreich für die automatisierte Konformationsanalyse der enzymatischen Tasche von Aldose Reduktase angewendet.

Für detaillierte Analysen der Form und der physikochemischen Eigenschaften von Proteinbindetaschen wurde in dieser Arbeit die Methode ***PocketShapelets*** entwickelt. Diese Technik ermöglicht strukturelle Alignments von extrahierten Bindevolumen durch Zerlegungen der Oberfläche von Proteinbindetaschen. Die Überlagerung gelingt durch die Identifikation strukturell ähnlicher Oberflächenkurvaturen zweier Taschen. *PocketShapelets* wurde erfolgreich zur Analyse der funktionellen Ähnlichkeit von Bindetaschen verwendet, die auf Betrachtungen physikochemischer Eigenschaften basiert.

Zur Analyse der topologischen Vielfalt von Bindetaschengeometrien wurde in dieser Arbeit die Methode *PocketGraph* entwickelt. Dieser Ansatz nutzt das Konzept des sog. „Wachsenden neuronalen Gases“ aus dem Bereich des maschinellen Lernens für eine automatische Extraktion des strukturellen Aufbaus von Bindetaschen. Ferner ermöglicht diese Methode die Zerlegung einer Bindestelle in ihre Subtaschen.

Die von *PocketPicker* charakterisierten Taschenvolumen bilden die Grundlage für die Methode ***ReverseLIQUID***. Dieses Programm wurde in dieser Arbeit weiterentwickelt und im Rahmen einer Kooperation zur Identifikation eines Inhibitors der Serinprotease HtrA des Erregers *Helicobacter pylori* verwendet. Mit *ReverseLIQUID* konnte ein strukturbasiertes Pharmakophormodell für das virtuelle Screening erstellt werden. Dieser Ansatz ermöglichte die Identifikation potenten Substanz (IC_{50} = 26 µM) für das Zielprotein.

Summary

Knowledge of the three-dimensional structure of therapeutically relevant target proteins provides valuable information for rational drug design. The constantly increasing numbers of available crystal structures enable qualitative and quantitative analysis of specific protein-ligand interactions *in silico*.

In this work, novel algorithms for the identification and the comparison of protein binding sites and their properties were developed and combined in the program *PocketomePicker*. The software combines the modules *PocketPicker*, *PocketShapelets* and *PocketGraph*. Furthermore, the method *ReverseLIQUID* was re-implemented in this work and used for the structure-based virtual screening with a cooperation partner.

The method ***PocketPicker*** is designed for the prediction of potential binding sites on protein surfaces. The technique implements a geometric approach based on the concept of "artificial grids" for the identification of continuous buried regions of the protein surface that might act as potential ligand binding sites. The method yields correct predictions of the actual binding site for 73% of the entries in a representative data set of protein structures. For 90% of the proteins the actual binding site is found among the top three predicted binding pockets. *PocketPicker* exceeds the predictive quality of other algorithms and enables correct binding site identifications on *apo* structures without significant drops of the prediction success
PocketPicker enables alignment-free comparisons of binding site shapes by encoding extracted binding volumes as correlation vectors. This approach was used for successful predictions of binding site functionality for homology models of APOBEC3C (*apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3*) and the glutamate dehydrogenase of the malaria pathogen *Plasmodium falciparum*. These projects were carried out with collaboration partners. Furthermore, *PocketPicker* correlation descriptors were used for automated analysis of binding site conformations of aldose reductase active sites.

The method ***PocketShapelets*** was implemented in this work for detailed analysis of shapes and physicochemical properties of protein binding sites. This approach enables structural alignments of extracted binding volumes by surface decomposition of protein binding sites. The structural superposition is achieved by identification of structurally similar surface curvatures of different binding pockets. *PocketShapelets* was successfully used for the analysis of functional similarity of binding sites based on observations of physicochemical properties.

PocketGraph was developed for the analysis of the structural diversity of binding site geometries. This approach uses the "Growing Neural Gas" concept used in machine learning for an automated extraction of the structural organization of binding sites. Furthermore, the method enables the decomposition of binding sites into subpockets.

The pocket volumes characterized by *PocketPicker* are the foundation of another program called ***ReverseLIQUID***. This method was refined in this work and used for the identification of a *Helicobacter pylori* serine protease HtrA inhibitor. This project was performed with a collaboration partner. A receptor-based pharmacophore model was derived using *ReverseLIQUID* and used for virtual screening. This approach led to the identification of a potent substance (IC_{50} = 26 µM) for the target structure.

1 Einleitung

1.1 Der Begriff des Rezeptors und die Wechselwirkung zwischen Proteinen und Liganden

„Corpora non agunt nisi fixata“

(Paul Ehrlich, 1913)

In Anlehnung an den alchimistischen Leitsatz *„Corpora non agunt nisi soluta“* stellte Paul Ehrlich fest, dass Substanzen nicht wirken, wenn sie nicht gebunden werden (Fuhrmann, 1994). Grundlage dieser Beobachtung ist das Konzept biochemischer Rezeptoren, dass Ende des 19. Jahrhunderts unabhängig von Paul Ehrlich und John Newport Langley entwickelt wurde. Während Langley eine „**rezeptive Substanz**“[1] beschrieb, die die Auslösung oder Hemmung von Kontraktionen einer Muskelzelle durch die Zugabe von Nikotin oder Curare beeinflusste (Langley & Dickinson, 1889; Langley, 1905), beschäftigte sich Ehrlich mit der hohen Spezifität der Antigen-Antikörper-Reaktion. Auf Basis dieser Arbeiten beschrieb er membranständige Rezeptoren als „**Seitenketten“**, die aufgrund chemischer und sterischer Eigenschaften spezifische Wechselwirkungen mit bestimmten Antikörpern ausbilden können (Ehrlich, 1897).

Obwohl weder Ehrlich noch Langley Rezeptoren analytisch nachweisen konnten, vermuteten sie Interaktionen zwischen Liganden und makromolekularen Strukturen als Ursache für die Spezifität biochemischer Reaktionen (Fuhrmann, 1994). Die Prinzipien, mit denen sich Moleküle hochspezifisch erkennen, waren zuvor von Emil Fischer als das **„Schlüssel-Schloss-Prinzip“** hypothetisch beschrieben worden (Fischer, 1894). Dieses formuliert die räumliche und chemische Komplementarität zweier Moleküle als Notwendigkeit für eine biologische Funktion. Das Konzept wurde von Emil Fischer bei der Untersuchung der enzymatischen Spaltung von Glucosiden entdeckt:

[1] Die von Langley charakterisierte „rezeptive Substanz“ beschreibt den nikotinischen Acetylcholinrezeptor.

„Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glucosid wie Schloss und Schlüssel zu einander passen müssen, um eine chemische Wirkung auf einander ausüben zu können."

(Emil Fischer, 1894)

Als Weiterentwicklung des Schlüssel-Schloss-Prinzip wurde die **„Induced-Fit-Theorie"** („induzierte Passform") von Daniel E. Koshland vorgestellt (Koshland, 1958). Dieses Modell berücksichtigt die konformationelle Flexibilität, sowohl des Liganden als auch des Rezeptors, die bei der Annäherung der Moleküle auftritt und die Ausbildung eines Komplexes ermöglicht. Die Konformationsänderung des Rezeptors wird durch den Liganden ausgelöst, der als flexibler Schlüssel wirkt. Die Anpassung einer Proteinbindetasche auf einen passenden Liganden durch Induced-Fit Effekte ist für die unterschiedlichen Rezeptoren mehr oder weniger stark ausgeprägt. Eine übermäßige Flexibilität des Rezeptors kann die Spezifität und Affinität der Komplexbildung negativ beeinflussen.

Die von Paul Ehrlich geforderte chemische Bindung als Voraussetzung für eine biologische Wirkung hat bis heute Gültigkeit. Der Bereich des Rezeptors an den ein Ligand bindet, wird dabei als Bindetasche oder Bindestelle bezeichnet. Mit wenigen Ausnahmen (etwa Anionentauschern[2]) wirken Arzneistoffe durch die Bindung an makromolekulare Zielstrukturen, von denen die meisten Proteine sind. Die biologische Wirkung entsteht als die Folge der Modulation des Rezeptors, zum Beispiel die Aktivierung oder Inhibierung eines Enzyms. Die Bindung eines Liganden an eine Zielstruktur geschieht dabei spontan, wenn die Energie des Rezeptor-Liganden Komplexes niedriger ist als die der ungebundenen Moleküle. Thermodynamisch wird die Komplexbildung durch die **entropischen** und **enthalpischen** Beiträge charakterisiert, die eine Bindung begünstigen oder verhindern können. Die Änderung der freien Bindungsenthalpie ΔG (auch „Gibbs Energie" genannt), die bei Protein-

[2] Der Anionentauscher Colestyramin wird nicht resorbiert und wirkt ausschließlich im Darm. Der Wirkstoff vermittelt dort eine Bindung an körpereigene Gallensäuren und verhindert deren Wiederaufnahme in den Organismus. Colestyramin wird bei Hypercholersterinämie eingesetzt, da der Körper gezwungen wird, vermehrt Cholesterin als Vorläufer neuer Gallensäuren zu verbrauchen.

Liganden Wechselwirkungen beobachtet werden kann, fasst diese Einflüsse zusammen (Gibbs, 1878; „Gibbs-Helmholtz Gleichung“, Gleichung 1).

$$\Delta G = \Delta H - T\Delta S. \quad (1)$$

Eine negative Gibbs Energie zeigt eine spontan ablaufende Komplexbildung an. Diese wird begünstigt durch steigende entropische Beiträge *S*, die den Freiheitsgrad eines gesamten molekularen Systems beschreiben. Zwar beschränkt die Bindung des Liganden in eine enge Bindetasche dessen konformationelle Freiheit, steigert jedoch die Freiheitsgrade durch die Bindung verdrängten Wassermoleküle. Ein genereller Anstieg der Entropie führt zu einer insgesamt niedrigeren freien Bindungsenthalpie des Systems und geht im Term $-T\Delta S$ negativ in die Gibbs Gleichung ein (in Abhängigkeit von der Temperatur *T*). Die Veränderung der *Enthalpie* ΔH beschreibt physikochemischen Wechselwirkungen zwischen Protein und Ligand, die eine Anziehung oder Abstoßung der Moleküle bewirken können. Negative enthalpische Beiträge begünstigen eine Bindung des Liganden.

Die spezifische Erkennung und die Affinität zwischen Molekülen werden maßgeblich durch nicht-kovalente Wechselwirkungen[3] beeinflusst. So können kleine Änderungen in der Konfiguration eines Liganden deutliche Abweichungen in der Affinität gegenüber einem Rezeptor verursachen oder andersartige Bindemodi des Moleküls hervorrufen. Als wichtigste nicht-kovalente Wechselwirkungen gelten Wasserstoffbrücken, Salzbrücken (auch Ionische Wechselwirkungen), aromatische und van-der-Waals-Interaktionen (Böhm *et al*., 2002).

Während enthalpische Beiträge zur freien Energie nur schwer zu quantifizieren sind, konzentrieren sich verschiedene Anwendungen der Chemie- und Bioinformatik auf die Abschätzung molekularer Wechselwirkungen zwischen Molekülen. Die Verwendung so genannter **künstlicher Gitter** bildet eine wichtige Methode der computergestützten

[3] Nicht-kovalente Wechselwirkungen beschreiben chemische Bindungen, die nicht durch die Wechselwirkung zwischen Außenelektronen zweier Atome gebildet werden, die ein sog. bindendes Elektonenpaar ausbilden.

Analyse molekularer Oberflächeneigenschaften und soll im folgenden Abschnitt vorgestellt werden.

1.2 Konzept künstlicher Gitter und Beschreibung molekularer Oberflächen

Biochemische Untersuchungen über die Strukturen, Eigenschaften oder das Bindeverhalten von Proteinen benötigen geeignete Repräsentationen von Molekülen. Quantenmechanische Beschreibungen interpretieren Atome und Moleküle als eine Summe von Elektronendichten, die grundlegende Eigenschaften wie molekulare Erkennung und Reaktionsverhalten bestimmen (Schneider & Baringhaus, 2008). Eine anschauliche und verbreitete Beschreibung intermolekularer Wechselwirkungen gelingt durch die Betrachtung molekularer Oberflächen (Böhm & Schneider, 2003). Mit dem Aufkommen der Computergraphik wurden Modelle zur Darstellung molekularer Oberflächen entwickelt, die bildhafte Darstellungen komplexer Strukturen und ihrer Eigenschaften ermöglichen.

Die Identifikation der für Liganden zugänglichen Bereiche eines Proteins bildet eine wichtige Anwendung der Darstellung molekularer Oberflächen. Hierbei spielt das von Lee und Richards entwickelte Modell der **lösungsmittelzugänglichen Oberfläche** eine zentrale Rolle (Lee & Richards, 1971). Diese Oberfläche wird durch das Abrollen einer virtuellen kugelförmigen Sonde über die van-der-Waals-Oberfläche (Corey & Pauling, 1953) eines Moleküls extrahiert (Abbildung 1 a)). Lee und Richards verwendeten dazu eine Sonde mit einem Radius von 1,4 Å, der die Dimension eines Wassermoleküls approximiert. Die Oberfläche, die beim Abrollen der Sonde von deren Mittelpunkt gezeichnet wird, zeigt die für das Wasser zugänglichen Bereiche eines Moleküls an (Abbildung 1 b)). Eine andere Darstellung beschreibt die **Conolly-Oberfläche**, die das Volumen in der Umgebung des Proteins anzeigt, dass vom Lösungsmittel nicht erreicht werden kann (Conolly, 1983). Dieses Volumen wird als diejenigen Bereiche charakterisiert, in die die Sondenkugel nicht eindringen kann (Abbildung 1 c)).

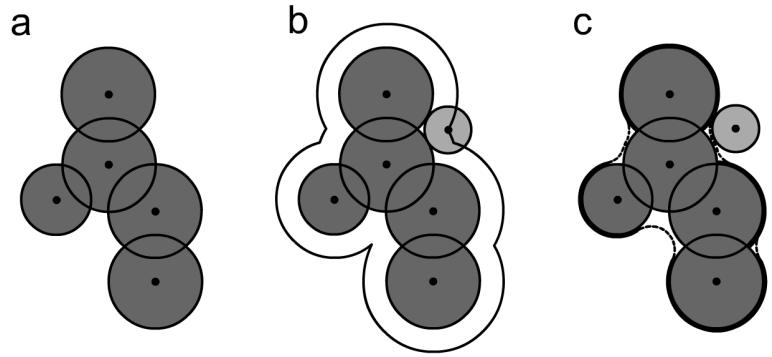


Abbildung 1: Repräsentation molekularer Oberflächen: a) Van-der-Waals-Oberfläche, b) Extraktion der lösungsmittelzugänglichen Oberfläche, c) Ermittlung der Conolly-Oberfläche.

Für detaillierte computergestützte Untersuchungen von Molekülen, ihrer chemischen und strukturellen Eigenschaften wird in der Chemieinformatik das Prinzip **künstlicher Gitter** angewendet (Eisenhaber *et al.*, 1995). Diese ermöglichen anschauliche Beschreibungen der Beziehungen benachbarter Objekte durch eine gleichmäßige Segmentierung des Raums. Ein künstliches Gitter ist ein gedachtes Gebilde von Kanten, die sich im Raum kreuzen und regelmäßige gitterartige Muster ausbilden. Eine in ein künstliches Gitter eingepasste Molekülrepräsentation kann so durch rechnergestützte Verfahren untersucht werden. Die Verwendung von künstlichen Gittern verschiedener Topologien und Auflösung ermöglicht qualitative Beschreibungen molekularer Eigenschaften für beliebige Auflösungen.

Bekannte Beispiele für die Verwendung künstlicher Gitter in der Chemieinformatik sind die Programme GRID (Goodford, 1985), GROW (Moon & Howe, 1991) oder LUDI (Böhm, 1991) für das *De Novo* Design von Wirkstoffen. Die genannten Verfahren ermitteln günstige Wechselwirkungszentren in Proteinbindetaschen für die Platzierung funktioneller Gruppen potentieller Liganden. Dazu analysieren virtuelle molekulare Sonden die Moleküloberfläche in künstlichen Gittern.

In dieser Arbeit wurden künstliche Gitter für die Identifikation und die Analyse von Proteinbindetaschen und ihrer Eigenschaften verwendet. Im folgenden Abschnitt sollen Methoden zur Bestimmung potentieller Bindestellen vorgestellt werden.

1.3 Vorhersage von möglichen Bindestellen für niedermolekulare Liganden auf Proteinstrukturen

Die computergestützte Identifikation potentieller Bindestellen auf der Oberfläche von Proteinen sowie in verdeckten Bereichen des Proteininnern bildet oftmals die Grundlage für den rezeptorbasierten rationellen Wirkstoffentwurf (Colman, 1994; Klebe 2000). Die Charakterisierung von Form und Volumen einer Ligandenbindestelle sowie die Identifikation der flankierenden Reste repräsentiert eine Schlüsselrolle für verschiedene Anwendungen wie das automatisierte Liganden-Docking und *in situ* Modelling (Hendlich *et al*., 1997). Ferner erlaubt der Vergleich des geometrischen Aufbaus von Bindetaschen verwandter Proteine wertvolle Hinweise für das Verständnis der Bindemechanismen der gebundenen Liganden und ihrer Wechselwirkungen mit ihren Rezeptoren.

Kenntnisse über Struktur und Funktion validierter Zielproteine (engl. *target proteins*) bieten wertvolle Informationen für den Entwurf neuer therapeutischer Wirkstoffe. Der starke Anstieg von hochaufgelösten Kristallstrukturen (> 56.000 Einträge, März 2009), die in der Protein Datenbank PDB (Bermann *et al.*, 2000) verfügbar sind, eröffnet hierbei neue Möglichkeiten für den rezeptorbasierten Wirkstoffentwurf (Gane & Dean, 2000; Klebe 2000). Die Identifikation potentieller Proteinbindestellen stellt dabei nach wie vor eine zentrale Aufgabe dar, da die Fähigkeit mit anderen Proteinen oder niedermolekularen Wirkstoffen zu interagieren die biologische Funktion eines Proteins bestimmt. Dies unterstreicht die Bedeutung der Untersuchung von Bindetaschen für den rationellen Wirkstoffentwurf, wie das Liganden-Docking oder das *de novo* Moleküldesign.

Verschiedene computergestützte Ansätze existieren für die Vorhersage potentieller Bindetaschen auf Proteinkristallstrukturen. Die Mehrzahl dieser Methoden stützt sich bei der Berechnung ausschließlich auf geometrische Betrachtungen, um Einstülpungen der Proteinoberfläche zu finden. Empirische Studien zeigten bereits, dass die größte Vertiefung dabei meist die tatsächliche Bindetasche eines betreffenden Proteins ausmacht (Sotriffer & Klebe, 2002; Campbell *et al.*, 2003). Das Programm SURFNET (Laskowski, 1995) berechnete die tatsächliche Ligandenbindestelle als die größte gefundene Tasche für 83% der Fälle für einen Datensatz von 67 monomeren Enzymen

(Laskowski *et al.*, 1996). SURFNET identifiziert Hohlräume zwischen zwei oder mehreren Proteinen, sowie interne Taschen durch das Einpassen von virtuellen Kugeln (engl. *spheres*) in die dem Lösungsmittel zugängliche Lücken (engl. *gaps*) zwischen Proteinatomen (Abbildung 2).

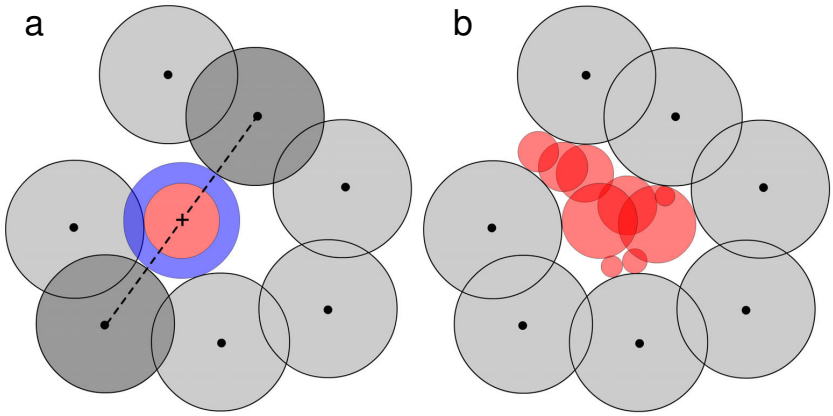


Abbildung 2: Zwei-dimensionale Darstellung der Taschenvorhersage mit SURFNET. a) Eine sog. *initial gap sphere* (Startkugel, blau) wird zwischen die van-der-Waals-Radien zweier Atome (dunkelgrau) eingepasst. Der Radius dieser Startkugel wird anschließend soweit reduziert, dass Überschneidungen mit anderen Atomradien vermieden werden. Die resultierende *gap sphere* ist in rot dargestellt. b) Weitere *gap spheres* werden ausgehend von den umgebenden Atomen platziert und angepasst. Die Anordnung der finalen *gap spheres* wird in SURFNET zur Beschreibung der Größe und Form von Proteinbindetaschen verwendet.

Das Programm CAST (Liang *et al.*, 1998; Binkowski *et al.*, 2003) nutzt den sog. *alpha shapes* Algorithmus (Edelsbrunner & Mücke, 1994; Edelsbrunner *et al.*, 1995) zur Vorhersage potentieller Bindetaschen. Grundlage dieses Verfahrens ist die räumliche Aufteilung von Vertiefungen der Oberfläche durch Voronoi Zerlegung (Aurenhammer, 1991) und Delaunay Triangulation (Lee & Schachter, 1980). Die Triangulation ermöglicht eine Segmentierung der Proteinoberfläche auf atomarer Ebene. Die sog. „Discrete Flow-Methode“ vereinigt benachbarte Segmente in Vertiefungen der Oberfläche und kennzeichnet Orte potentieller Bindetaschen (Abbildung 3).

CAST wurde auf 51 von 67 monomeren Proteinstrukturen getestet, die aus dem Datensatz zur Evaluation von SURFNET (Laskowski, 1995) abgeleitet wurden. Die Vorhersageroutine von CAST erreichte eine korrekte Vorhersage der tatsächlichen Bindetasche von 74% auf diesem Datensatz.

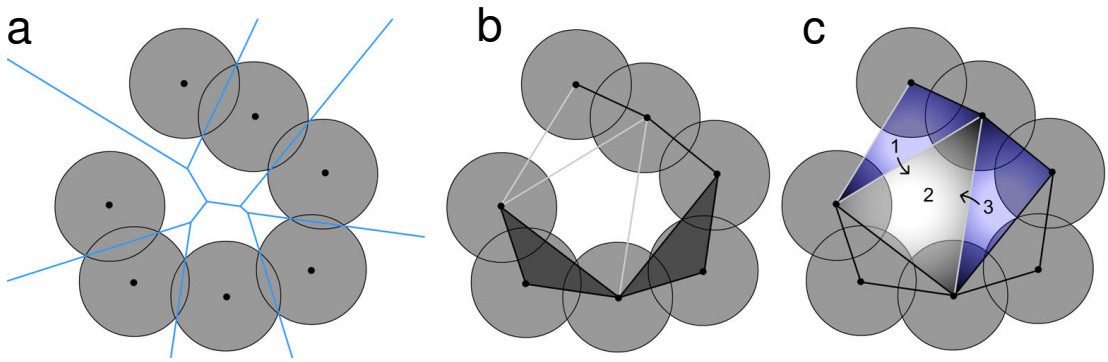


Abbildung 3: Zwei-dimensionale Beschreibung der Taschenvorhersage mit CAST. a) Darstellung einer Voronoi-Zerlegung der Atome (grau) einer Bindetasche. Die durch blaue Linien begrenzten Voronoi-Regionen zerteilen den Raum und ordnen diskrete Bereiche dem nächstgelegenen Atom zu. b) Die durch die Atomzentren verlaufenden Linien beschreiben die komplexe Hülle, die anhand des Voronoi-Diagramms in Delaunay-Dreiecke zerteilt wird. Die schattierten Dreiecke und schwarzen Linien beschreiben die „Alpha-Shape". Sie umschließt drei zentrale Dreiecke, die mindestens eine graue Begrenzung besitzen. c) Zwei spitzwinklige Dreiecke (1,3) werden dem stumpfwinkligen Dreieck (2) in der „Discrete-Flow-Methode" zugewiesen.

Ein weiterer Ansatz zur Bestimmung möglicher Bindetaschen ist im PASS (*Putative Active Sites with Spheres*) Algorithmus implementiert (Brady & Stouten, 2000). In einem iterativen Prozess werden Lagen von virtuellen Kugeln (engl. *spheres*) in vergrabenen Teilen der Proteinoberfläche verteilt. Der Algorithmus bestimmt zentrale Sphären für jede Vertiefung, die als „active site points" potentielle Orte für die Bindung von Liganden charakterisieren (Abbildung 4). Die Vergrabenheit wird als die Zahl der Proteinatome in einem Umkreis von 8 Å um eine Kugel bestimmt.

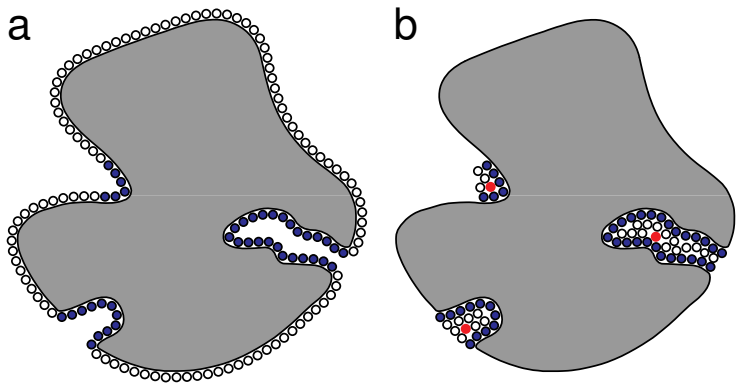


Abbildung 4: Vorgehensweise des PASS Algorithmus zur Bestimmung potentieller Bindetaschen a) Virtuelle Kugeln (engl. *spheres*) werden über die Proteinoberfläche verteilt und vergrabenen Bereiche identifiziert (blaue spheres). Die übrigen Kugeln (weiß) werden verworfen. b) Weitere Lagen von Kugeln werden hinzugefügt und zentrale „active site points" (rot) bestimmt.

Eine weitere geometrische Methode für die Vorhersage möglicher Bindetaschen wird durch das Programm POCKET (Levitt & Banaszak, 1992) realisiert. Der verwendete Algorithmus konstruiert ein regelmäßiges künstliches Gitter um die betrachtete Proteinstruktur und kennzeichnet die Gitterknoten dem Lösungsmittel zugängliche oder unzugängliche Bereiche. POCKET sucht entlang der *x*-, *y*- und *z*-Achse nach Gruppen von dem Lösungsmittel zugänglichen Gitterknoten, die von unzugänglichen Gitterknoten umschlossen werden. Solche Anordnungen werden als PSP-Ereignisse (engl. *Protein-Solvent-Protein events*) bezeichnet. Die Auswahl von nur drei Raumrichtungen stellte sich als nachteilig für den Erfolg der Vorhersage heraus, da Taschen mit einer Orientierung von 45° gegenüber dem Achsensystem nicht oder nur unvollständig erkannt werden konnten (Hendlich *et al.*, 1997). Das Programm LIGISTE wurde daher als Weiterentwicklung von POCKET vorgestellt, um diesen Nachteil auszugleichen (Hendlich *et al.*, 1997). So verfolgt LIGSITE eine zusätzliche Untersuchung des künstlichen Gitters entlang der vier kubischen Diagonalen. Dies erlaubt eine Taschensuche, die weniger abhängig von der Orientierung des Proteins im Gitter ist (Abbildung 5).

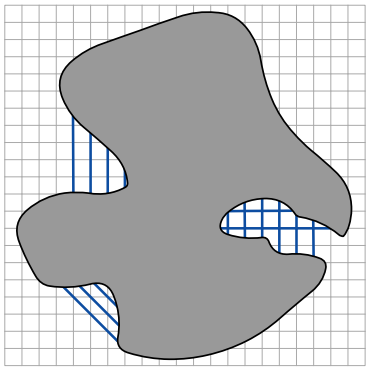


Abbildung 5: Prinzip der Taschenvorhersage von POCKET und LIGSITE. Das Protein wird in ein kubisches Gitter eingepasst, ein Suchalgorithmus sucht entlang der Achsen nach vom Protein umgebenen Bereichen, die für das Lösungsmittel zugänglich sind. Dabei betrachtet LIGSITE auch Orientierungen entlang der Gitterdiagonalen und kann so Taschen identifizieren, die von POCKET nicht gefunden werden (Tasche unten links).

LIGISITE^{cs} (*cs* = Conolly Surface) und LIGSITE^{csc} (*cs* = Conolly Surface and Conservation) repräsentieren ebenfalls Weiterentwicklungen des ursprünglichen POCKET Algorithmus, verfolgen jedoch die Suche nach SSS-Ereignissen (engl. *Surface-Solvent-Surface events*) anstelle der Suche nach PSP-Betrachtungen (Huang & Schröder, 2006).

Weitere Vorhersagemethoden, die sich ausschließlich auf geometrische Untersuchung der Proteinstruktur stützen sind Cavity Search (Ho & Marshall, 1990), VOIDOO (Kleywegt & Jones, 1994), APROPOS (Peters *et al.*, 1996) und Travel Depth (Coleman & Sharp, 2006).
Die Programme DrugSite (An *et al.*, 2004) und Pocket-Finder (Ruppert *et al.*, 1997) betrachten sowohl strukturelle als auch physikochemische Eigenschaften für die Identifikation potentieller Bindetaschen. Die Software QSiteFinder (Laurie & Jackson, 2005) verfolgt einen energiebasierter Ansatz zur Vorhersage. Dabei werden Wechselwirkungen zwischen dem Protein und einer van-der-Waals-Sonde untersucht, um energetisch günstige Bereiche der Oberfläche zu identifizieren, die die Bindung eines Liganden unterstützen.

Die vorgestellten Methoden bieten zudem Ansätze für genauere Charakterisierungen potentieller Bindetaschen. So ermöglichen Beschreibungen von vorhergesagten Bindevolumen durch Alpha-Shapes oder räumliche Anordnungen von Gittersonden maschinenlesbare Repräsentationen, die für weitere Analysen genutzt werden können. Von besonderem Interesse sind hier Ähnlichkeitssuchen von Bindetaschenformen und ihren Eigenschaften zur Abschätzung von unerwünschten Nebenwirkungen mit Bindestellen auf anderen Proteinstrukturen (Schmitt *et al.*, 2002).
Im Rahmen dieser Arbeit wurde eine Sammlung von Programmen erstellt, die die Vorhersage, sowie den Vergleich von Bindetaschen ermöglichen. Hierzu wurden Algorithmen entwickelt, die Ähnlichkeitssuchen anhand von zuvor berechneten Vorhersagen möglicher Taschen zulassen. Grundlage der entwickelten Methoden für den alignmentfreien Vergleich und die Analyse der Druggability von Bindetaschen bilden Autokorrelationsdeskriptoren, die nachfolgend vorgestellt werden.

1.4 Autokorrelationsdeskriptoren

Die Repräsentation molekularer Eigenschaften durch Korrelationsdeskriptoren findet vielfache Anwendung in der Bio- und Chemieinformatik (Pastor *et al.*, 2000; Durant *et al.*, 2002; Renner *et al.*, 2006; Tanrikulu *et al.*, 2007; Weisel *et al.*, 2007). Die

Kodierung struktureller Merkmale und pharmakophorer[4] Eigenschaften wirkstoffartiger Liganden als molekulare Deskriptoren ermöglicht das effiziente computergestützte Durchmustern von Substanzbibliotheken (engl. *Virtual Screening*). So wird die Abwesenheit oder das Auftreten bestimmter chemischer Fragmente in binären MACCS Keys Deskriptoren (Durant *et al.*, 2002) durch ‚0' oder ‚1' kodiert. Dieser binäre zweidimensionale Deskriptor ignoriert die topologische und räumliche Distanz der betrachteten Fragmente zueinander.

Topologische Autokorrelationsdeskriptoren wurden 1980 von Moreau und Broto entwickelt (Moreau & Broto, 1980; Broto *et al.*, 1984): Der **Autocorrelation of a Topological Structure** (ATS) Deskriptor repräsentiert Atome durch ihre molekularen Eigenschaften, etwa der atomaren Masse oder Partialladung. Der ATS Deskriptor interpretiert ferner die Länge des kürzesten Pfades, der zwei betrachteten Atome im Molekülgraph miteinander verbindet als deren topologische Distanz. Für eine gewählte topologische Distanz d wird der ATS Deskriptor wie in Gleichung 2 dargestellt berechnet.

$$ATS_d = \sum_{j=i+1}^{A} \sum_{i=1}^{A-1} \delta_{ij,d} \cdot \left(w_i w_j\right). \tag{2}$$

Hierbei beschreibt w die atomare Eigenschaft für die Anzahl A der Atome eines betrachteten Moleküls. Das Kronecker Delta δ ergibt 1 für alle Atompaare, die im Abstand d existieren. Zur Darstellung des vollständigen Deskriptors wird die ATS Autokorrelation über alle definierten Distanzen berechnet und zu einem Vektor $\{ATS_1, ATS_2, \ldots, ATS_D\}$, wobei D eine maximale betrachtete Distanz als Schwellenwert betrachtet.

[4] Ein Pharmakophor beschreibt die räumliche Anordnung der funktionellen Gruppen eines Moleküls, die dessen pharmakologische Wirkung bestimmen. „Ein Pharmakophor ist das Ensemble von sterischen und elektronischen Eigenschaften, das notwendig ist, um die optimalen supramolekularen Wechselwirkungen mit einer spezifischen biologischen Zielstruktur zu garantieren und dessen biologische Antwort auszulösen (oder zu blockieren)" – IUPAC Definition. Der Begriff des Pharmakophors wurde erstmals von Paul Ehrlich verwendet (Ehrlich, 1909).

Verschiedene Methoden nutzen den von Moreau und Broto vorgestellten Ansatz für den Vergleich von Molekülen und ihrer Eigenschaften über binäre Zeichenketten (engl. *Bitstrings*) und Beschreiber von höherer Komplexität. Dabei betrachten dreidimensionale Deskriptoren die Euklidische Distanz anstelle der topologischen Entfernung, wie etwa die Programme CATS3D (Renner *et al.*, 2006) oder LIQUID (Tanrikulu *et al*., 2007).

Das Konzept der Abstraktion kleiner Moleküle und ihrer Repräsentation durch Korrelationsdeskriptoren ist erfolgreich für die Auffindung biologisch aktiver Strukturen in Substanzbibliotheken eingesetzt worden (Schneider *et al.*, 1999; Fechner *et al*., 2003). In dieser Arbeit wurden Autokorrelationsdeskriptoren zur Beschreibung der Form und Vergrabenheiten von Proteinbindetaschen verwendet. Diese Repräsentation ermöglicht den effizienten alignmentfreien Vergleich von ligandenbindenden Bindestellen auch in großen Datenbanken. Ferner wurden diese Autokorrelationsdeskriptoren zur Untersuchung der Fähigkeit genutzt mit der potentielle Bindetaschen Liganden binden können. Diese Anwendung soll im folgenden Abschnitt beschrieben werden.

1.5 Beschreibung der Druggability von Bindetaschen

Die Identifikation von potentiellen Bindetaschen auf Proteinstrukturen bildet oftmals den ersten Schritt im rezeptorbasierten Moleküldesign. Größe und generelle physikochemische Eigenschaften einer Bindetasche dienen hierbei als grobe Filter in einer frühen Phase des strukturbasierten virtuellen Screenings. Verschiedene Vorhersagemethoden charakterisieren die Größe und Form von Proteinbindetaschen und kennzeichnen auf diese Weise diejenigen Moleküle einer Screeningbibliothek, die zu groß sind, um in die betrachtete Tasche zu binden. Energetische Betrachtungen ermöglichen darüber hinaus die Erkennung von Substanzen, deren Eigenschaften die Bindung an ein Zielprotein (engl. *Target*) verhindern. Die Fähigkeit wirkstoffartige Liganden zu binden wird als die **„Druggability“** eines Proteins beschrieben (Kubinyi 2002; Hopkins & Groom, 2002). Neben der Vorhersage möglicher Bindestellen ist die

Identifikation von **„druggable pockets“** eine zentrale Aufgabe des rezeptorbasierten Moleküldesigns (Hajduk *et al.*, 2005a; Hajduk *et al*., 2005b).

Therapeutisch relevante Zielproteine müssen druggable und krankheitsmodifizierend sein. So muss ein Wirkstoffkandidat nicht nur in der Lage sein, an ein Zielprotein zu binden, sondern muss darüber hinaus eine Veränderung des physiologischen Verhaltens des Proteins erzeugen, die sich auf das jeweilige Krankheitsbild auswirkt. Auf molekularer Ebene werden dabei Effekte wie die Steigerung oder Verringerung der enzymatischen Aktivität eines Enzyms oder die Beeinflussung von Wechselwirkungen mit anderen Proteinen verfolgt. Inwiefern diese Effekte das Krankheitsbild beeinflussen und welche unerwünschten Nebenwirkungen auftreten können, muss anschließend durch *in vivo* Untersuchungen und in klinischen Testverfahren bewertet werden.

Auf Grundlage der Sequenzierungsarbeiten im Rahmen des Humangenomprojekts wird die Zahl der proteinkodierenden Gene auf 20.000 bis 25.000 Einträge geschätzt (Collins, 2004). Dies wirft die Frage auf wie groß das „**Druggable Genome“** ist. Abschätzungen über die Zahl der therapeutisch relevanten Zielproteine stützen sich weitgehend auf die Analysen der Sequenzierung des Humangenoms. Dieser Ansatz ist jedoch zum Beispiel bereits dadurch limitiert, als dass eine Aussage über die Zahl möglicher Splicevarianten oder die Interaktionen der kodierten Proteine nicht von der Sequenzebene aus beantwortet werden kann. Auf Grundlage der ersten Version des sequenzierten Humangenoms wurde die Zahl der relevanten Zielproteine, die von niedermolekularen Verbindungen moduliert werden können auf etwa 5000 geschätzt (Imming *et al.*, 2006). Durch aufwändige Analysen der Ligandenbindung in bekannten strukturaufgeklärten Komplexen und der Identifikation ähnlicher Zielproteine im Humangenom wurde diese Zahl auf etwa 3000 Targets korrigiert (Hopkins & Groom, 2002).

Die Zahl an therapeutisch relevanten Zielproteinen, die heute von der pharmazeutischen Industrie genutzt wird, wird heute auf nur einige hundert geschätzt. Drews und Ryser schätzten die Zahl der Targets, die 1997 von allen auf dem Markt befindlichen Wirkstoffen adressiert werden auf nur 482 (Drews & Reyser, 1997). Die Diskrepanz

zwischen den tatsächlich genutzten Targets und der Zahl der vermuteten Zielproteine lässt Raum für viele neue druggable Targets. Es ist daher von besonderem Interesse, solche Proteine zu identifizieren, deren Taschen als druggable gelten. Gleichzeitig müssen Proteine, die nicht durch wirkstoffartige Liganden modifiziert werden können, bereits in frühen Phasen eines Designprozesses erkannt werden.

Verschiedene Ansätze wurden verwendet, um die Druggability von potentiellen Bindetaschen vorherzusagen. Gängige Methoden betrachten Taschengröße, Oberflächenrauheit oder polare und apolare Oberflächenanteile als Beschreiber für Druggability-Analysen. Frühere Arbeiten zeigen, dass die tatsächliche Bindestelle meist die räumlich größte Tasche mit größter Hydrophobizität und geometrischer Komplexität ist (Hajduk *et al.*, 2005a; Hajduk *et al.*, 2005b). Da keiner dieser Parameter für sich allein ausreichend ist, die Druggability einer Bindetasche korrekt zu beschreiben, werden Regressionsanalysen genutzt, um den Einfluss der einzelnen Variablen festzustellen (Hajduk *et al.*, 2005a; Hajduk *et al.*, 2005b).

Im Rahmen dieser Arbeit wurde eine Methode entwickelt, die die Druggability von Proteinbindetaschen berechnet. Zu diesem Zweck wurden Untersuchungen anhand von Korrelationsdeskriptoren durchgeführt, die aus zuvor extrahierten Bindevolumen errechnet wurden. Unsere Ergebnisse zu Klassifikationen der Druggability von Bindetaschen mit selbstorganisierenden Karten (SOMs; Kohonen, 1982) sollen in dieser Arbeit vorgestellt und diskutiert werden.

1.6 Ziel dieser Arbeit

Die rechnergestützte Analyse von Proteinbindetaschen setzt eine geeignete maschinenlesbare Repräsentation der Orte auf der Proteinoberfläche voraus, die als Bindestellen für wirkstoffartige Liganden dienen können. Das Programm *PocketPicker* (Weisel, 2006) ermöglicht die Vorhersage solcher Bindetaschen und charakterisiert deren Form und Vergrabenheit. Das Ziel dieser Arbeit ist die Weiterentwicklung dieser Software zur detaillierten Analyse der physikochemischen Eigenschaften, Druggability, sowie der topologischen Vielfalt von Proteinbindetaschen.

Die Untersuchung großer Mengen von Kristallstrukturen erfordert eine effiziente Berechnung der zu Grunde liegenden Daten. Eine Reimplementierung von *PocketPicker* in der Programmiersprache Java™ verspricht eine Beschleunigung der ursprünglich in Python entwickelten Software und bildet ein weiteres Ziel dieser Arbeit.

Die Abschätzung der Eignung möglicher Bindetaschen für den rechnergestützten Entwurf von Wirkstoffen wird als ‚Druggability' bezeichnet. Zur Untersuchung dieser Eigenschaft eignen sich die in *PocketPicker* implementierten Autokorrelations-deskriptoren. Die Analyse und Klassifikation von ligandenbindenden Taschen anhand von Selbstorganisierenden Karten (SOMs) ist Gegenstand dieser Arbeit.

Die Analyse der topologischen Vielfalt von Proteinbindestellen erfordert eine automatische Extraktion des Aufbaus und der Verzweigungen der Bindetaschen. Dies soll durch die Anwendung des Prinzips Wachsender Neuronaler Gase (GNG) gewährleistet werden.

Die Abschätzung der funktionellen Ähnlichkeit potentieller Bindestellen erfordert neben der Betrachtung von Form und Größe eine detaillierte Analyse der physikochemischen Eigenschaften einer Tasche. Die Auswertung elektrostatischer sowie lipophiler Potentiale in Proteinbindetaschen bildet in dieser Arbeit die Grundlage zur Unter-suchung funktioneller Ähnlichkeiten. Die korrekte strukturelle Überlagerung der zuvor berechneten Bindetaschen ist eine notwendige Voraussetzung für den Ähnlichkeitsvergleich und soll durch die Anpassung der Software *Shapelets* (Proschak *et al.*, 2007) erreicht werden.

Die in dieser Arbeit neu und weiter entwickelten Programme sollen im Programmpaket *PocketomePicker* zusammengefasst werden, um eine einfache Bedienung der Software für andere Nutzer zu ermöglichen. Das Paket beinhaltet eine Kommandozeilenversion für Berechnungen auf Servermaschinen sowie eine Ausführung mit Grafischer Benutzeroberfläche (GUI) und Molekülbetrachter für die Verwendung unter Microsoft Windows und Linux.

Die Leistungsfähigkeit der entwickelten Methoden wird durch Anwendung auf konkrete Projekte in Kooperationsarbeit getestet. Im Rahmen dieser Arbeit werden dazu Funktionsanalysen der vorhergesagten Taschen aus den Homologiemodellen der Proteine APOBEC3C und Glutamat Dehydrogenase druchgeführt. Ferner soll die Methode *ReverseLiQUID* für die Erstellung eines rezeptorbasierten Pharmakophormodells zum Virtuellen Screening nach Inhibitoren der Protease HtrA aus Helicobacter pylori angewendet werden.

2 Material und Methoden

2.1 Verwendete Programme und Methoden

2.1.1 Java

Die in dieser Arbeit neu entwickelten und reimplementierten Methoden wurden in der Programmiersprache Java™ (http://java.sun.com/), Version 5 realisiert. Außerdem wurden die frei verfügbaren Bibliotheken JGraphT (http://jgrapht.sourceforge.net/) in der Version 0.7, vecmath (https://vecmath.dev.java.net/) in der Version 1.5.0 und Chemistry Development Kit (Steinbeck *et al.*, 2003; Steinbeck *et al.*, 2006) in der Version 20060714 verwendet (http://almost.cubic.uni-koeln.de/cdk/).
Zur Programmierung einer grafischen Benutzeroberfläche wurden die externe OpenGL Bibliothek Jogl (https://jogl.dev.java.net/) in den Versionen 0.6.3 (Windows) und 0.7.1 (Linux), sowie die Grafikbibliotheken Swing (https://swing-layout.dev.java.net/) und JGoodies Looks (https://looks.dev.java.net/) in der Version 2.1.4 verwendet.

2.1.2 PyMOL

PyMOL (DeLano, 2002; http://pymol.sourceforge.net/) wurde als Open-Source Software zur Visualisierung von Molekülen im dreidimensionalen Raum entwickelt. In dieser Arbeit wurde die lizensierte Version PyMOL 1.1r1 zur Visualisierung von Ausgabedateien verwendet. PyMOL verfügt über einen integrierten Interpreter der Programmiersprache *Python* (http://www.python.org/). Dies erlaubt eine Erweiterung der in PyMOL vorhandenen Funktionen anhand eigens implementierter Python-Skripte.

2.1.3 JyMOL

JyMOL ist eine Java-basierte Variante von PyMOL und wurde speziell für die Integration in Java-Projekte entwickelt (http://delsci.com/jymol/). JyMOL ist als freie Software verfügbar und wurde in dieser Arbeit in den Versionen 0.63 (Windows) und 0.71 (Linux) zur Visualisierung von Moleküldarstellungen in der grafischen Benutzeroberfläche von *PocketomePicker* verwendet.

2.1.4 Reduce

Reduce (Word *et al.*, 1999; http://kinemage.biochem.duke.edu/software/reduce.php) ist ein Programm zum Anfügen von Wasserstoffatomen an Nukleinsäuren und Proteinstrukturen im Format der Protein Data Bank (PDB, Berman *et al.*, 2000). Wasserstoffatome werden dabei in standardisierten Geometrien angefügt, wobei die Orientierungen von OH, SH, NH_3^+, Methionin Methylgruppen, Histidinringen, sowie Amid-Seitenketten von Asparagin und Glutamin Seitenketten optimiert werden. Reduce kann darüber hinaus Wasserstoffatome an Liganden anfügen, die im PDB Eintrag mit *HET*-Bezeichnern (Heteroatoms) gekennzeichnet sind. Hierfür wird eine Datei bereitgestellt, die die Konnektivitäten der in der PDB abgelegten Liganden beinhaltet. Dies ist notwendig, da die Bindungswertigkeiten von Liganden in der PDB nicht angegeben sind.

In dieser Arbeit wurde Reduce in der Version 3.13 verwendet. Diese Version unterstützt die aktuell in der PDB verwendete ‚HET'-Nomenklatur für Liganden (Stand Oktober 2008).

2.1.5 PDB2PQR

PDB2PQR (Dolinsky *et al.*, 2004; http://pdb2pqr.sourceforge.net/) ist eine in Python implementierte Open-Source Anwendung zur Berechnung von Poisson-Boltzmann Elektrostatikbeiträgen. Dieses wird zur Berechnung des elektrostatischen Potentials in organischen Lösungen verwendet (Baker *et al.*, 2000). In dieser Arbeit wurde PDB2PQR in der Version 1.3.0 zur Berechnung von Partialladungen in Proteinstrukturen verwendet.

2.1.6 Cytoscape

Cytoscape (http://www.cytoscape.org; Shannon *et al.*, 2003) ist eine Open-Source Anwendung zur Abbildung von Interaktionen zwischen Datenpunkten in großen Datensätzen. Beziehungen zwischen einzelnen Datenbankeinträgen werden dabei als Kanten in Netzwerken dargestellt. Die Software wurde in dieser Arbeit für die

Visualisierung struktureller Verwandtschaften ähnlicher Bindetaschengeometrien verwendet.

2.1.7 MOLMAP

MOLMAP® (Schneider *et al.* 1998) wurde in dieser Arbeit verwendet, um die Druggability Proteinbindetaschen zu bewerten. MOLMAP® verwendet hierfür eine Variante des von Kohonen vorgestellten Algorithmus der **selbstorganisierenden Karten** (SOM; Kohonen, 1982). SOMs sind unüberwacht lernende **künstliche neuronale Netze** (engl. *Artificial Neural Networks*, ANNs), die eine Projektion hochdimensionaler Daten auf niedrigdimensionale Karten erlauben. Anders als bei überwachten neuronalen Netzwerken berechnet eine SOM keine Ausgabewerte für die Eingangsdaten. Die Neuronen der SOM lassen sich als Vektoren in einem Raum beschreiben, der die Dimensionalität der Eingabedaten besitzt. Die Gewichte der Neuronen geben die Position Neuronen im Raum an. Der in MOLMAP® verwendete Algorithmus zum Trainieren einer SOM ist nachfolgend dargestellt:

Initialisierung:

- Initialisieren der SOM mit S Neuronen c_i: $\mathscr{C} = \{c_1, c_2, ..., c_S\}$ mit zufällig in [-1,1] gewählten Gewichten w.
- Einrichtung der Nachbarschaftsbeziehungen zwischen den Neuronen aus $\mathscr{C}$ als ein rechteckiges Gitter ($S = X * Y$).
- Einstellung des Zeitparameters $t = 0$.

Rekursion:

- Wähle einen Eingabesignal x zufällig.
- Ermittle das Gewinnerneuron s mit kleinstem Abstand zu x.
- Aktualisiere die Gewichte w aller Neuronen c:

  $\Delta w_c = \varepsilon(t) h_{cs}(x - w_c)$,

  wobei die Euklidische Distanz d_2 zur Bestimmung der Abstände zwischen den Neuronen auf dem SOM-Gitter benutzt wurde. Die Nachbarschaft um das s wird durch eine Gauß-Funktion beschrieben:

$h_s = \exp\left(\frac{-d_2(c,s)^2}{2\sigma^2}\right)$, mit Standardabweichung:

$\sigma(t) = \sigma_{Start}\left(\frac{\sigma_{Start}}{\sigma_{Ende}}\right)^{\frac{t}{t_{\max}}}$ und Lernrate

$\varepsilon(t) = \varepsilon_{Start}\left(\frac{\varepsilon_{Start}}{\varepsilon_{Ende}}\right)^{\frac{t}{t_{\max}}}$.

- Erhöhe den Zeitparameter: $t = t + 1$

Termination:

- Beende Rekursion wenn $t = t_{\max}$

Vor Beginn des Trainings müssen der ‚Start' Wert und die Zahl der Zyklen in MOLMAP® definiert werden. Der ‚Ende' Wert berechnet sich als Zyklen / Anzahl Eingabedaten.

SOMs bieten eine Topologieerhaltende Projektion der Daten auf der niedrigdimensionalen Kohonen Karte, so dass die Nachbarschaftsbeziehung korrekt wiedergegeben werden. In dieser Arbeit wurden Projektionen auf rechteckige zwei-dimensionale toroidale Karten durchgeführt.

2.1.8 Shapelets

Für die strukturelle Überlagerung von Proteinbindetaschen wurde in dieser Arbeit der ***Shapelets***-Algorithmus (Proschak *et al*., 2008) verwendet und weiterentwickelt. Die Methode wurde ursprünglich zum formbasierten (engl. *shape-based*) **virtuellen Screening** von wirkstoffähnlichen Molekülen entworfen (Proschak *et al*., 2007). Der *Shapelets*-Algorithmus stand in dieser Arbeit als Quellcode zur Verfügung, was eine Anpassung der Methode zur Beschreibung von Bindetaschenoberflächen ermöglichte. Das Vorgehen der *Shapelets*-Methode soll nachfolgend dargestellt werden. Die in dieser Arbeit entwickelte Technik ***PocketShapelets*** wird im Abschnitt „Entwickelte Methoden" beschrieben (Kapitel 2.3).

2.1.8.1 Oberflächenbeschreibung mit Shapelets

Die *Shapelets*-Methode beschreibt ein Molekül durch eine Reihe von Flächenfunktionen, die charakteristische Stellen der Molekülоberfläche annähern. Der Algorithmus teilt sich in folgende Abschnitte:

- Beschreibung eines Moleküls als Gauß'sche Dichtefunktion
- Extraktion der Isofläche
- Zerlegung der Oberfläche

Das Verfahren verfolgt dabei ein etabliertes Konzept, das die dreidimensionale Form eines Moleküls *M* durch eine Summe von Gauß-Funktionen beschreibt (Blinn, 1982; Grant & Pickup, 1995). In diesem Ansatz werden die *N* vielen Atomkoordinaten $\vec{c}_n$ eines Moleküls *M* unter Berücksichtigung ihrer van-der-Waals Radien r_n in ein künstliches Gitter eingepasst (Gleichung 3). Ein Molekül wird für beliebige Raumkoordinaten $\vec{x}$ als Funktionswert einer Gauß'schen Dichtefunktion dargestellt.

$$M(\vec{x}) = \sum_{n=1}^{N} e^{\frac{-2(\vec{x}-\vec{c}_n)^2}{r_n^2}} . \qquad (3)$$

Für die Extraktion der Isofläche wird der ***Marching Cubes*** Algorithmus verwendet, der zur Berechnung triangulierter Oberflächen für drei-dimensionale Eingabedaten entwickelt wurde (Lorensen & Cline, 1987). Als Eingabe dienen dem Algorithmus die Funktionswerte von $M(\vec{x})$, die an jedem Knoten des zu Grunde liegenden kubischen Gitters berechnet wurden, sowie ein Isoflächenparameter *c* (engl. *cutoff*). Im Zuge der Berechnung wird für jede Zelle des Gitters ein achtstelliger binärer Index berechnet. Dabei trägt jeder Gitterpunkt einer Zelle eine 1 zum Index bei, wenn sein Funktionswert den cutoff *c* übersteigt. Den übrigen Gitterpunkten wird eine 0 zugewiesen, so dass der resultierende Index eine charakteristische Beschreibung des von der Zelle eingeschlossenen Volumens ermöglicht. Die Indizes der Zellen dienen darüber hinaus als Referenz für die Triangulation der Oberfläche (Abbildung 6).

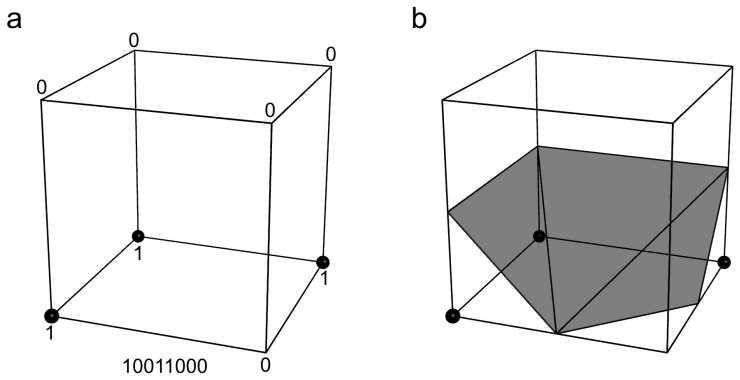


Abbildung 6: Extraktion der Isofläche für eine Zelle des künstlichen Gitters mit *Marching Cubes* Verfahren. a) Ecken der Zelle an denen der gemessene Funktionswert einen Schwellenwert überschreitet, erhalten einen Index von eins, sonst null. Die acht Ecken einer Zelle beschreiben einen eindeutigen Index. b) Die Indizierung dient der Extraktion einer triangulierten Oberfläche.

Für die Extraktion der Isofläche nutzt *Shapelets* ein kubisches Gitter mit einer Maschenweite von 0,5 Å. Trotz dieser hohen Auflösung wird die berechnete Oberfläche nur durch diskrete Punkte definiert, die ungleichmäßig verteilt sein können (Schroeder *et al.*, 1992). Die *Shapelets* Methode nutzt daher die einfache Glättungsmethode ***Welding Vertices*** (Dunn & Parberry, 2002), das rekursiv eine neue Koordinate für eng benachbarte Oberflächenpunkte interpoliert.

Zur Zerlegung der extrahierten Oberfläche nutzt *Shapelets* ein von Zachmann vorgestelltes Verfahren zur Approximation der lokalen Oberflächenkurvatur (Zachmann *et al.*, 1992). Ziel der Methode ist das Anpassen von Sattelfunktionen S (sog. **hyperbolische Paraboloide**) mit zuvor definierten Radien r_S an die triangulierten Punkte der zuvor berechneten Molekülоberfläche. Charakteristische Bereiche der Oberfläche werden so durch hyperbole Paraboloide beschrieben, die als *Shapelets* bezeichnet werden und namensgebend für die gesamte Methode sind.

Das Annähern der Paraboloide gelingt durch Minimieren einer Fehlerfunktion, die die Güte der Anpassung wiedergibt. In einem rekursiven Verfahren wird anschließend jeweils der Punkt mit dem geringsten berechneten Anpassungsfehler ausgewählt und alle Oberflächenpunkte innerhalb des Radis r_S aus der Datenstruktur entfernt. Auf diese Weise werden Paraboloide mit minimalem Fehler für lokal unabhängige Bereiche ausgewählt, bis keine Oberflächenausschnitte mit Radius $\geq r_S$ übrig sind (Abbildung 7).

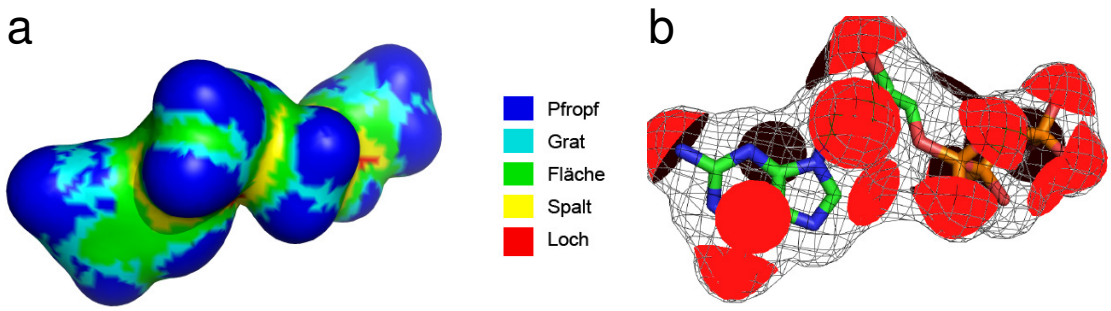


Abbildung 7: Anpassung von Paraboloiden an die Molekülоberfläche von Guanosintriphosphat. a) Einfärbung der extrahierten Isofläche anhand der Oberflächenkurvatur. b) *Shapelets*-Repräsentation der zerlegten Molekülоberfläche.

Aus der maximalen und minimalen lokalen Oberflächenkurvatur k_1 und k_2 an den Zentren p der Oberflächenausschnitte lässt sich der von Duncan und Olson definierte **Shape Index** *SI* (Gleichung 4) berechnen (Duncan & Olson, 1993; Proschak *et al.*, 2007).

$$SI(p_r) = \arctan\left(\frac{k_1 + k_2}{k_1 - k_2}\right). \tag{4}$$

Analog zur Kurvatur der Oberfläche (Abbildung 7a)) lassen sich fünf Klassen von *Shapelets* definieren, deren Form durch einen Shape Index *SI* aus dem Intervall [-1, 1] beschrieben werden (Exner *et al.*, 2002): *gleichmäßig konvex* (Pfropf, $SI \approx 1{,}0$), *langgestreckt konvex* (Grat, $SI \approx 0{,}5$), *flach* (Fläche, $SI \approx 0{,}0$), *langgestreckt konkav* (Spalt, $SI \approx -0.5$) und *konkav* (Loch, $SI \approx -1.0$).

Da Spalten und Löcher auf kleinen Molekülen nur selten vorkommen und Grate oftmals Unebenheiten der Molekülоberfläche annähern, werden im *Shapelets* Algorithmus ausschließlich Paraboloide der Klassen Pfropf (SI = [0,9, 1,0]) und Fläche (SI = [-0,1, 0,1]) verwendet.

Der *Shapelets* Algorithmus zur Formbeschreibung einer Molekülоberfläche ist nachfolgend in einer Übersicht dargestellt (Abbildung 8).

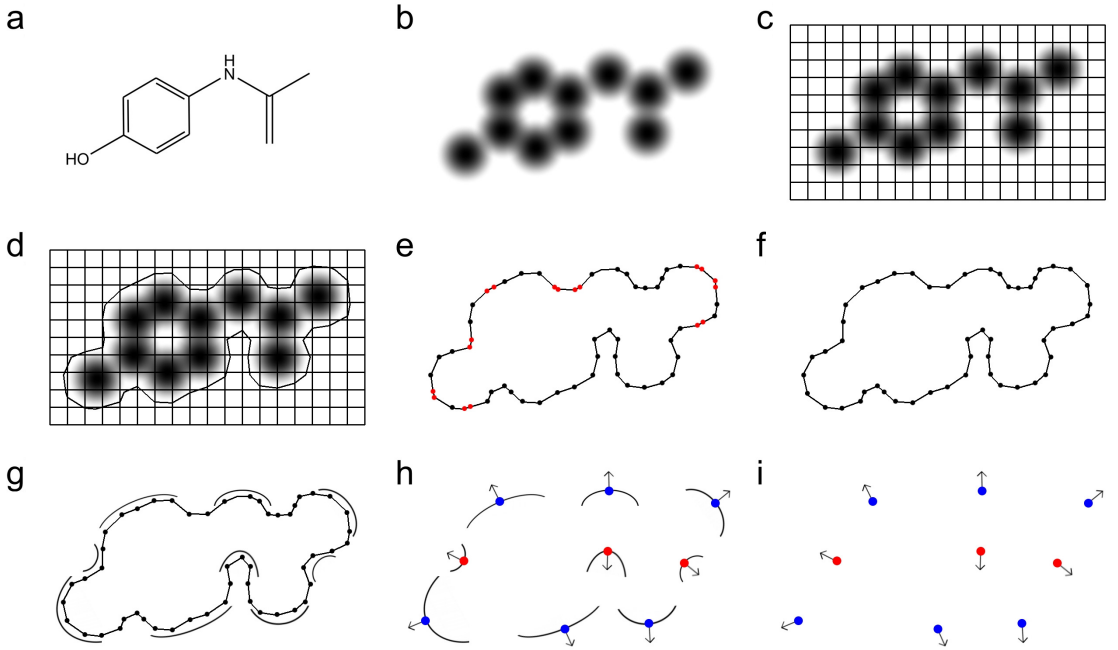


Abbildung 8: *Shapelet* Oberflächenzerlegung eines Moleküls. a) 2D-Molekülgraph von Paracetamol. b) Approximation der Atome durch Gauß'sche Funktionen. c) Projektion der Funktionswerte in ein künstliches Gitter. d) Extraktion der Isofläche mit dem *Marching Cubes* Algorithmus. e) und f) Glättung der Oberfläche mit dem *Welding Edges* Verfahren. g) Anpassung von hyperbolen Paraboloiden an die geglättete Oberfläche. h) und i) Repräsentation des Ausgangsmoleküls durch die Mittelpunkte und Vektornormalen der gezeigten Paraboloide.

2.1.8.2 Überlagerungen von Molekülen mit Shapelets-Methode

Die während der Oberflächenzerlegung berechneten Paraboloide wurden bereits erfolgreich zum virtuellen Screening von wirkstoffartigen Liganden eingesetzt (Proschak *et al.*, 2007). Für den Ähnlichkeitsvergleich nutzt die *Shapelets*-Methode dabei lediglich Beschreibungen der Form von Liganden, die durch die Koordinaten und Kurvaturen der extrahierten Paraboloide dargestellt werden. Der Vergleich verschiedener Ligandenformen gelingt dabei durch die Verwendung eines Verfahrens aus der Graphentheorie: Die **Cliquendetektion auf Assoziationsgraphen** ist ein etabliertes Verfahren zum Vergleichen molekularer Oberflächen (Gardiner *et al.*, 2000; Hofbauer *et al.*, 2004) und gelingt durch Interpretation der für ein Molekül berechneten *Shapelets* als Knoten *V* (engl. *vertices*) in einem **vollständigen Graphen** *G*. Ein Graph heißt vollständig, wenn jeder Knoten *V* aus *G* mit jedem anderen Knoten über Kanten *E* (engl. *edges*) verbunden ist (Gasteiger & Engel, 2003). Der Shape Index *SI* eines

Shapelets wird als Knoteneigenschaft μ_V interpretiert, die Entfernungen zwischen den verschiedenen *Shapelets* werden in den Kanteneigenschaften μ_E gespeichert.
Für den Ähnlichkeitsvergleich zweier Moleküle wird in *Shapelets* zunächst ein Assoziationsgraph $A = (V_a, E_a)$ mit $V_a \subseteq V \times V'$ aus den beiden vollständigen Graphen $G = (V, E, \mu_V, \mu_E)$ und $G' = (V', E', \mu_{V'}, \mu_{E'})$ erstellt:

- Prüfe die *SI* für jedes Paar von Knoten $v \in V$ und $v' \in V'$. Erstelle einen neuen Knoten v_a in A, wenn sich die *SI* von v und v' um weniger als c_{SI} unterscheiden.
- Zwei Knoten $v_a = (v, v')$, $v_b = (w, w') \in V_a$ werden durch eine Kante in A verbunden, wenn die Längen der dazugehörigen Kanten $(v,w) \in E$ und $(v',w') \in E'$ sich um weniger als einen Distanzwert c_D unterscheiden und zudem gilt: $v \neq w$ und $v' \neq w'$.

In einem weiteren Schritt wird eine Suchroutine angewendet, die Cliquen auf dem Assoziationsgraphen identifizieren soll. Eine **Clique** ist als vollständig verbundener Subgraph. Der Aufbau des Assoziationsgraphen dabei ist derart gewählt, dass eine Clique auf A einer räumlichen Übereinstimmung der *Shapelets* auf den zu Grunde liegenden Molekülen entspricht. Da das Problem der Cliquendetektion *NP*-vollständig ist (Hopcroft *et al.*, 2001), verwendet die *Shapelets*-Methode den **Bron-Kerbosch-Algorithmus** (Bron & Kerbosch, 1973). Dieser rekursive Backtracking-Algorithmus zählt alle maximalen Cliquen in einem Graphen auf und ist für die vorliegenden Graphgrößen ausreichend schnell (Brint & Willet, 1987; Gerhards & Lindenberg, 1979). Der Bron-Kerbosch-Algorithmus verfolgt einen sog. „branch-and-bound“ Ansatz, der Berechnungen frühzeitig abbricht, wenn sie nicht mehr zu einem Erfolg führen können.

Das Prinzip der Cliquendetektion auf dem Assoziationsgraphen ist in Abbildung 9 dargestellt. Die vorgestellte Methode findet alle Cliquen in einem Graphen, daher sind verschiedene Überlagerungen möglich.

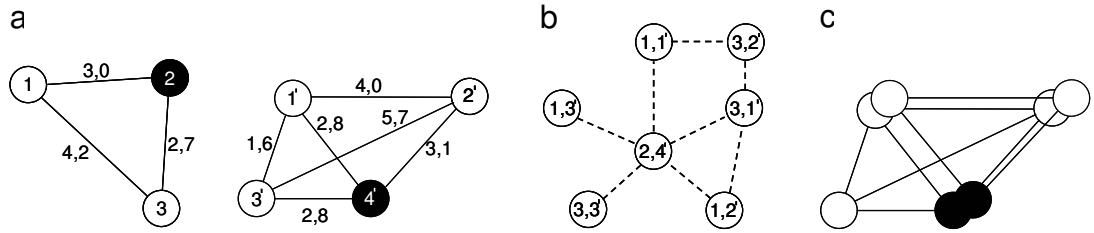


Abbildung 9: Prinzip der Cliquendetektion auf dem Assoziationsgraphen. a) Vollständig verbundene Graphen zweier Moleküle. Die Knoteneigenschaften μ_V sind durch Farben (schwarz und weiß) kodiert, die Kanteneigenschaften μ_E entsprechen den euklidischen Abständen. b) Assoziationsgraph der zu überlagernden Graphen mit c_D = 0,4. c) Bestmögliche Überlagerung der beiden Graphen mit drei gefundenen Cliquen in *A* ([2,4';3,1';1,2']).

Das Alignment der zu Grunde liegenden Moleküle gelingt mit dem Kabsch-Algorithmus (Kabsch, 1976), der eine rigide Überlagerung für die korrespondierenden Knoten aus den Molekülgraphen ermöglicht. Der Algorithmus verfolgt die Minimierung der kleinsten Fehlerquadrate und gliedert sich in zwei Schritte: (i) die Berechnung einer Rotationsmatrix für die gleichartige Orientierung der Knoten im Raum, (ii) die Translation der beiden Knotenmengen für die räumliche Überlagerung an einem Ort (Kabsch, 1976; Lin *et al.*, 2004).

Die hier vorgestellte Methode zur Überlagerung von Molekülen anhand struktureller Merkmale mit der *Shapelets*-Methode wurde in dieser Arbeit für die Überlagerung von Bindetaschen adaptiert und weiterentwickelt. Diese neue Technik ist nachfolgend im Abschnitt „Entwickelte Methoden“ dargestellt.

2.2 Verwendete Datenbanken

2.2.1 PDBbind

Die PDBbind Datenbank (Wang *et al.*, 2004; Wang *et al.*, 2005; http://www.pdbbind.org.cn/) ist eine aus der PDB abgeleitete Sammlung von repräsentativen Proteinkomplexen, für die experimentell bestimmte Affinitäten der jeweiligen Protein-Liganden-Wechselwirkungen vorliegen. In dieser Arbeit wurden die in der PDBbind Datenbank der Version 2007 beschriebenen Datensätze *Core Set* und *Refined Set* zur Abschätzung von Protein Druggability verwendet.

2.3 Entwickelte Methoden

2.3.1 Vorhersage von Proteinbindetaschen mit PocketPicker

Zur Vorhersage möglicher Bindestellen für kleine Moleküle auf Proteinstrukturen wurde der zuvor in einer Diplomarbeit entwickelte ***PocketPicker***-Algorithmus (Weisel, 2006) in dieser Arbeit weiterentwickelt und in der Programmiersprache Java reimplementiert. Die Routine zur Vorhersage potentieller Bindetaschen, sowie der in *PocketPicker* beschriebene Autokorrelationsdeskriptor ***PocketPicker* ShapeDeskriptor** bilden die Grundlage verschiedener Techniken die in der vorliegenden Arbeit entwickelt wurden:

- Beschreibung von Taschenvolumen und -formen
- Extraktion von Taschenoberflächen und Berechnung elektrostatischer Oberflächeneigenschaften
- Alignmentfreier Vergleich von Bindetaschenformen
- Vergleich von Subtaschenformen durch Oberflächenzerlegungen
- Abschätzung von Druggability und Selektivität verschiedener Bindetaschenformen

Nachfolgend sollen die Berechnungsschritte des in dieser Arbeit weiterentwickelten *PocketPicker* Algorithmus dargestellt werden:

PocketPicker nutzt Molekülstrukturen im PDB-Format (Berman *et al.*, 2000) als Eingabe und berechnet Volumina und Vergrabenheiten von potentiellen Bindestellen für kleine Moleküle auf der Proteinoberfläche. Die Information über Größe und Vergrabenheit von Taschenvolumen werden in einem Autokorrelationsdeskriptor ausgedrückt. Der *PocketPicker* Algorithmus gliedert sich in drei Berechnungsschritte:

(i) Identifikation vergrabener Bereiche auf der Proteinoberfläche
(ii) Gruppierung in unabhängige Taschenvolumen
(iii) Berechnung des ShapeDeskriptors

Der *PocketPicker* ShapeDeskriptor kodiert die Form und Vergrabenheit einer vorhergesagten Tasche in einem Autokorrelationsdeskriptor.

2.3.1.1 Berechnung der Vergrabenheit einer Bindetasche

Zu Beginn einer Berechnung wird die Proteinstruktur in ein künstliches Gitter eingepasst, welches das Molekül vollständig umfasst. *PocketPicker* verwendet standardmäßig eine Maschenweite des Gitters von 1 Å. Dieser Wert erlaubt eine Auflösung der Bindetasche mit einer Genauigkeit, die etwa der Länge kovalenter Bindungen im Protein entspricht. Darüber hinaus ermöglicht eine Maschenweite von 1 Å eine intuitive Volumenabschätzung, da ein Knotenpunkt des Gitters 1 $Å^3$ beschreibt.

Die Untersuchung der zu Grunde liegenden Proteinoberfläche gelingt mit Hilfe eines geometrischen Verfahrens, bei dem die molekulare Umgebung rund um oberflächennahe Knotenpunkte des Gitters abgetastet wird. Hierfür werden die Kreuzungspunkte des künstlichen Gitters in drei Klassen eingeteilt: Gitterknoten, die unter der Proteinoberfläche verborgen liegen und solche, die weit von der Oberfläche entfernt liegen, werden von weiteren Berechnungen ausgeschlossen, während an Knoten, die in einem Bereich nahe der Oberfläche liegen, geometrische Sonden installiert werden (Abbildung 10).

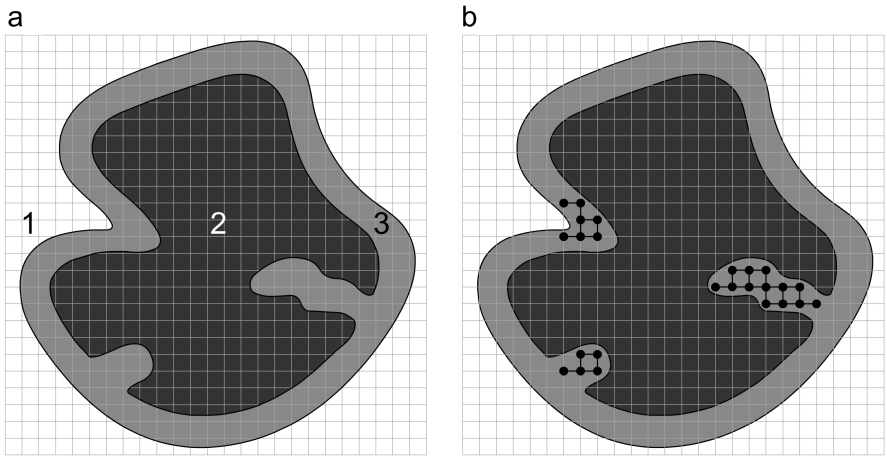


Abbildung 10: Einteilung des künstlichen Gitters in diskrete Bereiche. a) Gitterknoten, die mehr als 4,5 Å von der Oberfläche entfernt liegen (1), sowie Knoten, die unter der Proteinoberfläche (dunkelgrau) verborgen liegen (2) werden bei der Berechnung der Vergrabenheit nicht berücksichtigt. An Gitterknoten in oberflächennahen Bereichen werden Sonden installiert, die die Umgebung abtasten (3). b) Sonden, die in vergrabenen Bereichen der Oberfläche liegen, werden zu potentiellen Bindetaschen gruppiert.

Zur Berechnung ihrer Vergrabenheit untersuchen die installierten Gittersonden ihre Umgebung entlang 30 zuvor definierter Suchrichtungen. Dazu werden in jeder Sonde 30 Vektoren zentriert und so ausgerichtet, dass sie den umgebenden Raum möglichst gleichmäßig aufteilen. Anhand dieser Vektoren wird anschließend die Vergrabenheit festgestellt, die angibt, zu welchem Grad eine jeweilige Sonde mit Protein umgeben ist.
Die Ausrichtung der Vektoren ist ein nicht-triviales Problem und kann auch als die gleichmäßige Verteilung von n Punkten auf einer Kugeloberfläche[5] betrachtet werden (Saff & Kuijlaars, 1997). Im Gegensatz zum Kreismodell ist eine symmetrische Verteilung von Punkten für beliebige n auf der Kugel nicht möglich. Tatsächlich gibt es für das Kugelmodell nur drei komplett symmetrische Lösungen (für $n > 2$), die sich aus der Symmetrie der Platonischen Körper ableiten (Thurston, 1998): Betrachtet man Vektoren, die vom Ursprung zu den Eckpunkten dieser regulären Körper führen, so bieten nur die Vektoren des Tetraeders, des Oktaeders und des Icosaeders eine optimal symmetrische Konfiguration (Abbildung 11).

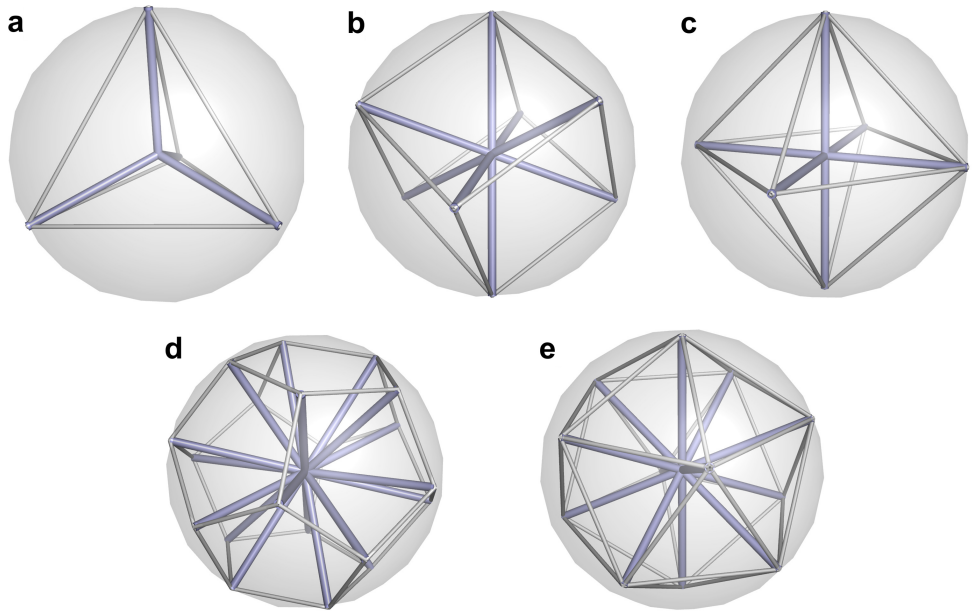


Abbildung 11: Geometrien der fünf Platonischen Körper. a) Tetraeder, b) Hexaeder, c) Oktaeder, d) Dodecaeder, e) Icosaeder. Allein die Ursprungsvektoren (violett), die die Strukturen von a), c) und e) beschreiben, bieten eine total symmetrische Verteilung von Vektoren auf der Kugeloberfläche.

[5] Die gleichmäßige Verteilung von Punkten auf der Kugel ist ein gut untersuchtes Problem in der Geometrie und ist auch als **Tammes' Problem** bekannt (Tammes, 1930; Conway & Sloane, 1988). Optimierungsverfahren verfolgen dabei die Maximierung der minimalen Abstände zwischen den Punkten zur Lösung des Problems.

Perfekt symmetrische Ausrichtungen von Vektoren auf der Kugel gibt es somit nur für die ausgewählten Mengen mit $n \in \{4, 6, 12\}$.

In der Literatur existieren verschiedene iterative Optimierungsverfahren[6], sowie nicht-iterative geometrische Verfahren, die eine annähernd gleichmäßige Verteilung von Punkten im Raum zum Ziel haben. Die Orientierungen der in *PocketPicker* verwendeten 30 Suchvektoren sind nach einem geometrischen Verfahren ermittelt worden, das auf der Geometrie der Platonischen Körper basiert: Die Flächen der Platonischen Körper sind regelmäßige Polyeder, die aus identisch großen gleichseitigen und gleichwinkligen Vielecken bestehen. Diese Flächen lassen sich durch **Triangulation** in kleinere Deckungsgleiche Dreiecke zerlegen (Thurston, 1998). Durch die iterative Triangulation wurden in den acht Dreiecksflächen des Oktaeders je drei Vektoren ausgerichtet, die sich zusammen mit den sechs Vektoren entlang der Raumachsen zu den gewünschten 30 Vektoren aufsummieren (Abbildung 12).

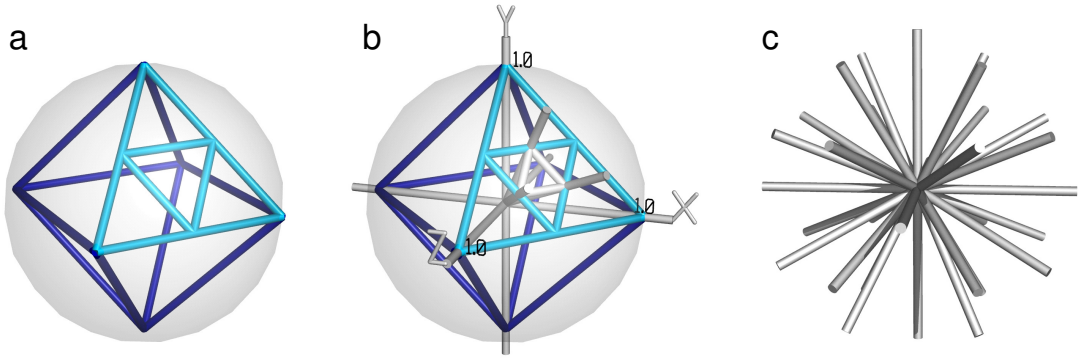


Abbildung 12: Ausrichtung der Suchvektoren auf der Einheitskugel S^2 in *PocketPicker*. a) Triangulation einer Fläche des Oktaeders in vier deckungsgleiche Dreiecke (hellblau). b) Durch Triangulation des mittleren Dreiecks (hellblau) wird ein neues zentrales Dreieck (weiß) beschrieben. Durch die Eckpunkte dieses Dreiecks werden Vektoren geführt und auf die Oberfläche von S^2 verlängert. c) Orientierung der 30 in *PocketPicker* verwendeten Suchvektoren.

Die Zugänglichkeit diskreter oberflächennaher Bereiche wird durch die zuvor installierten Gittersonden bestimmt, für die ein Vergrabenheits-Index berechnet wird. Eine Gittersonde untersucht ihre Umgebung anhand von 30 Suchstrahlen, deren

[6] Evolutionäre Strategien (Bäck & Schwefel, 1993) erzeugen neue Punktverteilungen als Nachkommen einer Startkonfiguration und optimieren schrittweise die minimalen Distanzen. Im Partikelmodell sind die Punkte als gleichmäßige Teilchen definiert, die sich gemäß der Coulomb'schen Wechselwirkungen (Rottler & Maggs, 2004) selbsttätig über die Kugeloberfläche verteilen (Morris *et al.*, 1995; Atiyah & Sutcliffe, 2002).

Orientierungen durch die Triangulation des Oktaeders bestimmt wurden. Diese Suchstrahlen sind mit einer Länge von 10 Å und einer Breite von 0,9 Å definiert. Der Vergrabenheits-Index einer Sonde entspricht nun der Anzahl der Suchstrahlen, in deren Ausdehnung mindestens ein Atom des Proteins detektiert wurde. Somit wird für jede Sonde ein Vergrabenheits-Index berechnet, dessen Werte zwischen null und 30 liegen und die Zugänglichkeit des jeweiligen Sondenpunktes beschreibt (Abbildung 13).

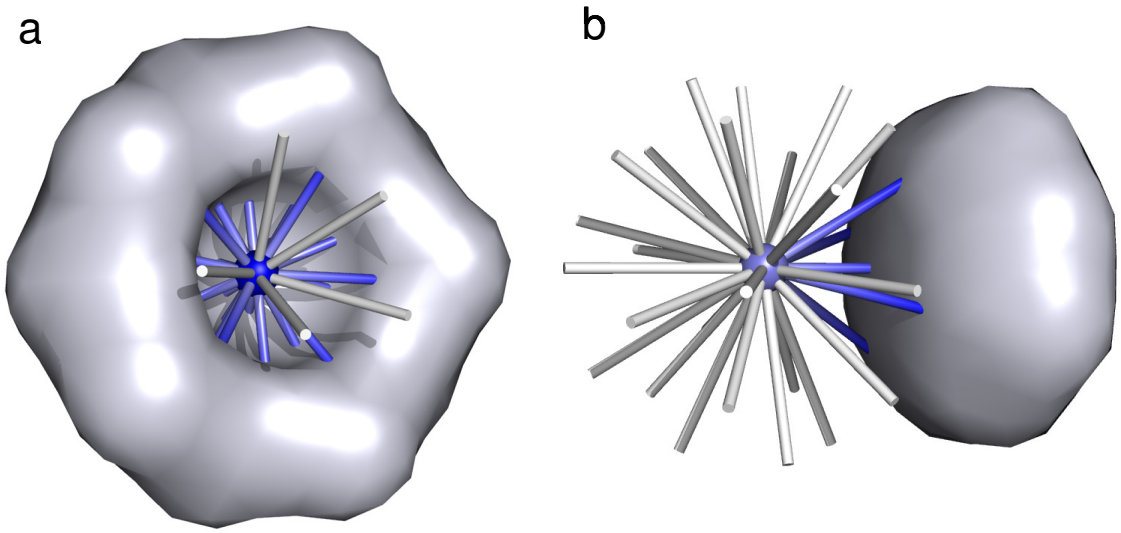


Abbildung 13: Berechnung der Vergrabenheit an konvexen und konkaven Modelloberflächen. a) Eine Sonde trifft mit 25 ihrer 30 Suchstrahlen auf Atome des Proteins (blaue Vektoren), woraus ein Vergrabenheits-Index von 25 resultiert. b) Auf einer konvexen Oberfläche zielen die meisten Suchstrahlen ins Leere (weiße Geraden), ein Vergrabenheits-Index von fünf wird ermittelt. Steigende Vergrabenheiten werden durch eine stärkere Blaufärbung der Sonde dargestellt.

Die algorithmische Umsetzung der Vergrabenheitsbestimmung gelingt durch 30 Richtungsvektoren $\vec{u}$ die entlang von 30 Geraden *G* (ermittelt durch Triangulation des Oktaeders) ausgerichtet und auf Länge 1 skaliert werden. Die in *PocketPicker* verwendeten Suchstrahlen werden in einer Sonde *P* (engl. *probe*, repräsentiert durch den Vektor $\vec{p}$) zentriert und auf die genannten Dimensionen skaliert. Ein nahe gelegenes Proteinatom Q ($\vec{q}$) wird vom Suchstrahl detektiert, wenn zwei Bedingungen erfüllt sind:

(i) Die Länge der orthogonalen Projektion *d* auf den jeweiligen Richtungsvektor darf die Breite des Suchstrahls nicht übersteigen.

(ii) Der Abstand des aus der Projektion resultierenden Punktes *X* zur Sonde darf die Länge des Suchstrahls nicht übersteigen.

Bei der Berechnung des orthogonalen Abstandes *d* (engl. *distance*) auf den Richtungsvektor kann auf eine Normierung (Division durch $|\vec{u}|$) verzichtet werden, da alle Vektoren die Länge eins besitzen (Gleichung 5).

$$d = |(\vec{q} - \vec{p}) \times \vec{u}| \text{, wenn } |\vec{u}| = 1. \tag{5}$$

Der Abstand zwischen der Sonde *P* und Punkt *X* wird als die Länge des Richtungsvektors $\vec{u}$ skaliert mit einem Faktor *t* berechnet (Gleichung 6).

$$t = (\vec{q} - \vec{p}) \cdot \vec{u} \text{, wenn } |\vec{u}| = 1. \tag{6}$$

Der Detektionsprozess ist in Abbildung 14 dargestellt. Sämtliche Koordinaten und Ortsvektoren beziehen sich auf den Kartesischen Ursprung *O* (engl. *origin*).

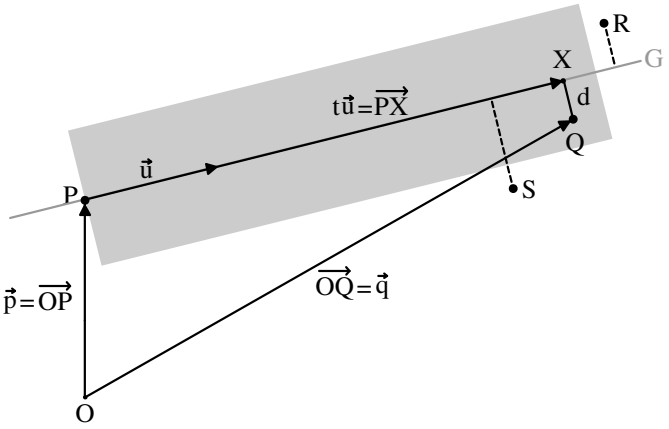


Abbildung 14: Berechnung des Vergrabenheits-Index einer Sonde. Ein Suchstrahl (grau) untersucht seine Umgebung nach Atomen und findet Atom *Q*. Atome *R* und *S* werden nicht erfasst, da sie außerhalb der Dimensionen des Suchstrahls liegen.

Bei der Berechnung der Vergrabenheit der Gittersonden wurde in dieser Arbeit ein Ansatz implementiert, der eine deutliche Vereinfachung des Rechenaufwandes darstellt. So beginnt die Berechnung nicht in den Sonden selbst, sondern iteriert über die Atome, die von den Sonden erfasst werden sollen (Abbildung 15).

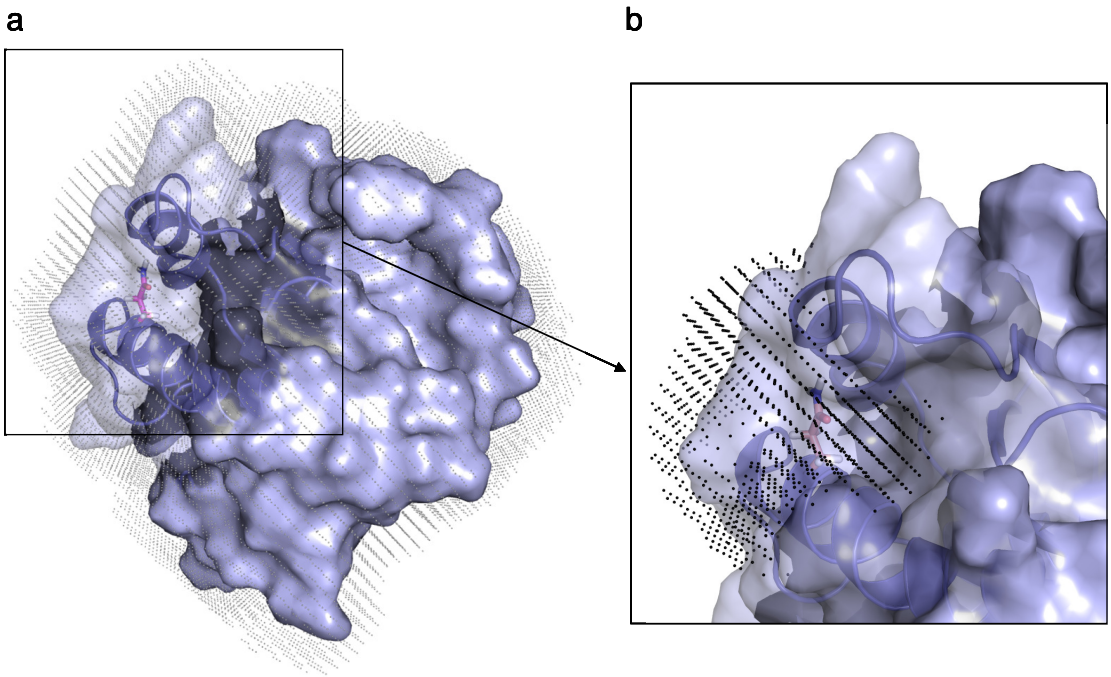


Abbildung 15: Einschränkung des Suchraums bei der Berechnung der Vergrabenheit. a) In oberflächennahen Bereichen eines Proteins (menschliches Krebsprotein P21-H-RAS, PDB: 121p) werden Sonden installiert (grau). b) Auswahl der Gittersonden im Radius von 10 Å um das C-β Atom des Glutamin-99 Restes (magenta). Der Vergrabenheits-Index einer Sonde (schwarz) wird um einen Zähler erhöht, falls ein Suchstrahl das aktuelle Atom ertastet. Der betreffende Suchstrahl wird für die Detektion weiterer Atome deaktiviert.

So werden ausgehend von einem aktuellen Proteinatom die umliegenden Gittersonden aktiviert, die ihrerseits prüfen, ob einer ihrer Suchstrahlen dieses aktuelle Atom erreicht. Ein Suchstrahl wird nach einer erfolgreichen Detektion deaktiviert und von weiteren Berechnungen ausgeschlossen. Dies soll verhindern, dass ein Suchstrahl eines Atoms in mehreren Aufrufen verschiedene Atome ertastet und den Vergrabenheits-Index wiederholt erhöhen kann. Auf diese Weise bleibt gewährleistet, dass eine Sonde mit ihren 30 Suchstrahlen einen Vergrabenheits-Index von höchstens 30 erreicht. Die Routine zur Berechnung der Vergrabenheit der Gittersonden ist nachfolgend in einem Pseudocode dargestellt:

```
Für alle Atome A_i:
      Identifiziere alle Gittersonden P_j, die nicht weiter als 10 Å
      (Länge der Suchstrahlen) von A_i entfernt sind.

      Für alle Suchstrahlen S_k von P_j:
            Falls S_k noch nicht markiert wurde:
                  Falls A_i innerhalb Länge und Breite von S_k:
                        Markiere S_k.
                        Erhöhe den Vergrabenheits-Index von P_j
                        um eins.
```

Eine weitere Beschleunigung der Taschenvorhersage wird durch die Ausrichtung der Suchstrahlen durch Triangulation des Oktaeders erreicht. Im Unterschied zur Verteilung der Suchvektoren anhand von Optimierungsverfahren garantiert die Triangulation eine spiegelsymmetrische Ausrichtung der Suchstrahlen. Dies vereinfacht die Berechnung erheblich, da die verwendeten 30 Suchrichtungen durch 15 Vektoren simuliert werden können. So genügen zur Bestimmung der Vergrabenheit die 15 Vektoren, die in eine Hemisphäre des die Sonde umgebenden Raumes gerichtet sind. Diese kann modellhaft als die ‚obere' Hemisphäre über einer gedachten *xz*-Ebene (genauer: $y \geq 0$, wenn $x > 0$; $y > 0$, sonst) aufgefasst werden (Abbildung 16). Die Berechnung, ob ein Atom von einem Suchstrahl erfasst wird, erfolgt nur noch für die Vektoren der oberen Hemisphäre. Wird bei der Berechnung des Abstandes zur Sonde ein Faktor *t* (Gleichung 6, Abbildung 14) mit negativem Vorzeichen für ein Atom berechnet, so befindet sich dieses in der unteren Hemisphäre der Sonde. Falls das Atom innerhalb der Längen- und Breitenbegrenzungen des aktuellen Suchstrahls liegt, wird nur sein punktgespiegelter Vektor auf der unteren Hemisphäre markiert und deaktiviert. Der Vergrabenheits-Index wird entsprechend um einen Zähler erhöht.

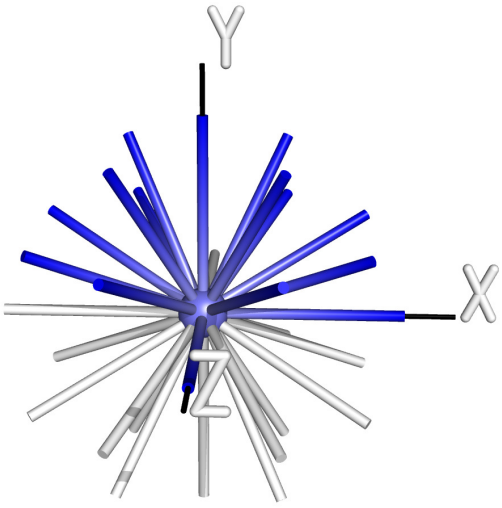


Abbildung 16: Aufteilung der des Suchraums in zwei Hemisphären. Die weißen Suchstrahlen sind punktsymmetrische Spiegelungen von Vektoren der ‚oberen' Hemisphäre (blau) am Koordinatenursprung.

Die Berechnung der Vergrabenheiten oberflächennaher Gittersonden dient der Beschreibung der Zugänglichkeit von diskreten Bereichen des Proteins. Anhand der

Vergrabenheits-Indizes lassen sich zusammenhängende Volumina auf der Proteinoberfläche identifizieren, die mögliche Bindestellen für niedermolekulare wirkstoffartige Liganden darstellen. Dies gelingt durch das Gruppieren (engl. *Clustering*) benachbarter Gittersonden mit erhöhter Vergrabenheit und ist nachfolgend geschildert.

2.3.1.2 Clustering von Gittersonden

Die Vorhersage von potentiellen Bindetaschen gelingt in *PocketPicker* über die Identifikation zusammenhängender vergrabener Bereiche auf der Proteinoberfläche. Aus diesem Grund werden für die anschließenden Berechnungen nur Sonden betrachtet, die zu mehr als 50% von Protein umgeben sind und somit einen Vergrabenheits-Index von mindestens 15 besitzen. Alle übrigen Gittersonden werden verworfen.

Potentielle Bindetaschenvolumen werden durch benachbarte Gittersonden mit erhöhter Vergrabenheit beschrieben. Die Verwendung eines kubischen künstlichen Gitters als Grundlage von *PocketPicker* ermöglicht dabei eine einfache Definition von Nachbarschaft, die auf aufwändige Abstandsberechnungen verzichtet und von dem Automatenmodell der **zellulären Automaten**[7] (Wolfram, 1983; Wolfram, 1984) abgeleitet ist. So beschreibt die sogenannte **von Neuman-Nachbarschaft** die orthogonal benachbarten Bereiche einer Zelle im Zellulären Automaten (auch diamantenförmige Nachbarschaft genannt), während die **Moore-Nachbarschaft** sämtliche angrenzende Zellen umfasst (Abbildung 17).

In dem hier beschriebenen Clusteringverfahren wird die Nachbarschaft einer Gittersonde als Moore-Nachbarschaft erster Ordnung (Radius $r = 1$) definiert. Für die in *PocketPicker* verwendeten dreidimensionalen Gitter bedeutet dies, dass die Nachbarschaft einer Gittersonde aus 26 angrenzenden Sonden besteht.

[7] Zelluläre Automaten modellieren mathematisch komplexe Probleme als abstrakte Simulation auf künstlichen Gittern. Diese können verschiedene Topologien besitzen und sind aus Zellen aufgebaut, die zuvor definierte Zustände besitzen. In diskreten Zeitschritten können die Zellen ihre Zustände anhand von festgelegten Übergangsregeln und unter dem Einfluss ihrer Nachbarzellen ändern.

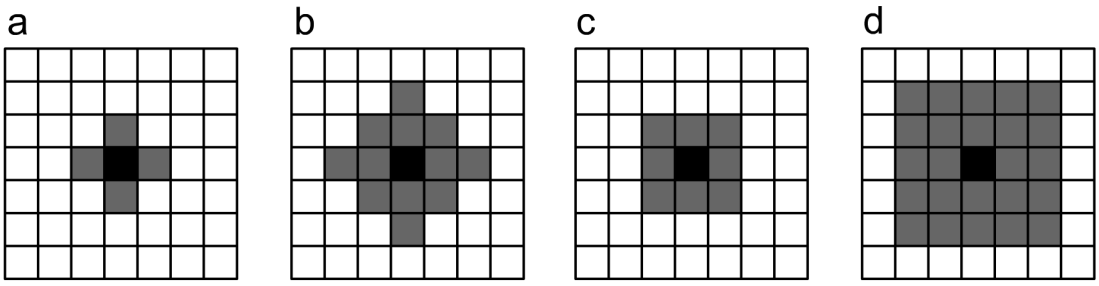


Abbildung 17: Nachbarschaftsdefinitionen in künstlichen Gittern Zellulärer Automaten. Die von Neumann-Nachbarschaft (a, b) betrachtet orthogonal benachbarte Zellen, während die Moore-Nachbarschaft (c, d) alle umgebenden Felder einbezieht. Dargestellt sind die Nachbarschaften mit Radius *r* = 1 und *r* = 2 in zwei-dimensionalen Gittern.

Ziel des Clusteringverfahrens ist das Gruppieren von Gittersonden mit Vergrabenheits-Indizes zwischen 15 und 26 in disjunkte Cluster, wobei die Sonden eines Clusters durch eine Moore-Nachbarschaft verbunden sein müssen. Zwei Cluster sind also unabhängig voneinander, wenn es keine Sonde gibt, die die beiden Punktwolken mit seiner Moore-Nachbarschaft verbindet.

Als Clusteringverfahren wurde in dieser Arbeit eine rekursive Methode implementiert, die alle Gruppen zusammenhängender Sonden erkennt und anschließend als Cluster absteigender Größe sortiert. Als Startpunkt der Prozedur dient eine der verbliebenen Sonden (mit Vergrabenheit zwischen 15 und 26), die zufällig ausgewählt wird. Ausgehend von diesem Punkt werden von den höchstens 26 direkten Nachbarn diejenigen Sonden ausgewählt, die ebenfalls bisher nicht aus dem Gitter entfernt wurden. Diese verbliebenen Nachbarn werden in einer Auswahl gespeichert, die nachfolgend als **Selection** bezeichnet wird. Im rekursiven Schritt werden für jede Sonde der Selektion eigene Suchen in der jeweiligen Nachbarschaft gestartet. Um Redundanzen zu vermeiden, werden die bereits in der Selection gespeicherten Sonden bei den rekursiven Aufrufen ignoriert, so dass die Sonden einer Selection sich nicht gegenseitig aufspüren können. Die Rekursion terminiert, wenn keiner der gestarteten Aufrufe neue Sonden in seiner Nachbarschaft finden kann. Sämtliche nun in der Selection gespeicherten Gittersonden sind Mitglieder eines Clusters. Unter den Sonden, die von dem ursprünglichen Startpunkt nicht erreicht wurden, weil sie mit diesem nicht über andere Sonden verbunden sind, wird eine zufällig gewählt und eine neue rekursive Nachbarschaftssuche gestartet. Die Methode terminiert, wenn jede Sonde einer

Selection zugewiesen wurde. Die Selections entsprechen genau der Menge der zusammenhängenden und voneinander disjunkten Clustern einer Berechnung. Die so identifizierten Cluster zeigen mögliche Bindetaschen auf der Oberfläche eines Proteins an. Der größte identifizierte Cluster wird als wahrscheinlichste Bindestelle interpretiert.
Das vorgestellte Verfahren zum Auffinden zusammenhängender Gruppen von Gittersonden entspricht einem Problem in der Graphentheorie. Es ähnelt dem sogenannten Flood-Fill-Algorithmus (auch Seed-Fill-Algorithmus; Feng & Soon, 1998) zur Identifikation von Zusammenhangskomponenten in ungewichteten, ungerichteten Graphen.
Durch die großzügig gewählte Definition der Nachbarschaft im dreidimensionalen Gitter kann es jedoch passieren, dass eng benachbarte Cluster miteinander verknüpft werden, wenn sie nur ‚lose' über eine Sonde verbunden sind (Abbildung 18). Auf diese Weise können zwei eigentlich getrennte Bindetaschen irrtümlich als eine große Bindestelle interpretiert werden, was die Genauigkeit der Vorhersage beeinträchtigt.

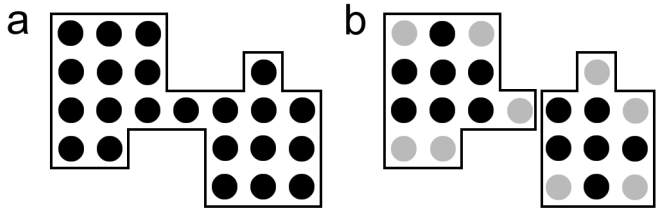


Abbildung 18: Clustering von Gittersonden. a) Gruppierung zweier benachbarter Cluster zu einer großen Tasche. b) Eine feinere Einteilung in zwei disjunkte Taschen gelingt durch den Ausschluss von Sonden, die nur wenige Nachbarn besitzen (graue Punkte). Diese werden im Anschluss an das Clustering dem jeweils größeren benachbarten Cluster zugewiesen.

Eine feinere Abgrenzung der Bindetaschen gegeneinander wird daher durch eine Modifikation des Clusteringverfahrens gewährleistet. So ist die rekursive Nachbarschaftssuche auf diejenigen Sonden beschränkt, die mehr als zehn weitere Sonden in ihrer Moore-Nachbarschaft besitzen. Gittersonden mit höchstens zehn direkten Nachbarn werden aber nicht aus dem Gitter entfernt, sondern im Anschluss an das Clusterverfahren an den jeweils größten Cluster in ihrer Nachbarschaft angeschlossen. Dieses modifizierte Clusterverfahren erlaubt die gewünschte Einordnung in disjunkte Punktwolken, die mögliche Bindevolumen auf der Proteinoberfläche markieren (Abbildung 19).

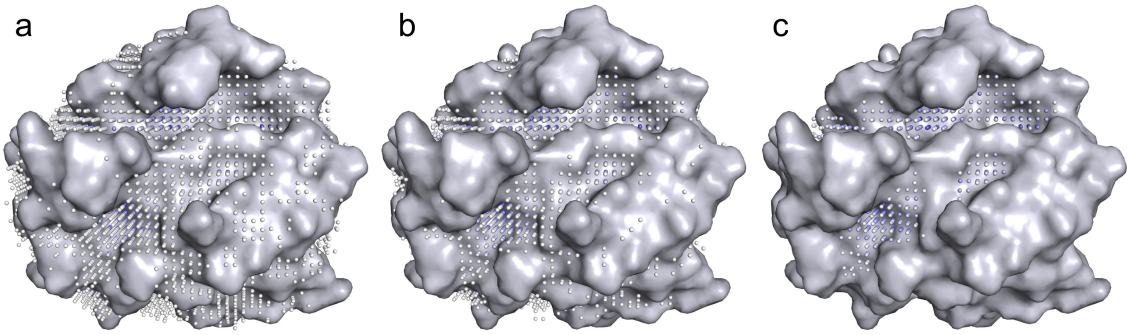


Abbildung 19: Rolle der Vergrabenheits-Indizes beim Clustering am Beispiel von P21-H-RAS (PDB: 121p). a) Visualisierung aller Sonden mit Vergrabenheit >10. b) Die Darstellung aller Sonden mit Vergrabenheit >12 lässt vergrabene Bereiche erkennen. c) Ergebnis der Taschenvorhersage mit modifiziertem Clusteringverfahren. Nur Sonden mit Vergrabenheit zwischen 15 und 26 werden berücksichtigt.

2.3.1.3 Kodierung von Bindetaschen mit Autokorrelationsdeskriptoren

Für einen schnellen und Alignment-freien Vergleich von Bindetaschen wurde in *PocketPicker* eine Methode verwirklicht, die die Form und Vergrabenheit einer vorhergesagten Tasche in einem Korrelationsdeskriptor kodiert. Dazu werden die zu Taschenvolumen gruppierten Gittersonden sechs Kategorien von A bis F mit ansteigenden Vergrabenheitswerten eingeteilt: **A**: 15–16, **B**: 17–18, **C**: 19–20, **D**: 21–22, **E**: 23–24, **F**: 25–26. Ein Shape-Deskriptor wurde entwickelt, der die Häufigkeiten der Abstände zwischen Sonden der verschiedenen Kategorien in einem Korrelationsdeskriptor zusammenfasst. In dieser Arbeit wurde ein Shape-Deskriptor verwendet, der das Auftreten für alle 21 möglichen Kombinationen dieser Kategorien über eine Distanz von bis 10 Å betrachtet. Der resultierende 210-dimensionale Deskriptor sortiert die beobachten Sondenpaare in sogenannte **Bins**, in die die Kategoriepaare wie folgt eingeteilt werden (Binpositionen in Klammern):

AA (1-10), **AB** (11-20), **AC** (21-30), **AD** (31-40), **AE** (41-50), **AF** (51-60),
BB (61-70), **BC** (71-80), **BD** (81-90), **BE** (91-100), **BF** (101-110),
CC (110-120), **CD** (121-130), **CE** (131-140), **CF** (141-150),
DD (151-160), **DE** (161-170), **DF** (171-180),
EE (181-190), **EF** (191-200),
FF (201-210).

So gibt das erste Bin eines Shape-Deskriptors an, wie viele Sondenpaare der Kategorie A im Abstand $d \leq 1$ Å zueinander in der betrachteten Bindetasche vorliegen. Das 210-te Bin listet die Häufigkeit von FF-Sondenpaaren mit Abstand 9 Å $< d \leq 10$ Å.
Eine von *PocketPicker* vorhergesagte Bindetasche kann so über einen Shape-Deskriptor ausgedrückt werden, der einem Vektor im 210-dimensionalen Raum entspricht. Ähnlichkeitsvergleiche zwischen zwei Bindetaschen r und s gelingen über die Berechnung des **Euklidischen Abstandes** d zwischen ihren Shape-Deskriptoren (Gleichung 7).

$$d = \sqrt{\sum_{i=1}^{210} (r_i - s_i)^2} \; . \tag{7}$$

Der vorgestellte Shape-Deskriptor kodiert Form und Vergrabenheit eines Taschenvolumens in einem Beschreiber, der für alignmentfreies Vergleichen von Bindetaschen und zur Vorhersage der Protein-Druggability verwendet werden kann. Die Berechnung eines Shape-Deskriptors für eine Bindetasche ist in Abbildung 20 gezeigt.

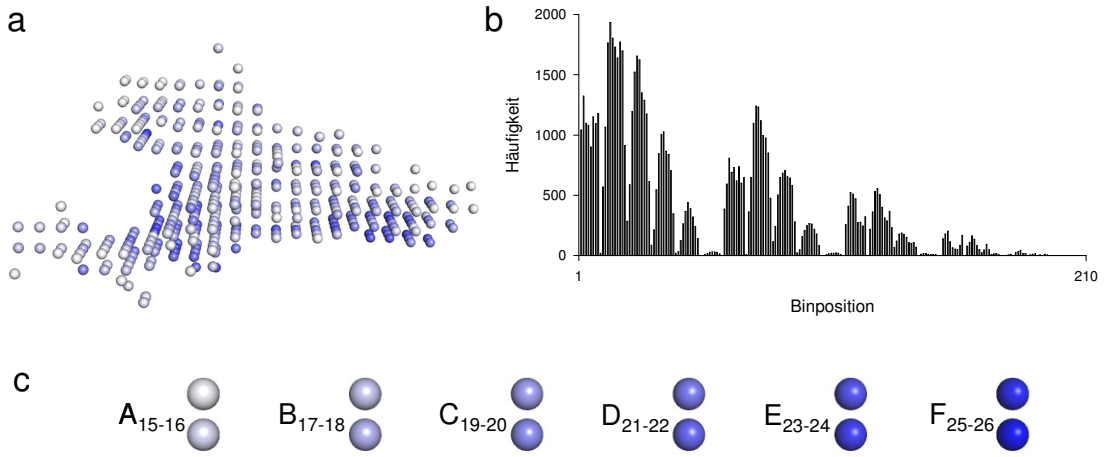


Abbildung 20: Berechnung des Shape-Deskriptors. a) Darstellung der größten mit *PocketPicker* gefundenen Tasche von P21-H-RAS (PDB: 121p). b) Shape-Deskriptor der gezeigten Tasche als Histogramm. c) Farblegende der Vergrabenheitswerte: Sonden mit größerer Vergrabenheit werden in kräftigeren Blautönen dargestellt.

Ein ähnlicher Ansatz wurde in einer vorhergehenden Arbeit zur automatischen Klassifikation von Enzymklassen verwendet (Stahl *et al.*, 2000).

2.3.2 Vergleich von Proteinbindetaschen mit PocketomePicker

Als eine Weiterentwicklung von *PocketPicker* wurde in dieser Arbeit das Programm ***PocketomePicker*** erstellt, welches das strukturelle Überlagern (engl. *Matching*) von Bindetaschen erlaubt. Dieses Programm verfolgt einen im Vergleich zu den Shape-Deskriptoren deutlich aufwändigeren Ansatz zum Ähnlichkeitsvergleich von Taschenformen, der Potentialeigenschaften von Bindestellen berücksichtigt und eine Visualisierung transformierter Taschen ermöglicht.

Grundlage des Verfahrens sind die mit *PocketPicker* vorhergesagten Taschenvolumen, die in *PocketomePicker* als Gauß'sche Dichtefunktion interpretiert werden. Diese Darstellung erlaubt die Extraktion einer Taschenoberfläche, die den Raum eingrenzt, der von potentiellen Liganden eingenommen werden kann.

Für die Überlagerung und Ähnlichkeitssuche von Bindetaschenformen wurde die Methode zur Oberflächenzerlegung ***Shapelets*** verwendet. Diese Technik wurde ursprünglich für das ligandenbasierte **Virtuelle Screening** von wirkstoffähnlichen Molekülen entworfen (Proschak *et al.*, 2007; 2008). Im Rahmen dieser Arbeit wurde die Methode ***PocketShapelets*** entwickelt, die den Formvergleich von Bindetaschenoberflächen auf Grundlage des *Shapelets*-Algorithmus ermöglicht. Zudem wurde in dieser Weiterentwicklung die Betrachtung von elektrostatischen und lipophilen Potentialen berücksichtigt, die in den *PocketShapelets* kodiert werden. Die Vorgehensweise dieser neuen Methode ist nachfolgend dargestellt.

2.3.2.1 Oberflächenbeschreibung von Bindetaschen mit PocketShapelets

Die Vorgehensweise der *PocketShapelets*-Methode zur Zerlegung von Proteinoberflächen und der strukturellen Überlagerung (engl. *Matching*) ähnlicher Taschenformen verhält sich analog zur Ablauf des *Shapelets*-Algorithmus (Kapitel 2.1.8). Die Unterschiede der beiden Methoden beschränken sich auf die Extraktion der Isoflächen: Während der *Shapelets*-Ansatz eine Zerlegung von Molekülоberflächen verfolgt, nutzt der *PocketShapelets*-Algorithmus Ergebnisse aus der Taschenvorhersage von *PocketPicker*, um eine Oberflächenbeschreibung vorhergesagter Volumina zu definieren. So werden Gauß'sche Funktionen in den von *PocketPicker* bestimmten

Gittersonden zentriert, die eine mögliche Bindetasche beschreiben. Die weiteren Berechnungsschritte gestalten sich analog zu den Berechnungsschritten im *Shapelets*-Algorithmus (Kapitel 2.1.8.1). Der Ablauf der *PocketShapelets*-Methode gliedert sich daher in die folgenden Schritte:

- Extraktion einer potentiellen Bindetasche mit *PocketPicker*
- Repräsentation des Volumens durch Installation von Gauß'schen Funktionen in den Gittersonden
- Extraktion der Isofläche mit *Marching Cubes* Algorithmus
- Zerlegung der Oberfläche, Darstellung durch hyperbole Paraboloide

Die Berechnungsschritte der *PocketShapelets*-Methode sind in Abbildung 21 dargestellt.

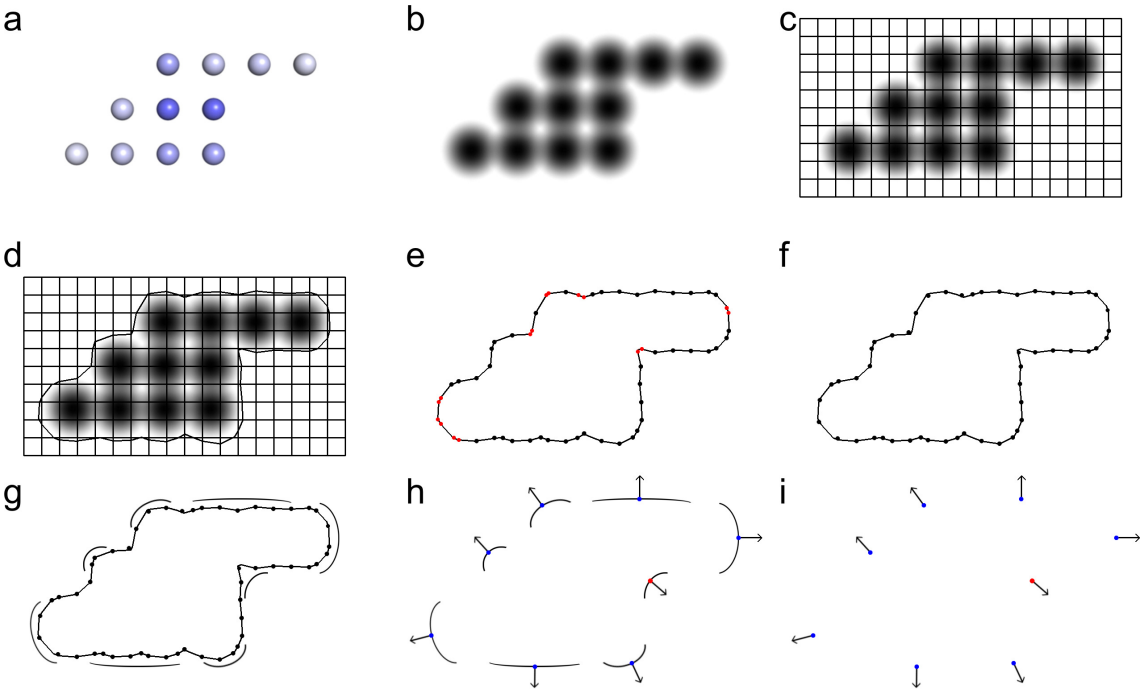


Abbildung 21: Beschreibung der Oberflächenkurvatur einer Bindetasche durch *PocketShapelets*. a) Zweidimensionale Darstellung einer Bindetasche durch die von *PocketPicker* verwendeten Gittersonden. b) Repräsentation der extrahierten Bindetasche durch Gauß'sche Funktionen. c) Projektion der Funktionswerte in ein künstliches Gitter. d) Extraktion der Isofläche mit dem *Marching Cubes* Algorithmus. e) und f) Glättung der Oberfläche mit dem *Welding Edges* Verfahren. g) Anpassung von hyperbolen Paraboloiden an die geglättete Oberfläche. h) und i) Repräsentation der Oberfläche der betrachteten Bindetasche durch die Mittelpunkte und Vektornormalen der gezeigten Paraboloide.

Die Beschreibung der Oberflächenkurvatur einer Bindetasche durch hyperbole Paraboloide (nachfolgend *PocketShapelets* genannt) erlaubt einen Ähnlichkeitsvergleich von Bindetaschen durch strukturelle Überlagerung gleichartiger Taschenformen. Dies gelingt durch Anwendung der Cliquendetektion auf dem Assoziationsgraphen, der als Bron-Kerbosch-Algorithmus (Bron & Kerbosch, 1973) in der *Shapelets*-Methode implementiert ist.

Im Unterschied zur Oberfläche eines Moleküls ist die Oberfläche einer Bindetasche keine vollständig geschlossene Hülle. Während die am Protein anliegende Seite eines Bindevolumens die Beschaffenheit der Struktur beschreibt, besitzt die dem Medium zugewandte Seite der Tasche keine definierte Form. Um ein Anpassen von *PocketShapelets* an diese offene Seite der Bindestelle zu verhindern, werden die lösungsmittelzugänglichen Teile der extrahierten Isofläche vor der Berechnung der hyperbolen Paraboloide entfernt. Die Unterscheidung zwischen strukturnahen und dem Medium zugewandten Bereichen der Taschenoberfläche gelingt durch die Betrachtung der Gauß'schen Oberfläche des Proteins. Das zu Grunde liegende Protein wird in ein künstliches Gitter eingepasst und durch die Summe von Gaußfunktionen repräsentiert, die in den Proteinatomen zentriert sind. Die Gaußfunktionen und ein Schwellenwert c_1 (engl. *cutoff*) sind so gewählt, dass sie die van-der-Waals-Oberfläche des Proteins approximieren (Gleichung 3). Gitterpunkte mit Funktionswerten $<c_1$ liegen außerhalb der Proteinhülle. Für das Abschneiden der zum Medium gewandten Bereiche der Taschenoberfläche wurde ein Cutoff c_2, mit $c_2 = c_1 - 0{,}12$ gewählt. Die Oberfläche der Tasche ist trianguliert und setzt daher sich aus Dreiecksflächen zusammen. Eine solches Dreieck wird aus der Oberfläche entfernt, wenn der Gauß'sche Funktionswert $<c_2$ ist. Der Schwellenwert c_2 wurde aufgrund visueller Betrachtungen gewählt und erlaubt eine genaue Abtrennung der nach außen gerichteten Teile der Taschenoberfläche (Abbildung 22).

Die vorgestellte Methode *PocketShapelets* erlaubt eine deterministische Zerlegung der mit *PocketPicker* extrahierten Bindetaschen. Die reduzierte Darstellung der Oberflächenkurvatur durch hyperbole Paraboloide ermöglicht zudem die strukturelle Überlagerung der Bindetaschen durch die in *Shapelets*-Methode vorgestellten Techniken.

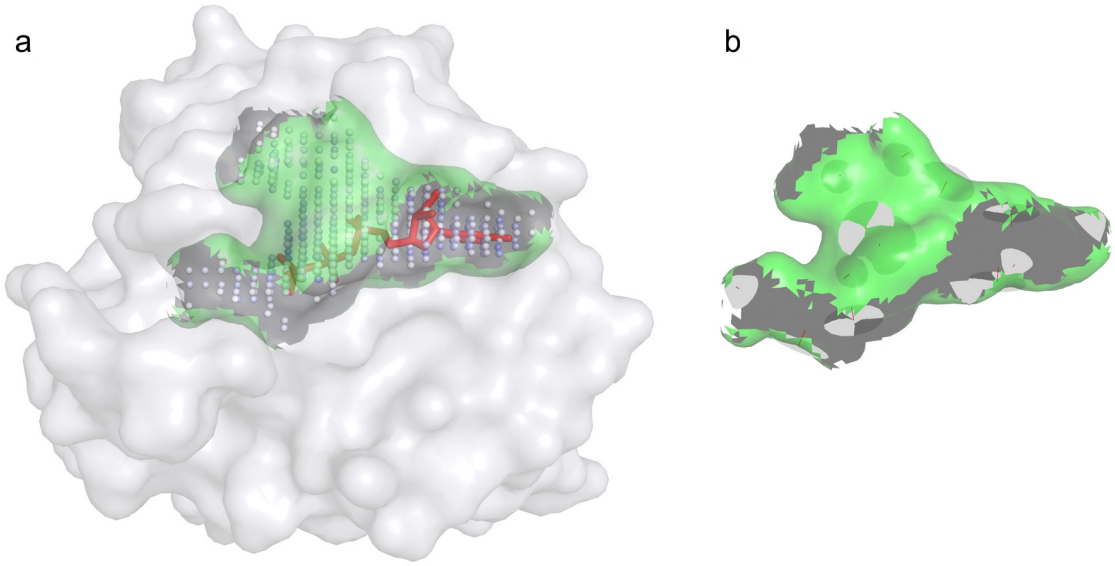


Abbildung 22: Extraktion der Taschenoberfläche für das Krebsprotein P21-H-RAS (hellgrau, PDB: 121p) und Berechnung der PocketShapelets. a) Die Bindetasche von GTP (rot, PDB: gcp) wird von *PocketPicker* korrekt identifiziert und die Gauß'sche Taschenoberfläche (grün) mit dem *Marching Cubes* Algorithmus extrahiert. b) Die dem Lösungsmittel zugewandeten Abschnitte der Taschenoberfläche werden vor dem Anpassen der PocketShapelets (hellgrau) entfernt.

In der *Shapelets*-Methode werden die formbeschreibenden Paraboloide durch Septupel kodiert. Dieses definiert ein Paraboloid über die *x*-, *y*- und *z*-Koordinaten der Shapeletzentren, die Koordinaten der Vektornormalen und den Shape Index *SI*, der die Form des Paraboloiden definiert. In dieser Arbeit wurden weitere Methoden zur Charakterisierung der in der Tasche vorherrschenden physikochemischen Eigenschaften implementiert. So gehen Abschätzungen des lipophilen/hydrophilen und des elektrostatischen Potentials als weitere Beschreiber in die *PocketShapelets* ein. Die Berechnung dieser Potentiale erfolgt am Zentrum eines Paraboloids und ist im folgenden Abschnitt beschrieben. Anwendungen zum Vergleich von Bindetaschenformen und ihren Eigenschaften sind im Abschnitt „Funktionsanalyse von Proteinbindetaschen mit *PocketShapelets*" dargestellt (Kapitel 3.6.2).

2.3.3 Analyse von Oberflächenpotentialen

Als Ergänzung zu den bisher vorgestellten Techniken, die der Beschreibung und dem Vergleich von Bindetaschenformen dienen, wurden in dieser Arbeit Methoden angewendet, die detaillierte Untersuchungen der in einer Bindetasche vorherrschenden

molekularen Eigenschaften ermöglichen. Dazu sind in *PocketomePicker* Verfahren zur Bestimmung des elektrostatischen Potentials und Abschätzung lokaler Lipophilie implementiert. Die Potentiale werden dabei an den Mittelpunkten der Shapelets gemessen, die somit neben der Form auch Eigenschaften von Bindetaschen beschreiben.

2.3.3.1 Berechnung des elektrostatischen Potentials

Zur Bestimmung des elektrostatischen Potentials $\varphi(\vec{r})$ wurde in dieser Arbeit ein Ansatz gewählt, der auf dem **Coulombschen Gesetz** (Leach 2001; Klapper *et al*., 1986) beruht (Gleichung 8). Das elektrostatische Potential wird unter Berücksichtigung der Partialladungen q_i der umliegenden N Atome ermittelt, sofern der Abstand d_i zwischen Atom und Shapeletmittelpunkt einen zuvor definierten Schwellenwert d_0 nicht überschreitet (Gleichung 8).

$$\varphi(\vec{r}) = \begin{cases} \frac{1}{4\pi\varepsilon} \cdot \sum_{i=1}^{N} \frac{q_i}{d_i}, \text{wenn } d_i \le d_0. \\ 0 \qquad\qquad\qquad\qquad , \text{wenn } d_i > d_0. \end{cases} \tag{8}$$

Die Partialladungen werden dabei in *PocketomePicker* mit PDB2PQR (Dolinsky *et al.*, 2004; http://pdb2pqr.sourceforge.net/) unter Verwendung der CHARMM Atomtypen berechnet (Brooks *et al*., 1983; MacKerell *et al*., 1998). Ferner wurde eine Dielektrizitätskonstante $\varepsilon = 1$ verwendet. Dieser Wert entspricht der Kenngröße des Vakuums und wurde gewählt, da ein verlässlicher Wert für die Permissivität an verschiedenartigen Proteinoberflächen nicht berechnet werden kann. Die berechneten elektrostatischen Potentiale wurden für eine geeignete Skalierung der in *PocketShapelets* verwendeten Scores mit einem Faktor 20 multipliziert.

2.3.3.2 Fragmentbasierte Bestimmung der Lipophilie

Der **hydrophobe Effekt** beschreibt die anziehenden Wechselwirkungen organischer Moleküle oder nicht-polarer Gruppen in wässriger Umgebung (Tanford, 1978). Zur Quantifizierung dieser Anziehungskraft dient die Änderung der freien Bindungsenthalpie (ΔG) beim Übergang einer organischen Substanz von einer wässrigen in eine

unpolare Phase. Da es kein einfaches Verfahren gibt, um $\Delta G_{transfer}$ abzuschätzen, werden empirische Betrachtungen zur Beschreibung relativer Lipophilie herangezogen (Heiden *et al.*, 1993). Als Grundlage dienen dabei die Verteilungskoeffizienten (log *P*, engl. *partition coefficient*) von Referenzmolekülen gemessen in polar-unpolaren Zweiphasensystemen[8].

In vergleichenden Studien konnte gezeigt werden, dass die Löslichkeit eines Moleküls durch die Summe der Lipophiliewerte seiner Fragmente f_i angenähert werden kann (Gleichung 9; Fujita *et al.*, 1964).

$$\log P = \sum_i f_i \,. \tag{9}$$

Verschiedene Methoden wurden entwickelt, die eine genauere Abschätzung der Lipophilie durch Betrachtung der Wechselwirkungen zwischen den Fragmenten eines Moleküls ermöglichen (Rekker & de Kort, 1979; Hansch & Leo, 1979).

Zur Charakterisierung lipophiler und hydrophiler Bereiche in Proteinbindetaschen wurde in dieser Arbeit ein solcher fragmentbasierter Ansatz gewählt, der den einzelnen Atomen eines Proteins Lipophiliewerte zuweist (Ghose & Crippen, 1986; Viswanadhan *et al.*, 1989; Ghose *et al.*, 1998). Analog zum elektrischen Potential wurde aus den Lipophiliewerten der atomaren Fragmente f_i ein **lipophiles/hydrophobes Potential** L_{HM} berechnet. Dieses wird an den Flächenmittelpunkten der Shapelets bestimmt, die die Form der Tasche beschreiben. Zum Potential tragen hierbei nur diejenigen Atome bei, deren Abstände d_i innerhalb eines Schwellenwerts *c* (engl. *cutoff*) zum Shapelet liegen. Zur Beschreibung der Lipophile der atomaren Fragmente wurden die in Ghose *et al.* 1998 vorgestellten Werte verwendet. Ferner wurde eine von Heiden und Mitarbeitern vorgeschlagene Abstandsfunktion g_{HM} implementiert (Heiden *et al.*, 1993), die den nicht-linearen Einfluss des Potentials L_{HM} bezogen auf den räumlichen Abstand berücksichtigt (Gleichung 10).

[8] Hierbei handelt es sich meist um Oktanol-Wasser-Phasensysteme für die Verteilungskoeffizienten für eine Vielzahl von Molekülen bekannt sind.

$$L_{HM} = \frac{\sum_{i=1}^{N} f_i \cdot g_{HM}(d_i, c, \delta)}{\sum_{i=1}^{N} g_{HM}(d_i, c, \delta)}. \qquad (10)$$

Hierbei ist δ ein Parameter, der die Steigung von g_{HM} beeinflusst. Das lipophile/hydrophobe Potential berechnet sich demnach wie in Gleichung 11 beschrieben.

$$g_{HM}(d_i, c, \delta) = \frac{e^{-2c/\delta} + 1}{e^{2(d_i - c)/\delta} + 1}. \qquad (11)$$

Die Funktion g_{HM} wurde als sigmoidale Fermifunktion (Reif, 1985) gewählt, die finite Werte für $d_i < c$ liefert und danach stark abfällt (Abbildung 23)

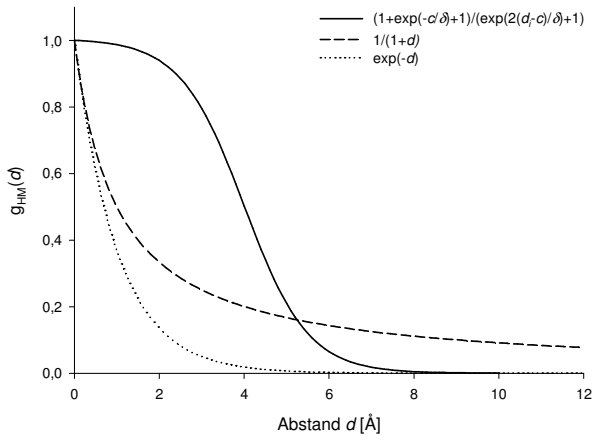


Abbildung 23: Abstandsfunktionen g_{HM} zur Gewichtung des räumlichen Einflusses des hydrophilen/lipophilen Potentials. In dieser Arbeit wurde g_{HM} als Fermifunktion implementiert (durchgezogene Linie). Andere gebräuchliche Abstandsfunktionen sind eingezeichnet. Die Funktionen sind für die Parameter c=4,0 und δ=1.5 dargestellt.

2.3.4 Analyse von Bindetaschentopologien mit PocketGraph

Die Darstellung niedermolekularer Proteinliganden als Molekülgraphen ermöglicht eine effiziente Analyse ihrer strukturellen und physikochemischen Eigenschaften durch computergestützte Verfahren. Im Unterschied dazu gibt es zur Untersuchung des Aufbaus und der Eigenschaftsverteilung auf Bindetaschenvolumen keine standardisierte

Repräsentation. In dieser Arbeit wurde die in *PocketPicker* (Weisel *et al.*, 2007) verwirklichte Darstellung von Bindestellen als Grundlage weiterer Analysen verwendet. Nachfolgend soll eine in *PocketomePicker* verwirklichte Technik namens ***PocketGraph*** vorgestellt werden, die eine automatische Extraktion der Topologie von Bindevolumen ermöglicht.

2.3.4.1 Automatische Topologieerkennung mit Wachsendem Neuronalem Gas

Für eine reduzierte Beschreibung von Bindetaschenvolumen wurde in *PocketomePicker* eine Technik implementiert, die als **wachsendes neuronales Gas** (engl. *Growing Neural Gas*, GNG) bekannt ist (Fritzke, 1994; Fritzke, 1995). Diese Methode realisiert ein unüberwacht lernendes ANN, welches sich im Vergleich zu SOMs und neuronalen Gas Netzwerken (Neural Gas Network, NGN; Martinetz & Schulten, 1991) dahingehend unterscheidet, dass es ohne vorherige Festlegung einer Netzgröße (Anzahl künstlicher Neuronen) auskommt. Das GNG ermöglicht eine adaptive Anpassung an eine Datenmenge durch das Einfügen und Löschen von Neuronen während der Lernphase. Als Abbruchkriterium dient ein Quantisierungsfehler (*QF*), der durch Adaption der Neuronen minimiert werden soll. In *PocketomePicker* wurde ein GNG implementiert, welches die Extraktion der Topologie einer Bindetasche zum Ziel hat. Nachfolgend ist der verwendete GNG-Ansatz als Pseudocode zusammengefasst:

GNG – Notation

Variablen:

A **Netzwerk** aus N Neuronen: $A = \{c_1, ..., c_n\}$.

c_i **Neuronen** deren räumliche Lage durch Referenzvektoren $w_c \in \Re^n$ definiert ist.

C **Nachbarschaftsverbindungen** zwischen den Neuronen des Netzwerks: $C \subset A \times A$.

ξ **Eingangssignal**, bestimmt durch eine **probabilistische Funktion** $p(\xi)$.

$s(\xi)$ **Siegerneuron** mit geringstem Abstand zum Eingangssignal: $\arg\min_{c \in A} \|\xi - w_c\|$.

E_c **Fehlervariable** eines Neurons.

$\|\cdot\|$ **Euklidischer Abstand**.

Konstanten:

λ **Zeitschritt** zum Einfügen neuer Neurone.
ε_b **Sprungweite** des Siegerneurons zum Eingangssignal (als Teil des Abstandes).
ε_n **Sprungweite** der topologischen Nachbarn des Siegerneurons zum Eingangssignal.
α, β **Skalierungsfaktoren** der Fehlervariablen.
a **Alter** eine Kante.
a_{max} **Maximales Alter** einer Kante.

GNG – Beschreibung des Algorithmus

Initialisation:

- Initialisiere ein Netzwerk A mit zwei Neuronen: $A=\{c_1,c_2\}$ mit zufällig gewählten Referenzvektoren gemäß $p(\xi)$.

Iteration:

- Wähle ein zufälliges Eingangssignal (Datenpunkt) ξ gemäß $p(\xi)$.
- Bestimme das Siegerneuron und das zweitnächste Neuron (s_1, $s_2 \in A$):
  $s_1 = _{c\in A}\left\|\xi - w_c\right\|$, $s_2 = _{c\in A\setminus\{s_1\}}\left\|\xi - w_c\right\|$.
- Verbinde s_1 und s_2 durch eine Kante, falls noch keine Verbindung besteht: $C = C \cup \{(s_1, s_2)\}$.
- Setze das Alter der Kante s_1-s_2 auf null: $a_{(s_1,s_2)} = 0$.
- Erhöhe die Fehlervariable des Siegerneurons um seinen quadratischen Abstand zum Eingangssignal:
  $\Delta E_{s_1} = \left\|\xi - w_{s_1}\right\|^2$.
- Aktualisiere die Referenzvektoren des Siegerneurons und seiner direkten topologischen Nachbarn:
  $\Delta w_{s_1} = \varepsilon_b(\xi - w_{s_1})$
  $\Delta w_i = \varepsilon_n(\xi - w_i) \qquad (\forall i \in N_{s_1})$,
  mit N_{s_1} als Menge der direkten topologischen Nachbarn von s_1.
- Erhöhe das Alter aller Kanten, die von s_1 ausgehen:
  $a_{(s_1,i)} = a_{(s_1,i)} + 1 \qquad (\forall i \in N_{s_1})$.
- Entferne die Kanten mit Alter größer als a_{max}.
- Entferne die Neuronen, die keine eingehenden Kanten mehr besitzen.
- Füge ein neues Neuron ein, wenn die Zahl der gezogenen Eingangssignale ein Vielfaches von λ ist:
  - Bestimme dazu das Neuron q mit dem bisher größten Fehler:
    $q = \arg\max_{c\in A} E_c$.
  - Bestimme unter den Nachbarn von q das Neuron f mit größtem Fehler:
    $f = \arg\max_{c\in N_q} E_c$.

- Interpoliere den Referenzvektor des neuen Neurons *r* aus den Koordinaten von *q* and *f* und füge es in das Netzwerk *A* ein:
  $A = A \cup \{r\}, \quad w_r = (w_q + w_f)/2$.
  Verbinde das Neuron *r* über Kanten mit den Neuronen *q* and *f* und entferne die Kante zwischen *q* and *f*:
  $C = C \cup \{(r,q),(r,f)\}, \qquad C = C \setminus \{(q,f)\}$.
- Verringere die Fehlervariable von *q* and *f* um den Faktor α:
  $\Delta E_q = -\alpha E_q, \quad \Delta E_f = -\alpha E_f$.
- Interpoliere die Fehlervariable von *r* aus *q* und *f*:
  $E_r = (E_q + E_f)/2$.

- Aktualisiere die Fehlervariable aller Neurone:
  $\Delta E_c = -\beta E_c \quad (\forall c \in A)$.

Termination:

- Beende die Rekursion, wenn ein zuvor gewählter *QF* erreicht wurde.

Der *QF*, der als Approximationsfehler des GNG Trainings verwendet wird, wird als **mittlerer quadratischer Quantisierungsfehler** (engl. *Mean Squared Quantization Error*, MSQE) ausgedrückt und beschreibt den durchschnittlichen quadratischen Abstand eines Datenpunktes zum nächstgelegen Neuron des GNG. In dieser Arbeit wurden die von *PocketPicker* errechneten Taschenpunkte, die eine potentielle Bindetasche anzeigen als Datenpunkte verwendet. Eine reduzierte Beschreibung der so charakterisierten Bindetaschenformen gelingt durch Anwendung eines GNG Ansatzes auf diese Taschenpunkte. Die Interpretation der Neurone und Verbindungen eines so trainierten GNG als Knoten und Kanten eines Graphen erlaubt eine deutlich vereinfachte Darstellung von Taschengeometrien. Diese Methode wurde in dieser Arbeit als ***PocketGraph*** entwickelt und in *PocketomePicker* integriert.

Um lineare, ringfreie Repräsentationen zur erhalten, wurde eine Variante einer Technik zur Erkennung des **kleinsten Satzes kleinster Ringe** (engl. *Smallest Set of Smallest Rings*, SSSR) implementiert (Balducci & Pearlman, 1994). Diese Methode realisiert eine **Tiefensuche** auf ungerichteten, gewichteten Graphen (Tarjan, 1972) mit anschließender Identifikation von kürzesten Wegen auf dem Tiefensuchgraph (Abbildung 24).

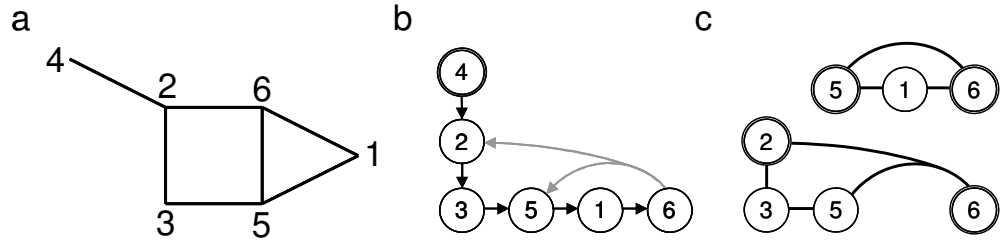


Abbildung 24: Identifikation der kleinsten Ringe auf einem Graphen. a) Ein Graph mit zufälliger Benennung der Knoten. Dieser Graph enthält drei Ringe (2,3,5,6; 1,5,6; 2,3,5,1,6). b) Resultierender Tiefensuchegraph einer in Knoten 4 gestarteten Tiefensuche (eingekreist) mit Rückwärtskanten (grau), die Zyklen anzeigen. c) Kleinste Ringe als kürzeste Wege zwischen Start- und Endpunkten (beide eingekreist) der Rückwärtskanten mit Abkürzungen über Rückwärtskanten.

Im Anschluss an eine Approximation einer Tasche durch ein GNG werden durch Anwendung der SSSR-Technik die kleinsten Ringe im resultierenden GNG-Graphen identifiziert und die jeweils älteste Kante eines Ringes entfernt. Dies erlaubt eine Darstellung der Topologie einer Tasche durch linearisierte Graphen (Abbildung 25).

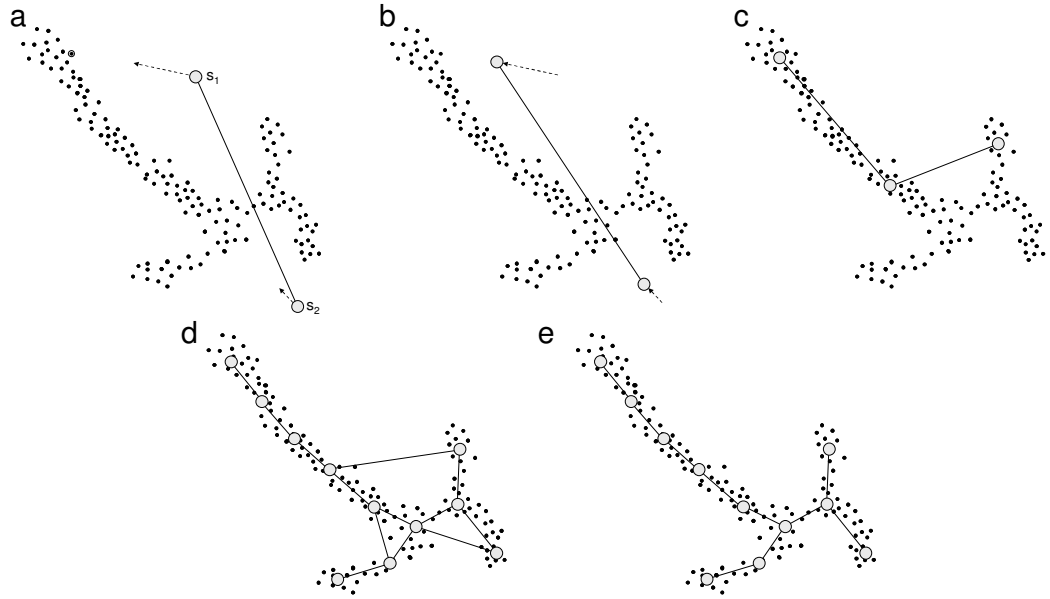


Abbildung 25: Schematischer Ablauf einer Approximation mit GNG. a) Das Netzwerk wird mit zwei Neuronen initialisiert. Ein Datenpunkt wird zufällig ausgewählt (eingekreist) und das Siegerneuron s_1, sowie das zweitnächste Neuron s_2 bestimmt. b) Die Referenzvektoren der Neurone werden angepasst (mit verringerter Lernrate für s_2). c) Weitere Neurone werden in diskreten Zeitschritten eingefügt. d) Das Training terminiert, wenn ein geforderter QF erreicht wurde. e) Durch eine SSSR Technik werden Ringe identifiziert und die ältesten Kanten eines Ringes entfernt.

Die zufällige Auswahl von Taschenpunkten für das Trainieren der Neuronen bedeutet die Einführung eines Nichtdeterminismus in *PocketGraph*. Dieser ist insofern

unerwünscht, als dass er verschiedene Graphdarstellungen für wiederholte Anwendungen des Algorithmus auf derselben Tasche generiert (Abbildung 26).

Zur Unterdrückung des Nichtdeterminismus bei der zufälligen Auswahl der Taschenpunkte wurde eine Technik in *PocketGraph* implementiert, die eine deterministische und balancierte Auswahl simuliert. Die Erzeugung von zufälligen Zahlenfolgen erfolgt durch deterministische Algorithmen, die in den Bibliotheken von Programmiersprachen verfügbar sind. So nutzt die in Java™ integrierte Klasse java.util.Random eine Zahl vom Typ Long Integer als sogenannten „Keim“ (engl. *Seed*), um Folgen von Zufallszahlen für ein gewünschtes Intervall zu erzeugen. Aus dem Keim entspringen Nachfahren von Zahlen, die nicht zufällig sind, sondern einem mathematischen Verfahren gehorchen (Lineare Kongruenz Methode; Knuth, 1998). Die Klasse java.util.Random erzeugt den Seed aus der aktuellen Uhrzeit (ausgedrückt in Millisekunden), so dass für zwei zeitliche folgende Eingaben verschiedene Zahlenreihen generiert werden. Die Klasse bietet jedoch die Möglichkeit, den Seed fest einzustellen. Diese Option wird in *PocketGraph* genutzt, um eine balancierte Auswahl der Taschenpunkte zu simulieren, die ein gleiches Ergebnis für wiederholte Anwendungen garantiert.

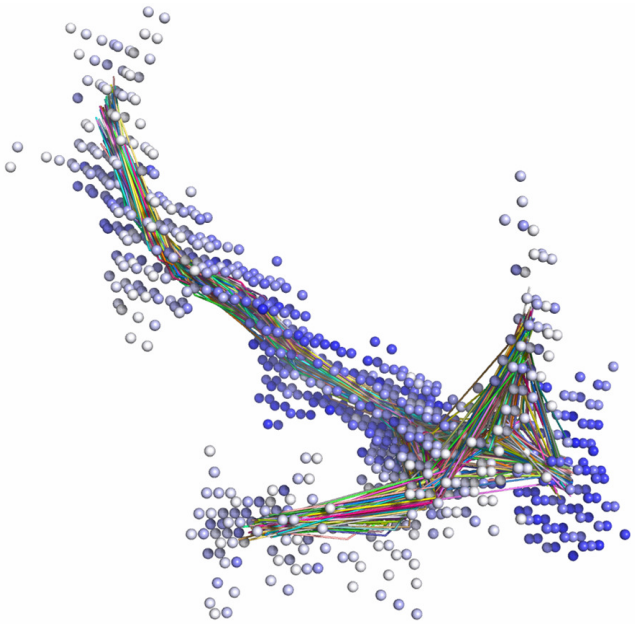


Abbildung 26: Nichtdeterminismus bei der Berechnung von PocketGraphen für die ligandenfreie Bindetasche einer Kristallstruktur der Aldose Reduktase (PDB: 1iei). Die zufällige Auswahl der Taschenpunkte für das Training des GNG produziert bei wiederholter Anwendung verschiedene PocketGraphen mit unterschiedlichen Topologien. Gezeigt sind die ermittelten Graphdarstellungen von 100 Berechnungen auf der gezeigten Tasche.

Die deterministisch erzeugte Graphdarstellung eines Bindevolumens in *PocketPicker*-Darstellung durch linearisierte PocketGraphen erlaubt eine reduzierte Repräsentation von Bindetaschentopologien. Diese Technik kann u.a. für die Zerlegung einer Bindestelle in ihre Subtaschen genutzt werden. Dieser Ansatz ist nachfolgend dargestellt.

2.3.4.2 Zerlegung von putativen Bindestellen in Subtaschen

Die Extraktion von Taschentopologien durch GNG Netzwerke erlaubt eine reduzierte Beschreibung des strukturellen Aufbaus von Bindstellen. So charakterisieren endständige Positionen von PocketGraphen Subtaschen, die die Verankerung potentieller Liganden im Bindevolumen ermöglichen. Die Zerlegung von Taschen in disjunkte Kompartimente erfolgt an den Schnittstellen eines PocketGraphen von denen mehrere endständige Abschnitte des Graphen abgehen.

Die Zerlegung von Bindestellen durch Betrachtung von PocketGraph-Topologien erlaubt eine Dekomposition der mit *PocketPicker* berechneten Taschenvolumen. Darüber hinaus ermöglicht diese Technik eine Repräsentation der Form und physikochemischen Eigenschaften von Subtaschen durch *PocketShapelets*, die anhand der GNG-Netzwerken gruppiert werden (Abbildung 27).

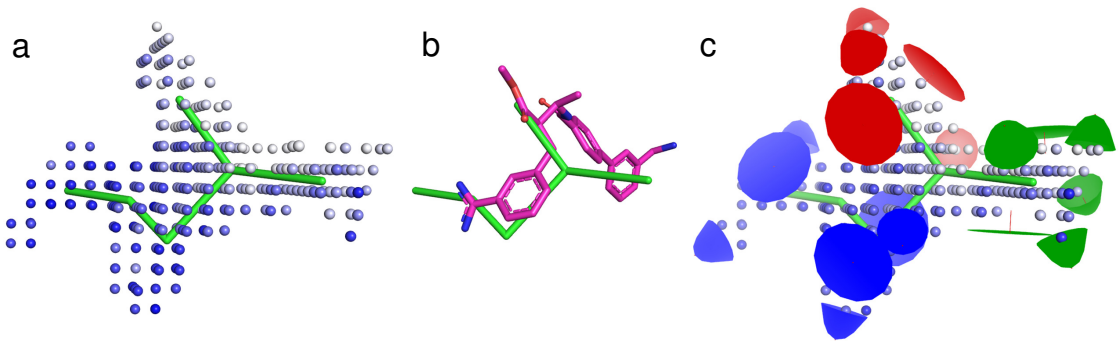


Abbildung 27: Zerlegung der größten Bindetasche von humanem Gerinnungsfaktor Xa (PDB: 1ezq). a) Erkennung der Topologie der mit *PocketPicker* extrahierten Tasche durch *PocketGraph* (grün). b) Reduzierte Darstellung des Taschenaufbaus durch *PocketGraph* (grün). Diese Repräsentation charakterisiert Subtaschen, die vom Inhibitor RPR128515 (magenta, PDB: rpr) besetzt werden. c) Die Topologieerkennung durch *PocketGraph* ermöglicht die Zuordnung der formbeschreibenden *PocketShapelets* zu den drei Subtaschen (blau, rot, grün) der Bindestelle.

3 Ergebnisse und Diskussion

Ziel der vorliegenden Arbeit war die Analyse von Form, Eigenschaften und Druggability von Proteinbindetaschen. Als Grundlage der in dieser Arbeit erstellten Algorithmen zum Vergleich von Bindetaschen und ihren Eigenschaften diente das Programm *PocketPicker* (Weisel, 2006). Die neu entwickelten Methoden wurden in dem Programmpaket *PocketomePicker* zusammengefasst, welches die folgenden Module beinhaltet:

- PocketShapelets – strukturelle Überlagerung von Bindetaschen zum Vergleich von elektrostatischen und lipophilen Oberflächenpotentialen.
- PocketGraph – Anwendung wachsender neuronaler Gase (GNG) zur Extraktion der Topologie von Proteinbindetaschen.

Anwendungen zu den in dieser Arbeit entwickelten Methoden werden nachfolgend vorgestellt und Ergebnisse zum Ähnlichkeitsvergleich von Bindetaschenformen und ihren Eigenschaften analysiert. Ferner sind nachfolgend Anwendungsbeispiele für die Vorhersage der Druggability und Funktion von Bindetaschen dargestellt. Zudem soll eine von *PocketPicker* abgeleitete Anwendung für ein erfolgreiches rezeptorbasiertes Ligandendesign vorgestellt werden.

Das in *PocketPicker* verwirklichte gitterbasierte Konzept zur Vorhersage potentieller Bindestellen wirkstoffartiger Liganden bildet die Grundlage der in dieser Arbeit entwickelten Methoden. Nachfolgend sollen daher die Genauigkeit der Suchroutine und die Eignung der *ShapeDeskriptoren* für die Analyse von Proteinbindetaschen erläutert und diskutiert werden.

3.1 Vorhersage und Analyse von Bindetaschen mit PocketPicker

In diesem Abschnitt soll die Leistungsfähigkeit von PocketPicker dargestellt werden, mit der das Programm Bindetaschen für Liganden auf Proteinkristallstrukturen korrekt erkennt. Für diesen Zweck wurde der Algorithmus auf einem repräsentativen Datensatz

getestet und Ergebnisse mit denen anderer etablierter Verfahren zur Identifikation von Bindetaschen verglichen.

3.1.1 Evaluation der Taschenvorhersage mit PocketPicker

Um die Güte der in *PocketPicker* implementierten Suchroutine zur Identifikation von Bindetaschen zu bewerten, wurde eine Evaluationsmethode gewählt, die zuvor in bereits in anderen Studien verwendet wurde (Brady & Stouten, 2000; Huang & Schröder, 2006). Folglich wird eine Vorhersage als erfolgreich bezeichnet, wenn das geometrische Zentrum einer berechneten Tasche in einem Abstand von höchstens 4 Å zu einem beliebigen Atom des gebundenen Liganden liegt.

Die Qualität der Vorhersageroutine von *PocketPicker* wurde auf einem Datensatz getestet, der 48 Protein-Liganden-Komplexe, sowie deren entsprechenden *apo*-Kristallstrukturen umfasst. Diese Datensammlung (siehe Anhang, Tabelle A1) wurde zuvor in einer Arbeit verwendet, die die Leistungsfähigkeit verschiedener etablierter Vorhersageverfahren untersucht (Huang & Schröder, 2006). Die Ergebnisse aus diesem Vergleichstest wurden als Referenz zur Evaluation der Taschenvorhersage von *PocketPicker* verwendet.

Die hier betrachteten Methoden bedienen sich geometrischer Verfahren zur Vorhersage und beschreiben die größte Vertiefung auf der Proteinoberfläche als wahrscheinlichste Bindetasche. Daher werden Ergebnisse der Taschenvorhersage in zwei Kategorien eingeteilt. Korrekte Charakterisierungen von Bindetaschen sollen nachfolgend „Top1-Treffer“ genannt werden, wohingegen „Top3-Treffer“ Vorhersagen beschreibt, bei denen der betrachtete Ligand in einer der drei größten gefundenen Taschen vorliegt. Eine vollständige Darstellung der Vorhersage mit *PocketPicker* ist in Tabelle 1 gegeben.

Tabelle 1: Erfolgsrate von *PocketPicker* auf 48 Komplexen und ihren *apo*-Strukturen. Als Treffer werden die Taschen aufgeführt, deren geometrisches Zentrum weniger als 4 Å vom nächsten Atom des Referenzliganden entfernt sind (D_{Ligand}). Klammern kennzeichnen Treffer, die den Schwellenwert von 4 Å übersteigen.

Komplex	Treffer	D_{Ligand} [Å]	*apo*-Struktur	Treffer	D_{Ligand} [Å]
1bid	1	2,0	3tms	1	3,9
1cdo	1	1,9	8adh	1	1,3
1dwd	1	1,5	1hxf	1	0,7

1fbp	1	0,7	2fbp	6	0,7
1gca	1	1,5	1gcg	1	1,6
1hew	1	0,9	1hel	1	0,9
1hyt	1	0,5	1npc	1	0,8
1inc	1	0,6	1esa	1	3,9
1rbp	1	0,6	1brq	1	0,1
1rob	1	0,2	8rat	1	1,1
1stp	1	1,2	1swb	1	1,2
1ulb	1	1,4	1ula	1	1,6
2ifb	1	1,3	1ifb	1	1,3
3ptb	1	0,9	3ptn	2	0,7
2ypi	4	0,8	1ypi	3	2,8
4dfr	(1)	8,1	5dfr	1	1,8
4phv	1	0,7	3phv	(2)	4,2
5cna	7	0,8	2ctv	8	0,5
7cpa	1	0,8	5cpa	1	1,1
1a6w	3	1,3	1a6u	4	1,3
1acj	(1)	4,4	1qif	(1)	4,4
1apu	1	1,1	3app	1	0,8
1blh	1	1,7	1djb	1	1,6
1byb	1	2,3	1bya	1	2,2
1hfc	1	0,8	1cge	1	0,7
1ida	1	0,8	1hsi	1	1,8
1igj	2	0,3	1a4j	1	0,7
1imb	3	2,0	1ime	1	2,0
1ivd	2	0,6	1nna	1	1,0
1mrg	(1)	4,4	1ahc	1	3,4
1mtw	2	0,7	2tga	4	1,2
1okm	2	1,9	4ca2	1	2,1
1pdz	1	0,9	1pdy	1	1,8
1phd	1	1,1	1phc	1	1,4
1pso	1	0,7	1psn	1	1,1
1qpe	1	0,8	3lck	1	1,6
1rne	1	0,6	1bbs	1	0,7
1snc	1	2,8	1stn	2	1,2
1srf	1	0,9	1pts	2	0,3
2ctc	1	0,6	2ctb	1	0,9
2h4n	1	1,5	2cba	1	1,4
2pk4	2	0,4	1krn	2	1,0
2sim	2	0,9	2sil	2	1,0
2tmn	1	0,5	1l3f	1	0,8
3gch	1	0,5	1chg	2	1,9
3mth	1	1,0	6ins	2	1,4
5p2p	1	0,9	3p2p	1	1,0
6rsa	1	0,9	7rat	1	0,8

Vor Beginn der Berechnungen wurden Wasserstoffatome mit PyMOL (DeLano, 2002) an die PDB-Dateien angefügt und die *apo*-Strukturen mit dem *align*-Befehl in PyMOL strukturell überlagert, um eine Ausrichtung auf die Komplexe zu erreichen. Die Erfolgsraten der Taschenvorhersage mit *PocketPicker* sind in Tabelle 2 im Vergleich mit anderen Vorhersagemethoden zusammengefasst.

Tabelle 2: Vergleich der Vorhersagequalität verschiedener Methoden auf einem Datensatz von 48 Komplexen und ihren entsprechenden *apo*-Strukturen.

	Top1-Treffer [%]		**Top3-Treffer [%]**	
	Komplex	***apo*-Struktur**	**Komplex**	***apo*-Struktur**
PocketPicker)[1]	73	71	90	88
LIGSITE^{cs})[2]	69	60	87	77
LIGSITE)[3]	69	58	87	75
CAST)[4]	67	58	83	75
PASS)[5]	63	60	81	71
SURFNET)[6]	54	52	78	75
LIGSITE^{csc})[2,7]	79	71	-	-

[1]Weisel *et al*., 2007. [2]Huang & Schröder, 2006. [3]Hendlich *et al*. 1997. [4]Liang *et al*. 1998, Binkowski *et al.*, 2003. [5]Brady & Stouten, 2000. [6]Laskowski, 1995. [7]Der Vergleich der Erfolgsraten von LIGSITE^{csc} mit denen der übrigen Methoden ist nur beschränkt möglich, da LIGSITE^{csc} ein nicht ausschließlich geometrisches Verfahren ist.

PocketPicker übertrifft die Vorhersagen von CAST (Liang *et al.*, 1998; Binkowski *et al.*, 2003), PASS (Brady & Stouten, 2000) und SURFNET (Laskowski, 1995) deutlich und zeigt Vorteile gegenüber den Methoden LIGSITE (Hendlich *et al.*, 1997) und LIGSITE^{cs} (Huang & Schröder, 2006). Die Erfolgsraten von LIGSITE^{csc} sind nur der Vollständigkeit halber aufgeführt und nur bedingt mit den übrigen Methoden vergleichbar. LIGSITE^{csc} stellt eine Weiterentwicklung des ursprünglichen LIGSITE Algorithmus dar und nutzt neben geometrischen Verfahren Information über die Konserviertheit der Reste einer Bindetasche. So findet unter den drei größten vorhergesagten Taschen eine Sequenzanalyse der Reste statt, die die Oberfläche der Tasche bilden. Als wahrscheinlichste Bindestelle wird dann die Tasche interpretiert, die größtmögliche Konserviertheit zu einem Eintrag der ConsSurf-HSSP Datenbank (Glaser *et al.*, 2005) besitzt. Dieses Wissen bedeutet einen Vorteil gegenüber den anderen Methoden des Vergleichs, beschränkt den Einsatz von LIGSITE^{csc} aber auf Proteine, für die Sequenzhomologe in der ConsSurf-HSSP Datenbank bekannt sind.

Die korrekte Vorhersage von Bindetaschen auf *apo*-Strukturen ist von besonderem Interesse bei geometrischen Suchverfahren, da die Abwesenheit eines Taschen-

induzierenden Liganden die Berechnungen erschweren kann. Dieser Effekt wird in der Literatur als *induced fit* beschrieben (Koshland, 1994) und stört die Taschenvorhersage in mehreren Proteinen des Datensatzes (Abbildung 28).

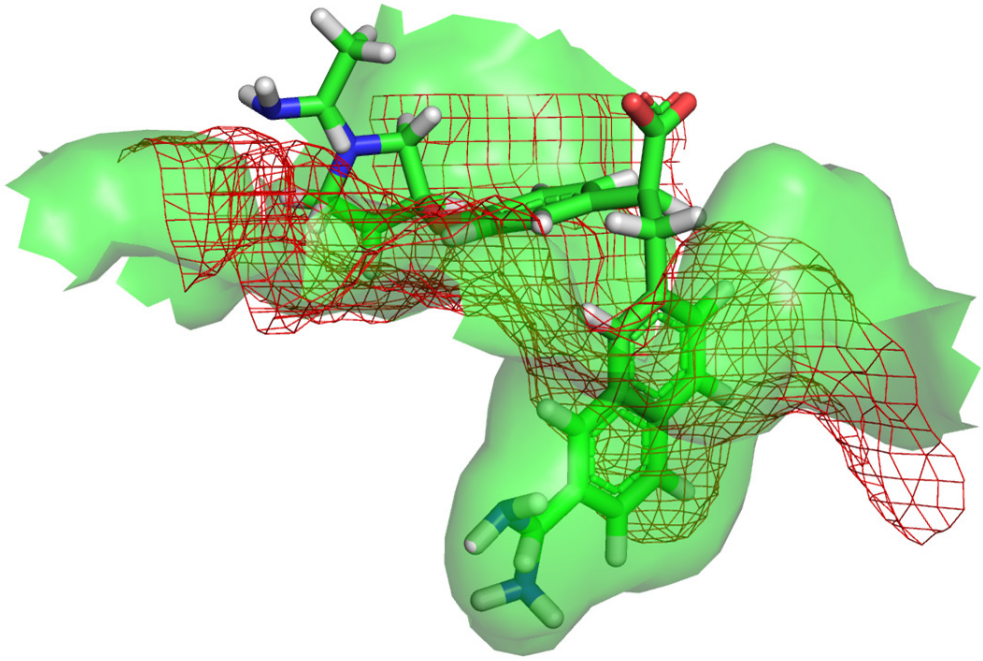


Abbildung 28: *Induced-Fit* Phänomen in Trypsin. Die Abbildung zeigt einen Ausschnitt der Molekülоberflächen aus einem rigiden Strukturalignment des Komplexes (grün, PDB: 1mtw) mit gebundenem Liganden (PDB: dx9) und der korrespondierenden *apo*-Struktur (Gitterdarstellung, rot, PDB: 2tga). Der Ligand bewirkt eine deutliche Vertiefung der Tasche, die in der *apo*-Struktur nur schwach ausgeprägt ist.

Während bei den meisten anderen Methoden des Vergleichs ein deutlicher Abfall des Vorhersageerfolges auf *apo*-Strukturen zu beobachten ist, zeigt *PocketPicker* hier eine erkennbare Robustheit. Dies mag an der vergleichsweise hohen Auflösung liegen, die durch die Verwendung von 30 Suchstrahlen erreicht wird. So kann *PocketPicker* auch in flachen Taschen Atome in Bereichen aufspüren, die von anderen Methoden möglicherweise nicht ausreichend erfasst werden. Dieser Effekt konnte bei der Entwicklung von LIGSITE beobachtet werden (Hendlich *et al.*, 1997). Diese Methode ist eine direkte Weiterentwicklung des Programms POCKET (Levitt & Banaszak, 1992) und erreicht eine Verbesserung der Vorhersage durch die Betrachtung einer vergrößerten Zahl an Raumrichtungen.

Die Genauigkeit der Taschenvorhersage ist abhängig von der Beschaffenheit der Proteinoberfläche. So werden fein abgegrenzte Taschen oder vollständig vergrabene

Bereiche besser erkannt als flache Bindestellen auf der Proteinoberfläche (Abbildung 29).

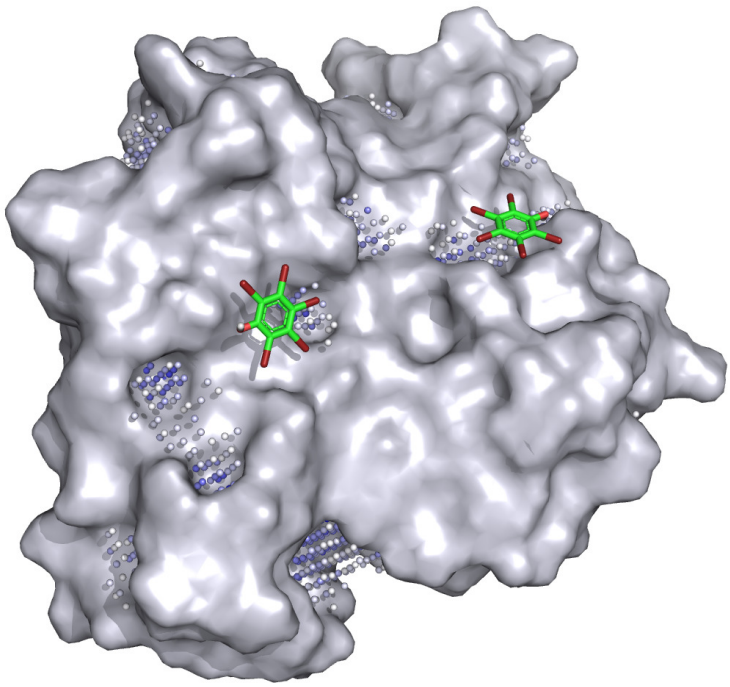


Abbildung 29: Taschenvorhersage mit *PocketPicker* für Transthyretin (PDB: 1e4h) mit gebundenem Bromphenol (PDB: pbr). Die Liganden binden an nicht vergrabene Teile der Oberfläche und können daher von geometrischen Verfahren nicht erkannt werden.

Die Erkennung von Bindestellen in flachen oder nicht vergrabenen Teilen der Oberfläche gestaltet sich schwierig für geometrische Verfahren. Ein alternativer Ansatz wäre die Vorhersage von Bindestellen unter Berücksichtigung physikochemischer Eigenschaften. So verzichtet das Programm Q-SiteFinder (Laurie & Jackson, 2005) auf geometrische Betrachtungen und bestimmt potentielle Bindetaschen als Bereiche der Oberfläche, die energetisch günstige Bereiche für die Bindung eines niedermolekularen Liganden bieten. Eine Erweiterung der geometrischen Vorgehensweise von *PocketPicker* unter Berücksichtigung molekularer Eigenschaften bietet daher einen vielversprechenden Ansatz für eine weitere Steigerung der Vorhersage.

Manuelle Inspektionen zeigen, dass die Suchroutine von *PocketPicker* auf kleineren monomeren Proteinen (< 5000 Atome) die besten Erfolgsraten erreicht (Daten nicht gezeigt). Größere multimere Proteine, die aus identischen Untereinheiten zusammengesetzt sind, erschweren hingegen die Vorhersage. So entstehen an den Kontaktflächen der Protein-Protein-Komplexe oftmals Hohlräume, die durch das

geometrische Suchverfahren fälschlicherweise als Bindetasche interpretiert werden können (Abbildung 30).

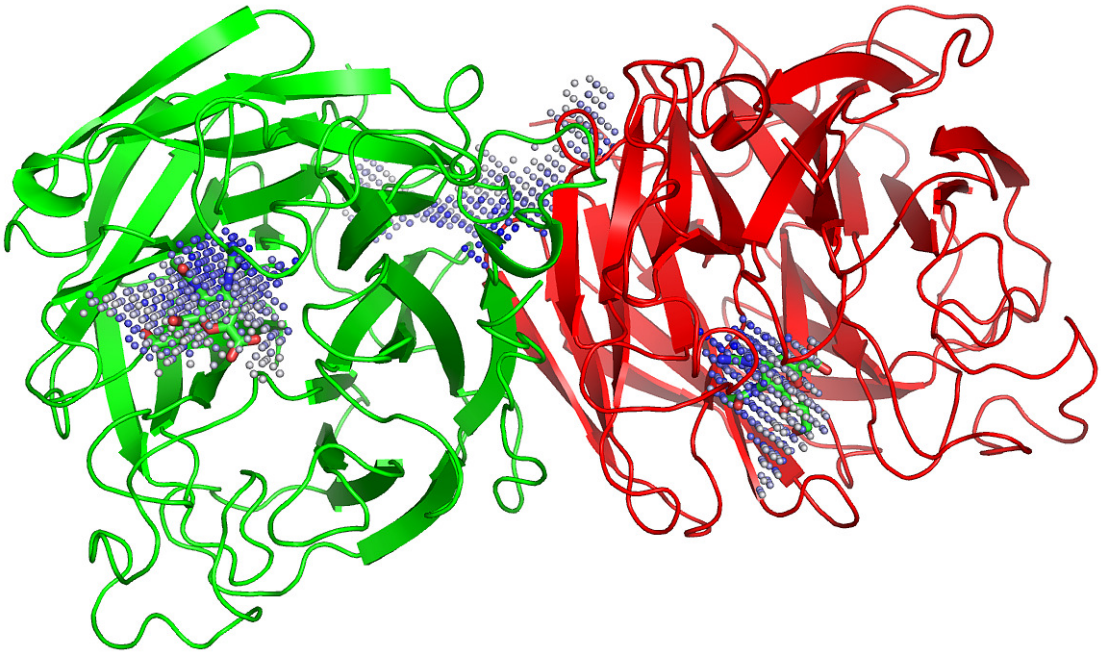


Abbildung 30: Vorhersage von Bindetaschen mit *PocketPicker* für Influenza-Virus Neuramidase (PDB: 1a4g). Eine Lücke zwischen den Ketten A und B wird als größte Tasche identifiziert und daher irrtümlich als aktives Zentrum interpretiert. Die tatsächlichen Bindestellen der Liganden *Zanamivir* (PDB: zmr) werden als zweit- und drittgrößte Tasche gefunden.

Weiterhin gestaltet sich die Berechnung von Bindetaschen auf großen Proteinen (> 8000 Atome) schwierig, da disjunkte Taschen durch unter der Oberfläche verlaufende „Tunnel“ verbunden sein können. Diese Beobachtung wird durch Ergebnisse einer früheren Arbeit bestätigt, die einen umgekehrt proportionalen Zusammenhang zwischen dem Molekulargewicht und der Dichte von Proteinen feststellt (Fischer *et al*., 2004). So konnte in theoretischen wie in experimentellen Versuchen gezeigt werden, dass größere Proteine eine geringere Packungsdichte aufweisen als kleinere Proteine (Quillin & Matthews, 2000; Tsai *et al*., 1999; Liang & Dill, 2001; Fischer *et al*., 2004). Hohlräume zwischen Sekundärstrukturen größerer Proteine können daher die Taschenvorhersage mittels geometrischer Verfahren beeinträchtigen.

Das Kriterium zur Bewertung des Vorhersageerfolges wurde bereits in früheren Arbeiten anderer Gruppen vorgestellt (Brady & Stouten, 2000; Huang & Schröder, 2006) und auch zur Evaluation von *PocketPicker* verwendet. Der zur Bewertung der

Vorhersagequalität verwendete Distanz-Schwellenwert von 4 Å ist dabei zufällig gewählt und kann das Ergebnis nachteilig beeinflussen (Abbildung 31).

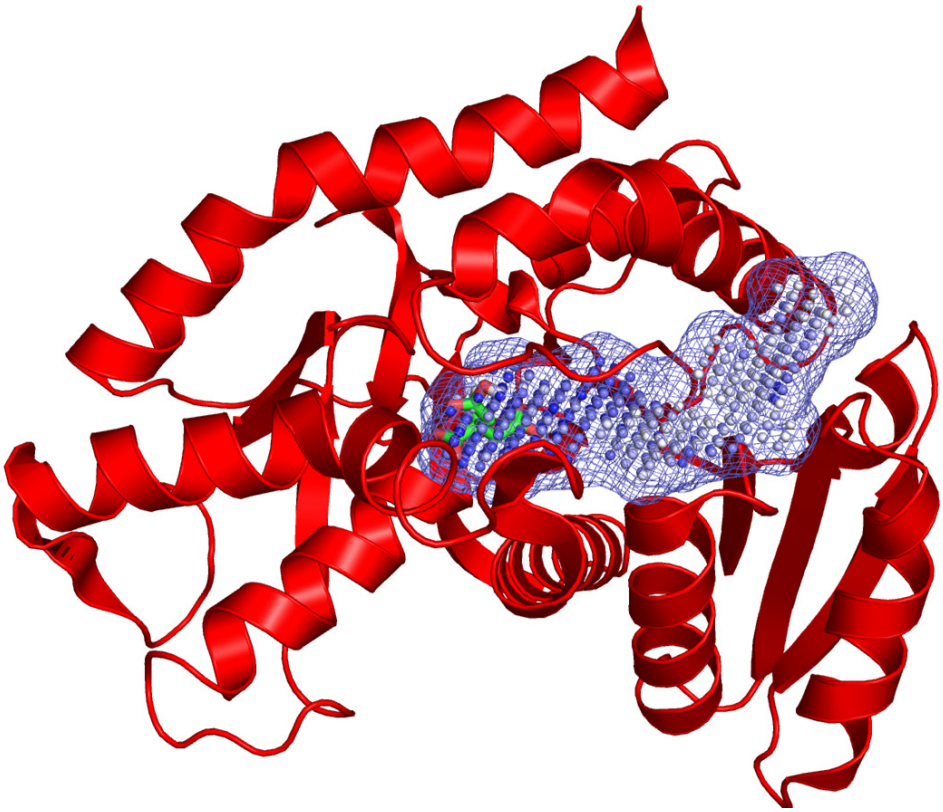


Abbildung 31: Größte von *PocketPicker* identifizierte Tasche für Malat-Dehydrogenase (PDB: 2cmd). Der gebundene Ligand *Citrat* (PDB: cit) besetzt den entlegenen Teil einer langgestreckten Tasche (Gitterdarstellung). Aufgrund der besonderen Taschenform wird dieses Beispiel nicht als korrekte Vorhersage gewertet, da das dem Zentrum der Tasche nächstgelegene Ligandenatom den Schwellenwert von 4 Å überschreitet.

Eine geringfügige Erhöhung des Schwellenwertes um 0,5 Å auf 4,5 Å würde eine Steigerung auf 73% bzw. 77% für die Top1-Vorhersage für die *apo*- und Komplex-strukturen bedeuten (vgl. Tabellen 1 und 2).

Die vorgestellte Software *PocketPicker* stellt eine leistungsfähige Alternative zu den etablierten Verfahren zur Vorhersage von Bindetaschen dar und bildet die Grundlage verschiedener Methoden, die im Rahmen dieser Arbeit entwickelt wurden. Zudem bietet die Darstellung von Bindetaschen als eine Sammlung von Sondenpunkten mit berechneten Vergrabenheiten Ansatzstellen für Techniken, die in anderen Programmen nicht verfügbar sind. So ermöglicht *PocketPicker* den alignmentfreien Vergleich von Bindetaschen durch ShapeDeskriptoren, als auch detaillierte Analysen und Überlagerungen von Bindetaschen unter Berücksichtigung ihrer Eigenschaften mit

PocketShapelets. Weiterführende Techniken umfassen die Extraktion der Topologie von Bindetaschen durch eine GNG Implementierung und die Betrachtung von Subtaschen. Ergebnisse und Anwendungen dieser Techniken sollen in den folgenden Abschnitten vorgestellt und diskutiert werden.

3.1.2 Laufzeitanalyse von PocketPicker

PocketPicker wurde ursprünglich in der Programmiersprache *Python* entwickelt (Weisel *et al.*, 2006), um die Integration in das Programmpaket zur Molekülvisualisierung *PyMOL* (DeLano, 2002; http://pymol.sourceforge.net/) zu ermöglichen, welches seinerseits auf Python basiert. Im Rahmen dieser Arbeit wurde eine neue Version von *PocketPicker* entwickelt und in der Programmiersprache *Java*™ (http://java.sun.com/) realisiert. Für den Leistungsvergleich der beiden Implementierungen wurden Vorhersagen für Bindetaschen auf Kristallstrukturen verschiedener Größe ausgeführt (Tabelle 3). Die Größe der betrachteten Proteine wird hier durch die Anzahl ihrer Atome ausgedrückt, da diese die Laufzeit maßgeblich bestimmen.

Tabelle 3: Vergleich der Laufzeiten (LZ) von *PocketPicker* in verschiedenen Implementierungen der Programmiersprachen Python und Java™. Die Berechnungen wurden auf einer AMD Opteron Workstation mit Zweikernprozessor durchgeführt (2*2GHz). Wasserstoffe wurden mit dem „h_add“ Befehl von PyMOL angefügt. PocketPicker wurde in der öffentlich zugänglichen Python-Version (http://www.modlab.de) als Plugin für PyMOL v1.1r1 und in der neu entwickelten Java™-Variante auf Java™ der Version 1.5 verwendet.

PDB-ID	Bezeichnung	Anzahl Atome	LZ (Python)	LZ (Java™)
121p	Krebsprotein P21	2614	2262 s	< 15 s
1ads	Aldose-Reduktase	5042	4765 s	< 27 s
2a8j	Taspase (Dimer)	8998	13254 s	< 42 s

Die Laufzeitanalyse offenbart einen deutlichen Vorsprung der neu implementierten Variante von *PocketPicker*. Dies ist auf die wesentlichen Unterschiede der beiden Implementierungen zurückzuführen:

(i) In Vergleichstests (sog. *benchmark tests*) konnte gezeigt werden, dass die Architektur der verwendeten Programmiersprache Java™ für vielfältige Aufgaben-stellungen schnellere Berechnungen ermöglicht als die Skriptsprache Python (Fourment

& Gillings, 2008). In Verbindung mit Erweiterungsmodulen (etwa Bausteine der Programmiersprachen C/C++ oder Fortran) sind zwar deutliche Laufzeitbeschleunigungen in Python möglich (Cai *et al.*, 2005), jedoch wurde dieser Ansatz in der vorliegenden Arbeit nicht umgesetzt.

(ii) Die Java™-Variante von *PocketPicker* verfügt über eine verbesserte Methode zur Berechnung der Vergrabenheit von Gittersonden. Diese neue Methode sucht ausgehend von den Atomen eines Proteins nach umliegenden Sonden, die diese Atome mit Hilfe ihrer Suchstrahlen identifizieren können. Dies ermöglicht eine Laufzeit in Abhängigkeit von wenigen tausend Atomen pro Protein (Tabelle 3). Hingegen erfolgt die Bestimmung der Vergrabenheit in der älteren Python-Version von *PocketPicker* durch aufwändigere Abstandsbestimmungen in den Gittersonden zu den Proteinatomen. Dies führt zu einer Laufzeit, die abhängig von vielen zehntausend Sonden ist, die in oberflächennahen Bereichen der Proteinoberfläche installiert sind (27874 Gittersonden für PDB 121p, (45717 [1ads], 77376 [2a8j]).

Der Auswahl einer effizienten Suchroutine kommt besondere Bedeutung zu, da Operationen auf dreidimensionalen Gittern meist kubische Laufzeiten produzieren. Eine Verdoppelung der Kantenlänge des zu Grunde liegenden Gitters resultiert in einer Verachtfachung der Gitterpunkte. Dabei ist zu berücksichtigen, dass der Bereich der Gittersonden, in denen die Berechnung der Vergrabenheit stattfindet auf Regionen entlang der Oberfläche des Proteins, sowie auf Einschlüsse im Proteininnern beschränkt ist. Weiterhin beeinflusst die Globularität eines betrachteten Proteins die Laufzeit, so dass die Rechenzeit nicht direkt aus dessen Größe abgeschätzt werden kann. Für ein Protein mit zerklüfteter Oberfläche und vielen Einschlüssen im Innern entstehen daher größere Laufzeiten als für globuläre Proteine mit vergleichsweise glatter Oberfläche.
Die in dieser Arbeit durchgeführte Implementierung von *PocketPicker* bedeutet eine starke Laufzeitbeschleunigung (Tabelle 3), die eine Berechnung von großen Datenmengen erst ermöglicht.

3.2 Anwendungen der Taschenvorhersagefunktion von PocketPicker.

3.2.1 Identifikation von stabilen Taschen für rationales Moleküldesign in Moleküldynamiksimulationen von TAR-RNA

Die Flexibilität von Makromolekülen spielt eine wichtige Rolle für die molekulare Erkennung und das Bindungsverhalten von Makromolekülen. Dies gilt insbesondere für RNA-Strukturen, deren Zucker-Phosphat-Rückgrat eine ausgeprägte konformationelle Flexibilität ermöglicht, was den rationalen Wirkstoffentwurf erheblich erschweren kann (Abbildung 32). So stellt die Identifikation zeitstabiler Bindestellen einen wichtigen Schritt für das rezeptorbasierte Moleküldesign auf RNA-Molekülen dar.

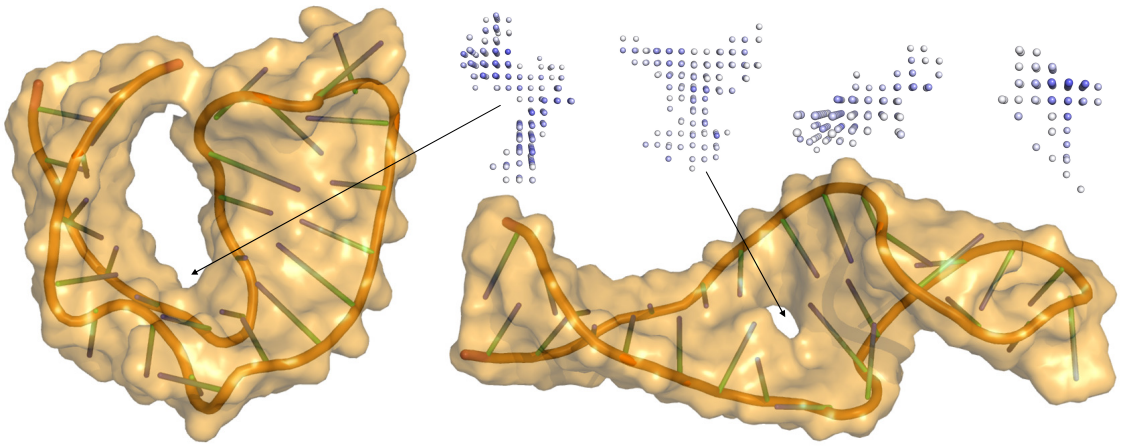


Abbildung 32: Berechnung von Bindetaschen auf TAR-RNA Strukturen. Die gezeigten Moleküle repräsentieren die beiden unterschiedlichsten von 20 Konformationen der NMR-Struktur PDB 1anr. Die große Flexibilität des gezeigten RNA-Moleküls erzeugt vorhergesagte Bindetaschen mit deutlichen strukturelleren Unterschieden. Gezeigt ist die jeweils größte Tasche von vier Konformationen der NMR-Struktur 1anr.

PocketPicker wurde im Rahmen einer Kooperation mit dem Sonderforschungsbereich 579 „RNA-Liganden-Wechselwirkungen“ (Projekt A11.2, gefördert durch die Deutsche Forschungs-gesellschaft, DFG) zur Vorhersage potentieller Bindetaschen auf HIV-TAR RNA verwendet. Die Bildung des Komplexes aus viralem Tat-Protein (engl. *transactivator of transcription*) und TAR-RNA (engl. *transactivation response region*) führt zu einer deutlichen Steigerung der Transkriptionseffizienz des HI-Virus (Dingwall *et al.*, 1989). Aus diesem Grund ist die Inhibierung der TAR-RNA eine mögliche

Anwendung für den rationalen Wirkstoffentwurf von AIDS-Therapeutika (Dingwall *et al*., 1989; Tanrikulu *et al*., 2007).

Zur Repräsentation der konformationellen Flexibilität der TAR-RNA wurden Moleküldynamiksimulationen (MD) für vier in der PDB annotierte NMR-Strukturen (PDB: 1anr, 1arj, 1qd3, 1lvj) durchgeführt. Die MD wurde mit dem Software-Paket AMBER 8 (University of California, San Francisco) mit einer modifizierten Version des „Cornell *et al*. 94 all atom" Kraftfeld (Cheatham *et al.*, 1999) gerechnet (Nietert, 2008). Für 200.000 in der MD erzeugten Konformere der TAR-RNA wurden insgesamt 923.369 potentielle Bindetaschen mit *PocketPicker* extrahiert und in einer Datenbank abgelegt. Informationen über Volumen und Vergrabenheit der vorhergesagten Bindestellen wurden zur Untersuchung der Dynamik zeitstabiler Bindetaschen auf der Zielstruktur verwendet (Nietert, 2008). Ferner konnten durch diesen Ansatz Bereiche auf der TAR-RNA identifiziert werden, die als Ankerpunkte für den Wirkstoffentwurf durch *in silico* Screeningverfahren dienen. So wurde ein sog. „Taschenoberflächenbildungspotential" für die RNA-MD-Strukturen anhand der Taschenvorhersagen mit *PocketPicker* berechnet. Dieses Potential kennzeichnet Teile der Moleküloberfläche, die besonders häufig Begrenzungen für die vorhergesagten Taschen bilden. Diese Oberflächenbereiche werden für den rationellen Wirkstoffentwurf empfohlen (Nietert *et al.*, 2009).

Die in der Datenbank abgelegten Strukturinformationen der mit *PocketPicker* berechneten Bindestellen ermöglichte ferner eine Verknüpfung Bindestellen innerhalb der MD-Konformer-Trajektorien. Anhand dieser Daten lässt sich somit die zeitliche Entwicklung der potentiellen Bindestellen in die Auswahl für geeignete Bindestellen für das Virtuelle Screening integrieren (Nietert, 2008).

3.2.2 Virtuelles Screening nach Inhibitoren der Serinprotease HtrA von *Helicobacter pylori* mit ReverseLIQUID

Helicobacter pylori (*H. pylori*) ist ein Gram-negatives, mikroaerophiles Bakterium (Taylor, 2006). Es kolonisiert die menschliche Magenschleimhaut, wobei ca. 50% der Menschheit befallen ist. Als Pathogen begünstigt es die Entstehung von Magengeschwüren und Magenkrebs (Kwok *et al.*, 2007).

Experimentelle Befunde deuten darauf hin, dass *H. pylori* während der Infektion Kontakt zu Membranproteinen der Wirtszellen aufnimmt, um ein Typ IV Sekretionssystem aufzubauen und den primären Virulenzfaktor **CagA** (engl. *Cytotoxin associated antigen A*) in die Wirtszelle zu translozieren. Diese **Integrine** genannten Membranproteine werden bei polaren Epithelzellen allerdings bevorzugt basolateral exprimiert (Kwok *et al.*, 2007). Außerdem können extrazellulär geschnittene **E-Cadherin**-Fragmente im Medium mit *H. pylori* infizierter Zellkulturen nachgewiesen werden (Weydig *et al*., 2007). Beide Beobachtungen legen den Schluss nahe, dass eine Protease von *H. pylori* sekretiert wird und die Zell-Zell-Kontakte degradiert, um *H. pylori* den Zugang zur basolateralen Seite der Wirtszellen zu ermöglichen (Löwer *et al*., 2008). Auch für eine Metastasierung entarteter Wirtszellen kann ein solcher Prozess entscheidend sein.

Das vom Gen Hp1019 des Stammes *H. pylori* 26695 codierte Protein **HtrA** konnte im Rahmen einer Kooperation mit dem Paul-Ehrlich-Institut in Langen im Überstand von *H. pylori* mit proteolytischer Aktivität nachgewiesen werden (Löwer *et al.*, 2008). Um den Einfluss dieser extrazellulären Protease auf die Infektion von Kulturzellen mir *H. pylori* zu untersuchen, sollte ein niedermolekularer Inhibitor für HtrA gefunden werden. Hierzu wurde ein Homologiemodell erstellt, das auf der aktiven Konformation der Protease HtrA von *Escherichia coli* (PDB: 3cs0) basiert.

PocketPicker wurde für die Charakterisierung potentieller Bindetaschen auf diesem Homologiemodell eingesetzt und bildet ferner die Grundlage für eine Methode, die ein **strukturbasiertes Virtuelles Screening** für die berechneten Taschen ermöglicht. Diese „**ReverseLIQUID“** (Weisel, 2006) genannte Technik verbindet die geometrische Darstellung von Bindetaschen durch *PocketPicker* mit der Beschreibung physikochemischer Eigenschaften durch LUDI-Atomtypen (Böhm, 1991). Dies gelingt durch die komplementäre Projektion der Eigenschaften der Proteinatome, die die Bindetasche einrahmen auf die Sondenpunkte, die diese charakterisieren. ReverseLIQUID identifiziert Bereiche der Tasche, die eine Bindung von Wasserstoffbrücken-Akzeptoren oder -Donoren oder die Wechselwirkungen von lipophilen Molekülfragmenten mit den Resten des Proteins begünstigen und fasst diese zu sog. **Potentiellen Pharmakophor-**

Punkten (*PPP*s) zusammen (Abbildung 33). Dies ermöglicht die Darstellung einer extrahierten Tasche als die räumliche Verteilung ihrer Eigenschaften über PPPs. Ferner erlaubt dieser Ansatz die Kodierung eines rezeptorbasierten Pharmakophormodells durch Autokorrelationsdeskriptoren, wie sie in der Software LIQUID zur Beschreibung von ligandenbasierter Pharmakophore verwendet werden (Tanrikulu *et al*., 2007). Auf diese Weise können die mit ReverseLIQUID extrahierten LIQUID-Deskriptoren zur Suche in kommerziellen Substanzbibliotheken verwendet werden.

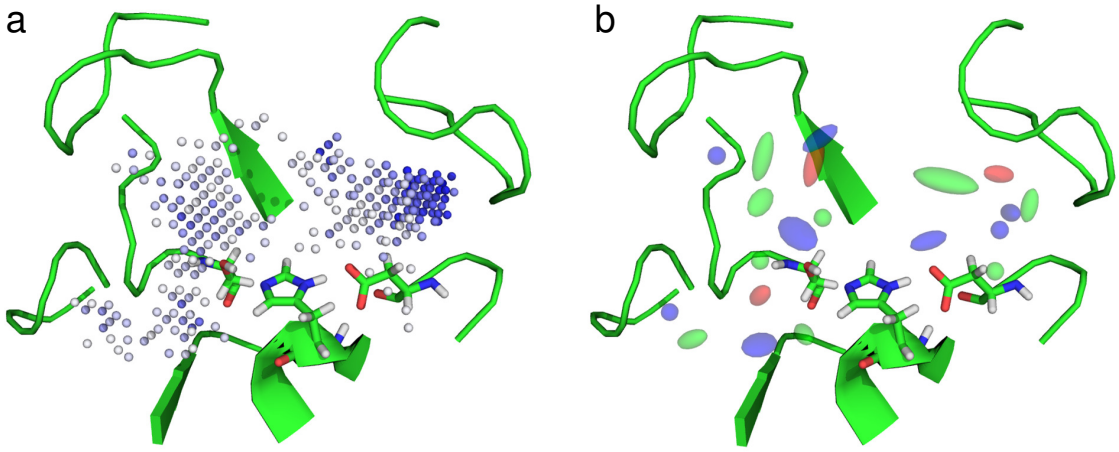


Abbildung 33: Extraktion eines rezeptorbasierten Pharmakophormodells mit ReverseLIQUID. a) Bestimmung möglicher Bindestellen mit *PocketPicker* für das Homologiemodell der Protease HtrA (Ausschnitt in Cartoondarstellung). b) LUDI Atomtypen der taschenflankierenden Reste werden komplementär auf die Gittersonden projiziert. ReverseLIQUID gruppiert Bereiche mit ähnlichen Eigenschaften zu PPPs, die als trivariate Gaußfunktionen ein „fuzzy" Pharmakophormodell beschreiben. Dieser Ansatz ermöglicht ein virtuelles Screening in Substanzbibliotheken über LIQUID-Autokorrelationsdeskriptoren.

In retrospektiven Studien konnte die Funktion dieser Berechungen für eine Auswahl an pharmakologisch wichtigen Proteinen aus verschiedenen Strukturklassen validiert werden (nicht gezeigt). Dabei stellte sich vor allem eine Abhängigkeit der Güte der Modelle von der Güte der Vorhersage von *PocketPicker* heraus, was den Schluss zulässt, dass eine möglichst genaue Definition der Bindetasche für das Gelingen eines strukturbasierten virtuellen Screening unerlässlich ist.

Für die Protease HtrA von *H. pylori* konnten drei strukturabgeleitete Pharmakophormodelle berechnet werden, wobei jeweils verschiedene von *PocketPicker* vorhergesagte Bindetaschen einbezogen wurden. Die Molekülkataloge der Firmen Asinex und Specs wurden nach Ähnlichkeit zu diesen Modellen sortiert und nach

Begutachtung der jeweils ähnlichsten 100 Substanzen wurden 26 Substanzen ausgewählt und bestellt. In einem *in vitro* Assay mit der rekombinanten Protease HtrA inhibierten sechs Substanzen den Verdau eines rekombinanten Substrats. Die beste Verbindung erreichte in dem Assay eine maximale Inhibition von ca. 78% bei einer mittleren inhibitorischen Konzentration bei halbmaximaler Inhibition (IC_{50}) von 26 µM (Löwer *et al*., 2009).

3.3 Verwendung von ShapeDeskriptoren zur Vorhersage der Druggability von Proteinbindetaschen

In der vorliegenden Arbeit wurde eine Technik entwickelt, die versucht die „Druggability" einer zuvor mit *PocketPicker* ermittelten Proteinbindetasche vorherzusagen (Weisel *et al*., 2009). Die Druggability beschreibt die Eignung einer Bindestelle einen Wirkstoff zu binden. Für die Untersuchung dieser Eigenschaft wurden Bindetaschen für einen repräsentativen Datensatz von Protein-Liganden Komplexen berechnet, für die experimentell bestätigte Bindeaffinitäten bekannt sind. Zur Einschätzung der Druggability wurde eine Klassifizierung mit SOMs (Kohonen, 1989) auf den ShapeDeskriptoren der extrahierten Bindetaschen durchgeführt.

3.3.1 Verwendete Protein-Datensätze für Druggability-Untersuchungen

Für die Untersuchung der Druggability wurden zwei **Protein-Datensätze** A und B zusammengestellt, um einerseits ein repräsentatives Profil des Taschen-Universums (Pocketom) zu charakterisieren und zugleich eine sequentielle Diversität zu garantieren, die eine Beeinflussung der Vorhersage durch Homologien in den betrachteten Strukturen ausschließt. Beide Datensätze wurden vom sog. *Refined Set* der PDBbind Datenbank (Wang *et al.*, 2004; Wang *et al.*, 2005) abgeleitet. Diese Sammlung umfasst 1300 Komplexe für die Bindeaffinitäten ihrer Liganden bekannt sind. Das *Refined Set* wurde seinerseits aus der PDB abgeleitet und umfasst Protein-Liganden Strukturen, die eine Reihe von Qualitätskriterien erfüllen müssen. So müssen die Kristalle eine

Auflösung von ≤ 2,5 Å besitzen und eine sterisch günstige Bindung des Liganden ohne Überschneidungen im Komplex ermöglichen.
Datensatz A wurde in zwei Schritten aus dem *Refined Set* abgeleitet:

(i) Monomere Strukturen und solche, die in ihre jeweiligen monomeren Varianten überführt werden konnten, wurden in einer Sammlung von 909 Komplexen zusammengefasst. Druggability-Analysen wurden ausschließlich auf Monomeren durchgeführt, da multimere Strukturen Einschlüsse zwischen ihren Ketten beinhalten können, die von *PocketPicker* als Bindetaschen missinterpretiert werden können.

(ii) Die Sammlung wurde weitergehend auf diejenigen Komplexe beschränkt, bei denen das Volumen der vorhergesagten ligandenbindenden Taschen das Volumen des betreffenden Liganden nicht um mehr als 50% überschreitet.

Der resultierende Datensatz A umfasst 623 Proteine mit ebenso vielen Bindetaschen für die im *Refined Set* beschriebenen Referenzliganden, Bindestellen von anderen Liganden und leeren Taschen.
Datensatz B wurde vom sog. *Core Set* der PDBbind abgeleitet, das seinerseits eine Auswahl der im *Refined Set* annotierten Komplexe darstellt, die jedoch phylogenetisch divers hinsichtlich ihrer Aminosäurekonserviertheit sind: Das *Core Set* umfasst 210 Komplexe mit einer gegenseitigen Sequenzidentität von weniger als 75%. Datensatz B besteht aus den 98 Komplexen des *Core Sets*, die bereits in Datensatz enthalten sind. Datensatz B wurde in dieser Arbeit verwendet, um zufällige Gruppierungen von Daten auf der SOM zu verhindern, die durch Ähnlichkeiten homologer Proteine bedingt sein können.

3.3.2 Beschreibung des Liganden-Datensatzes

Für jeden Komplex eines Protein-Datensatzes wurde eine Bindetasche ausgewählt, die dem jeweiligen PDBbind „Referenzliganden“ am nächsten gelegen ist. Diese Tasche muss nicht notwendigerweise der Top1-Treffer einer Vorhersage mit *PocketPicker* sein. Neben den Referenzliganden können weitere Liganden einer Struktur enthalten sein, die

als „Zusatzliganden“ in den **Liganden-Datensatz** aufgenommen wurden. Ionen und Wassermoleküle wurden nicht in diese Sammlung einbezogen, um zu verhindern, dass die zugehörigen Bindetaschen als *druggable* interpretiert werden. Ein Vergleich mit der Collection of Bioactive Reference Analogues (COBRA) Datenbank (Schneider & Schneider, 2003) wurde vorgenommen, um die Mindestgröße annotierter wirkstoffartiger Moleküle zu bestimmen. Von den 8311 enthaltenen Molekülen waren alle Liganden innerhalb von zwei Standardabweichungen ($2\sigma = 20{,}4$) um den Mittelwert (29,4 Schweratome) aus mindestens neun Schweratomen aufgebaut. Folglich wurden alle Moleküle mit weniger als neun Atomen aus der Sammlung der Zusatzliganden entfernt. Aufgrund ihres ubiquitären Bindeverhaltens (Qasba, 1999) wurden Zucker ebenfalls nicht in den Liganden-Datensatz aufgenommen. Dieser umfasst 623 Referenzliganden und 19 Zusatzliganden für Datensatz A, sowie 98 Referenzliganden und 7 Zusatzliganden für Datensatz B. Abbildung 34 zeigt die Charakterisierung der Taschen eines Komplexes.

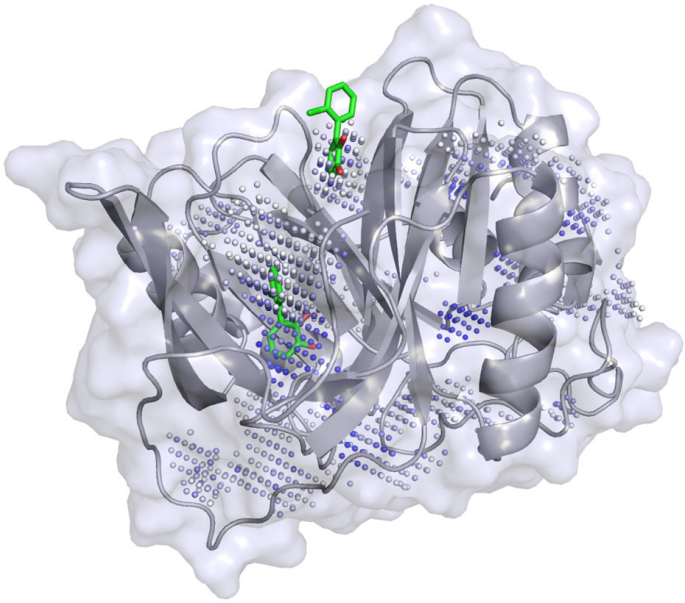


Abbildung 34: Charakterisierung der mit *PocketPicker* vorhergesagten Bindetaschen der Dihydroxybiphenyl-Dioxygenase (DHBD, PDB: 1lgt). Der in der PDBbind beschriebene Referenzligand 2'-Chloro-Biphenyl-2,3-Diol (PDB: bp3), sowie ein Zusatzligand (hier ebenfalls bp3) kennzeichnen ihre Taschen als druggable.

Für die 623 Taschen von Datensatz A wurden 13859 potentielle Bindetaschen extrahiert (2257 für Datensatz B). Der Vorhersagequalität von *PocketPicker* liegt hier bei 77% für Top1-Treffer und erreicht eine korrekte Identifikation von 90% für Top3-Treffer.

3.3.3 Analyse der Größe und Vergrabenheit potentieller Bindetaschen

Es konnte gezeigt werden, dass Bindetaschen, die in der Lage sind wirkstoffartige Moleküle zu binden größer sind als leere Taschen und eine erhöhte Vergrabenheit aufweisen (Weisel *et al.*, 2007). Diese Beobachtung trifft auch für die vorhergesagten Taschen der Datensätze A und B zu (Abbildung 35).

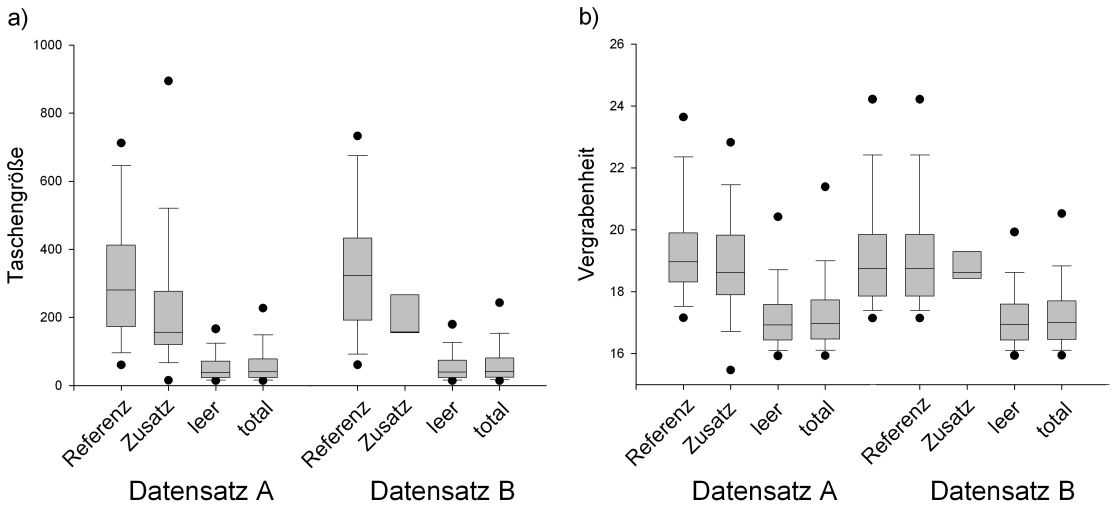


Abbildung 35: Boxplot-Darstellungen der Taschengröße (a) und Vergrabenheit der Taschen aus den Datensätzen A und B. Ergebnisse werden für Referenzliganden (*Referenz*), Zusatzliganden (*Zusatz*), leere Taschen (leer) und den vollständigen Datensatz (*total*) gezeigt. Die schwarzen Kreise kennzeichnen das 5. und 95. Perzentil.

Die Taschen, die mit Referenz- oder Zusatzliganden besetzt sind, weisen deutlich größere Vergrabenheit und Volumen an. Dessen ungeachtet kennzeichnen die jeweils größten Taschen der Datensätze A und B leere Taschen mit Volumen von 2121 $Å^3$ (Carbamoyl Phosphat Synthetase, PDB: 1bxr) und 2010 $Å^3$ (Hepatitis C Virus RNA Polymerase, PDB: 1nhu). Ferner wurde die größte durchschnittliche Vergrabenheit mit einem Wert von 25,5 ebenfalls für eine leere Tasche eines Proteins festgestellt, das in

beiden Datensätzen enthalten ist (Beta Secretase, PDB: 1fkn). Die Durchschnittswerte für Vergrabenheiten und Größe der Taschen sind in Tabelle 4 zusammengefasst.

Tabelle 4: Durchschnittliche Größe und Vergrabenheit der Taschen von Datensatz A und B. Taschengrößen werden als Anzahl der Gittersonden interpretiert, die das Bindevolumen repräsentieren. Vergrabenheitswerte beziehen sich auf *PocketPicker* Vergrabenheits-Indizes.

	Datensatz A		**Datensatz B**	
	Ø Größe	Ø Vergrabenheit	Ø Größe	Ø Vergrabenheit
Taschen mit Referenzliganden	323,0	19,4	349,8	19,2
Taschen mit Zusatzliganden	229,4	18,9	223,6	18,8
Leere Taschen	59,5	17,4	62,6	17,3
Gesamter Datensatz	71,5	17,5	75,6	17,4

Neben einer ausreichenden Taschengröße wird eine hinreichende Vergrabenheit als notwendige Voraussetzung für die Ausbildung nicht-kovalenter Wechselwirkungen zwischen Ligand und Rezeptor angesehen (Weisel *et al.*, 2009). Eine Analyse der Taschengröße und Vergrabenheit für die verwendeten Datensätze ist in Abbildung 36 dargestellt. Die Abbildung verdeutlicht, dass bekannte Ligandenbindetaschen nicht gleichmäßig hinsichtlich ihrer Größe und Vergrabenheit verteilt sind, sondern vielmehr bestimmte Bereiche des Ergebnisraums bevorzugen.

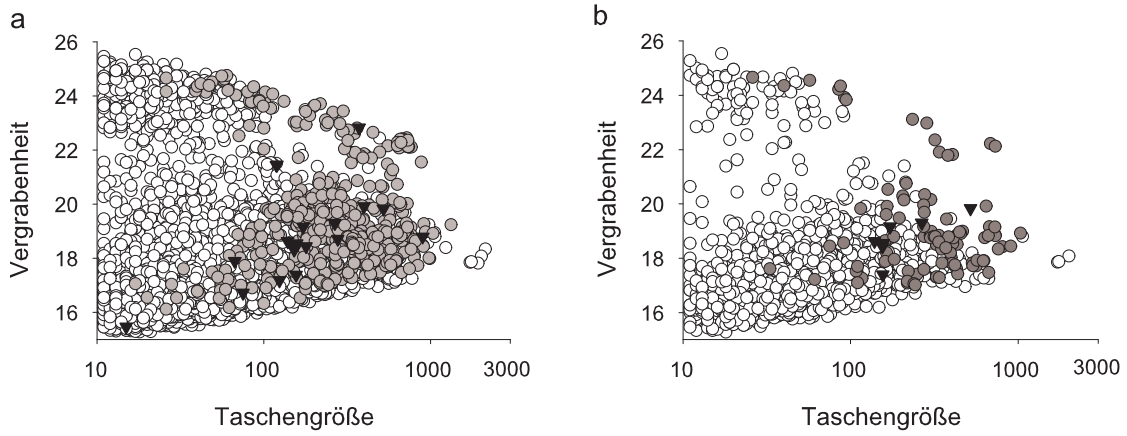


Abbildung 36: Untersuchung der Taschengrößen und Vergrabenheiten für die Datensätze A (a) und B (b). Gezeigt sind die Verteilungen für leere Taschen (weiße Kreise), Zusatzliganden (schwarze Dreiecke) und Referenzliganden (schwarze Kreise).

Ligandenbindende Taschen zeichnen sich durch ein vergrößertes Volumen und eine erhöhte Vergrabenheit aus (Tabelle 4). Die Ergebnisse zeigen weiterhin, dass es für beide Datensätze je zwei unabhängige Populationen hinsichtlich der Vergrabenheit gibt. Diese Gruppen beschreiben flache Bindestellen (durchschnittliche Vergrabenheit < 20), sowie sehr tief vergrabene Taschen (durchschnittliche Vergrabenheit > 22,5). Der dazwischen liegende Bereich ist dagegen relativ dünn besiedelt, obgleich dieser Raum mit steigender Taschengröße (Volumen > 100 $Å^3$) viele ligandenbindende Taschen enthält. Dies mag darin begründet sein, dass Proteine eine abnehmende Kompaktheit bei steigendem Molekulargewicht aufweisen (Fischer *et al.*, 2004). So beschreiben tief vergrabene Taschen auf großen Proteinen oftmals Tunnel, die unterhalb der Oberfläche verlaufen. Die Ergebnisse zeigen, dass viele dieser vergrabenen Tunnel Liganden binden und somit die Funktion des Proteins bestimmen. Dies bestätigt die Einschätzung, dass große Taschen mit erhöhter Wahrscheinlichkeit als Ligandenbindestellen fungieren können (Weisel *et al.*, 2007). Darüber hinaus empfehlen diese Ergebnisse vergrabene tunnelartige Taschen als mögliche Startstellen für den rezeptorbasierten Wirkstoffentwurf.

Eine weitere Besonderheit stellt die Verteilung der ligandenbindenden Taschen mit durchschnittliche Vergrabenheit > 22,5 dar. Unter diesen 55 Taschen (Abbildung 36) treten Häufungen der folgenden Proteinfunktionen auf: Zuckerbindend/Zuckertransport (n = 15), aminosäurebindend (n = 12), Hormon-/Wachstumsfaktor-Rezeptoren (n = 8), Hydrolasen/Lysozyme (n = 8). Ergebnisse früherer Arbeiten bestätigen, dass Zucker für gewöhnlich mit geringer Affinität an flache Taschen binden, die von Schleifenregionen ausgebildet werden, wohingegen die Bindung von Sacchariden durch Zuckertransporter meist in vergrabenen Taschen mit hoher Affinität gelingt (Qasba, 1999). Dieser Befund wird durch die berechneten hohen Vergrabenheitswerten bestätigt.

3.3.4 Klassifikation der Druggability mit Selbstorganisierenden Karten

Für detaillierte Protein-Druggability-Analysen wurden das Programm molmap® (Schneider & Wrede, 1998) verwendet, um SOMs auf den von *PocketPicker*

berechneten ShapeDeskriptoren der verwendeten Datensätze A und B zu trainieren. Für die Projektion der Daten wurden SOMs mit 10x15 Neuronen in toroidaler Netzwerktopologie gewählt. Das Training wurde über 400.000 Epochen unter Verwendung einer Gauß'schen Nachbarschaftsfunktion (Startradius 7 mit linearer Adaption) durchgeführt. Für das Clustering der Daten wurde die Euklidische Distanzmetrik verwendet. Die Projektion der ShapeDeskriptoren auf die SOMs zeigte eine deutliche Gruppierung der ligandenbindenden für beide verwendeten Datensätze (Abbildung 37).

Die Ergebnisse dieser SOM-Projektionen verdeutlichen die Leistungsfähigkeit von *PocketPicker* ShapeDeskriptoren zwischen ligandenbindenden und unbesetzten Taschen zu unterscheiden. Dabei erlaubt die SOM eine Identifikation einer vergleichsweise kleinen Zahl von ligandenbindenden Taschen (623 von 13859 Einträgen für Datensatz A) innerhalb einer vielfach größeren Menge an leeren Taschen.

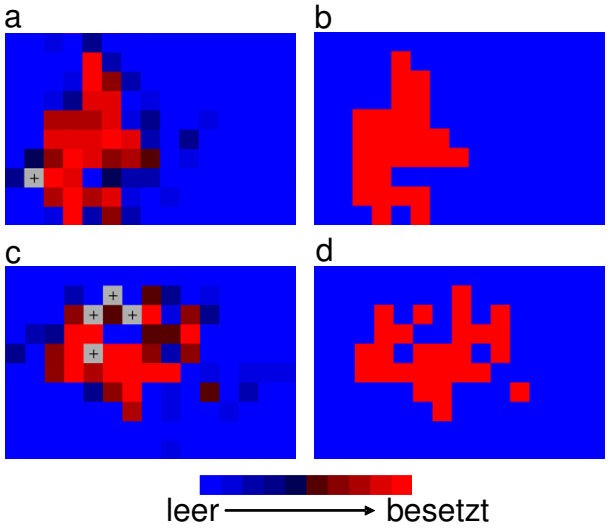


Abbildung 37: Projektion der extrahierten Taschen aus den Komplexen der Datensätze A (a, b) und B (c, d). Relative Häufigkeiten sind in (a) und (c) gegeben. Diese SOMs zeigen rote Felder für Neurone die ausschließlich von ligandenbindenden Taschen besetzt sind; blaue Bereiche repräsentieren Neuronen, die lediglich von leeren Taschen besetzt sind. Neurone, die Mischungen beider Taschentypen enthalten gemäß der jeweiligen relativen Häufigkeit gefärbt. Neurone denen keine Taschendeskriptoren zugewiesen wurden sind mit ‚+' markiert. Binäre Klassifikationen sind in (b) und (d) gegeben. Neurone werden hier entweder als rote oder blaue Felder dargstellt, wenn sie mehrheitlich durch Beschreiber von ligandenbindenden oder leeren Taschen besetzt sind.

Zur **Abschätzung der Klassifikationsgüte** wurden Matthews Korrelationskoeffizienten *cc* (engl. c*orrelation coefficient*) für die Projektionen der SOMs berechnet (Gleichung

12). ligandenbindende Taschen wurden hierbei als „positiv korrekt“ (*P*) bewertet, wenn sie auf rote Neurone projiziert wurden und als „unterbewertet“ (engl. *underpredicted*, *U*) eingeschätzt, wenn sie in blaue Bereiche der binären SOM abgebildet wurden. Die Projektionen der ShapeDeskriptoren leerer Taschen auf blaue Felder der binären SOM wurden als „negativ korrekt“ (*N*) erachtet, wohingegen Abbildungen auf rote Bereiche als „überbewertet“ (engl. *overpredicted*, *O*) bezeichnet wurden.

$$cc = \frac{P \cdot N - O \cdot U}{\sqrt{(P+O)(P+U)(N+O)(N+U)}}. \tag{12}$$

Die für die Datensätze A und B berechneten SOMs erreichten Matthews Korrelationskoeffizienten von 0,72 und 0,76 (*cc* = 1 kennzeichnet 100% korrekte Zuweisungen, s.a. Anhang, Tabelle A2). Diese Ergebnisse bestätigen die Eignung der vorgestellten Methode zur Unterscheidung von ligandenlbindenden und leeren Taschen.

3.3.5 Untersuchung der Bioverfügbarkeit und Verteilung der Drug-Likeness von Liganden für Taschen mit hoher Drug-Likeness

Zur Beschreibung der Druggability wurde in dieser Arbeit ein Konzept verwendet, das die Fähigkeit eines Proteins untersucht, Liganden mit hoher Affinität zu binden. Weitere Untersuchungen wurden durchgeführt, um die Eigenschaften zu untersuchen, die einen Liganden als wirkstoffartig (engl. *drug-like*) charakterisieren. So wurden Lipinski-Deskriptoren für die Liganden der verwendeten Datensätze mit der Software MOE (*Molecular Operating Environment*, Chemical Computing Group, Montreal, Canada, http://www.chemcomp.com) berechnet. Zur Bewertung der **oralen Bioverfügbarkeit** wurde hierbei untersucht, inwieweit die betrachteten Liganden den von Lipinski aufgestellten „Rule of Five“ (*Ro5*) entsprechen (Lipinski *et al.*, 1997). Diese auch als **Lipinski-Regeln** bekannten Konventionen beschreiben Richtwerte, die Vielfache von fünf oder gleich fünf sind. Diese Richtlinien sagen eine potentiell schlechte passive Absorption oder Permeation für oral verabreichte Substanzen voraus, die mehr als eine der folgenden Regeln verletzen:

- Molekulargewicht < 500 Da
- begrenzte Lipophilie (ausgedrückt als Verteilungskoeffizient *clogP* < 5)
- höchstens 5 Wasserstoffbrücken-Donoren (Summe von OH und NH Gruppen)
- höchstens 10 Wasserstoffbrücken-Akzeptoren (Summe von O- und N-Atomen)

In diesem Sinne besitzen 80% der Liganden aus Datensatz A (79% für Datensatz B) günstige Lipinski-Eigenschaften. Da die *Ro5* lediglich die potentielle orale Bioverfügbarkeit von Molekülen betreffen, wurde eine genauere Analyse der Drug-Likeness durchgeführt. Hierzu wurde ein Konzept verwendet, das Neuronale Netzwerke zur Vorhersage der Drug-Likeness von Substanzen einsetzt (Sadowski & Kubinyi, 1998; Ajay *et al*., 1998). In dieser Arbeit wurde eine Reimplementation dieses Ansatzes genutzt, welches Werte zwischen null (nicht *drug-like*) und eins (*drug-like*) für gegebene Liganden berechnet (Schneider & Schneider, 2004). Die Vorhersagen zeigen eine annähern balancierte Verteilung der Drug-Likeness der Liganden aus den verwendeten Datensätzen (Abbildung 38).

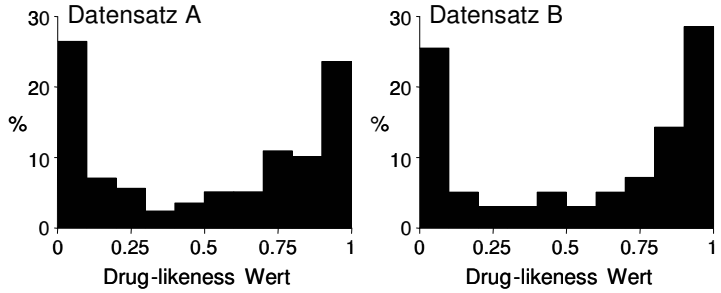


Abbildung 38: Berechnete Drug-Likeness-Werte (ANN Vorhersage) für die Liganden der Datensätze A und B.

Die Ergebnisse dieser Vorhersagen weisen mindestens 50% der Daten einen Drug-Likeness-Wert von > 0,5 zu. Darüber hinaus erlaubte diese Analyse eine Beschreibung der verwendeten Moleküle als „besonders drug-like“ (Wert > 0,9) und „wenig drug-like“ (Wert < 0,1). In weiteren Untersuchungen wurden diese Informationen genutzt, um die Verteilung der ShapeDeskriptoren auf den trainierten SOMs hinsichtlich der Drug-Likeness ihrer korrespondierenden Liganden zu analysieren. So wurden Teilmengen von ShapeDeskriptoren aus Datensatz B abgeleitet, für deren Liganden Drug-Likeness-Werte von > 0,9 (n = 28) oder < 0,1 (n = 25) vorhergesagt wurden.

Diese Untermengen wurden als repräsentative Wirkstoff- bzw. Nichtwirkstoff-Daten interpretiert und auf die SOMs zurückprojiziert, die zuvor mit den vollständigen Datensätzen A und B trainiert worden waren (Abbildung 39).

Da für die Rückprojektion Deskriptoren ligandenbindender Taschen gewählt wurden, die aus den Datensätzen A und B abgeleitet sind, können nur Neurone belegt werden, die bereits auf den ursprünglichen SOMs von Taschen mit Referenz- oder Zusatzliganden besetzt waren (Abbildung 37). Die Taschen mit wirkstoffartigen Liganden besiedeln erkennbar die zentralen Bereiche der Druggability-Inseln der ursprünglich auf den Datensätzen A und B trainierten SOMs.

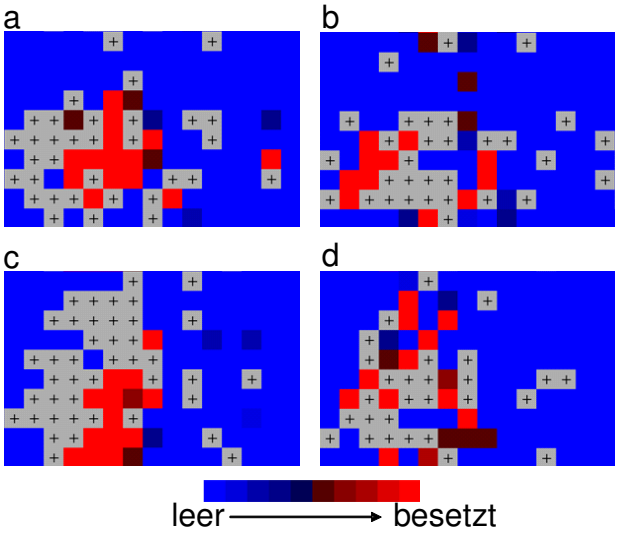


Abbildung 39: Projektion der ShapeDeskriptoren der Taschen aus Datensatz B, die Liganden mit sehr großer oder sehr niedriger vorhergesagter Drug-Likeness beinhalten auf die SOMs aus Abb. 37. *Obere Reihe*: Datensatz B, *untere Reihe*: Datensatz A. Taschen mit wirkstoffartigen Liganden (ANN-Vorhersagewert > 0,9) besiedeln zentrale Bereiche der Druggability-Inseln (a), während Taschen mit nicht-wirkstoffartigen Liganden Randbereiche besetzen. Dieser Effekt ergibt sich auch für Projektionen der Deskriptoren auf die mit Datensatz A trainierten SOM (c, d). Blaue Bereiche enthalten Deskriptoren leerer Taschen, Beschreiber von ligandenbindenden Taschen erscheinen rötlich. Neurone denen keine Taschendeskriptoren zugewiesen wurden sind mit ‚+' markiert.

Die Taschen, die Liganden mit nur geringer Drug-Likeness enthalten belegen hingegen Randbereiche und zeichnen die Umrisse der Druggability-Inseln nach. Dieser Effekt lässt sich besonders gut an den entsprechenden binären SOMs der Wirkstoff- / Nichtwirkstoff-Projektionen ablesen (Abbildung 40).

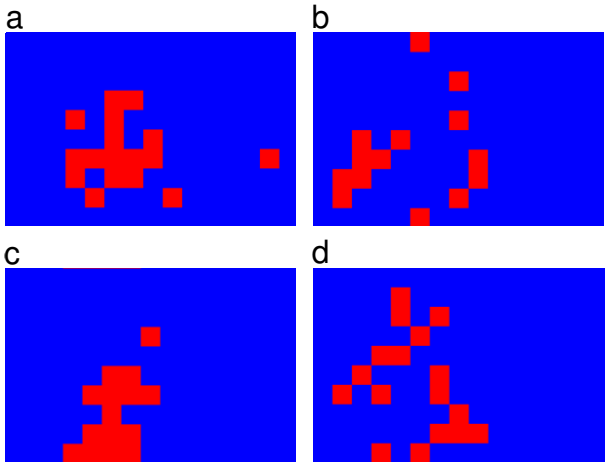


Abbildung 40: Binäre Klassifikation der als besonders drug-like (a, c) bzw. wenig drug-like (b, d) vorhergesagten Taschen des Datensatzes B. Die obere Reihe zeigt die Projektion auf die mit Datensatz B trainierte SOM, die untere Reihe die Abbildung auf die SOM des Datensatzes A. Taschen, die wirkstoffartige Liganden enthalten besetzen zentrale Bereiche der SOMs, während die Deskriptoren von Taschen, die Liganden mit geringer Drug-Likeness binden die Ränder der Druggability-Inseln nachzeichnen.

Die Ergebnisse verdeutlichen, dass der vorgestellte Ansatz nicht nur in der Lage ist ligandenbindende Taschen von leeren zu unterscheiden (Druggability), sondern darüber hinaus qualitative Aussagen über die Fähigkeit der Taschen treffen kann, wirkstoffartige Moleküle zu binden (Drug-Likeness). Bemerkenswert ist die Fähigkeit dieses Ansatzes, ligandenbindende Taschen aus einem Datensatz von potentiellen Bindetaschen zu identifizieren und darüber hinaus eine deutliche Unterscheidung der Drug-Likeness innerhalb der Druggability-Inseln zu gewährleisten. Es soll jedoch erwähnt werden, dass ein Ligand mit niedrigem Drug-Likeness-Wert nicht notwendigerweise ein Nicht-Wirkstoff sein muss. So sind alle Liganden des verwendeten Datensatzes in der PDBbind annotiert und binden mit experimentell bestätigter Affinität an ihre jeweiligen Rezeptoren.

Die für die Datensätze A und B berechneten Druggability-Inseln sind ähnlich in ihrer Ausdehnung, besitzen jedoch eine unterschiedliche Gestalt (Abbildung 37 a), c)). Dies ist darin begründet, dass die SOMs auf verschiedenen Datensätzen trainiert wurden, die unterschiedliche Inhalte zur Klassifikation darstellen. Darüber hinaus beschreibt das kompetitive Lernen als Grundprinzip der selbstorganisierenden Karten eine nicht-deterministische Methode des maschinellen Lernens (Fritzke, 1994; Fritzke *et al.*, 1995). Dennoch ist die SOM in der Lage, dicht besiedelte Inseln mit ligandenbindenden

Taschen zu visualisieren. Die SOMs zeigen ferner eine starke Anreicherung von wenigen Hundert ligandenbindenden Taschen als eine Insel (rote Bereiche) umgeben von einem Ozean aus vielen Tausend leeren Taschen (blaue Flächen). Diese Betrachtung bestätigt erneut die diskriminative Qualität selbstorganisierender Karten. So erreichen die Projektionen der Taschen mit Liganden von hoher Drug-Likeness (ANN-Vorhersagewert > 0,9) Korrelationskoeffizienten von *cc* = 0,88 und *cc* = 0,89 für Abbildungen auf die binären SOMs der Datensätze A und B (siehe Anhang, Tabelle A3). Projektionen der Taschen mit Liganden von niedriger vorhergesagter Drug-Likeness (ANN-Vorhersagewert < 0,1) erreichen Korrelationskoeffizienten von *cc* = 0.88 und *cc* = 0,79 für die Datensätze A und B (siehe Anhang, Tabelle A3).

3.3.6 Vorhersage der Druggabilty für *apo*-Strukturen

Die Suchroutine von *PocketPicker* ist in der Lage, Bindetaschen von *apo*-Strukturen mit hoher Genauigkeit (71% korrekte Vorhersagen, 88% korrekte Vorhersage als eine der drei größten Taschen) vorherzusagen (Tabelle 2). Die Datensätze, die für Druggability-Vorhersagen mit selbstorganisierenden Karten verwendet wurden, bestehen aus Protein-Liganden-Komplexen, und nutzen ligandenbindende Taschen zur Charakterisierung der Druggability. Für den strukturbasierten Wirkstoffentwurf ist es von besonderem Interesse speziell die Druggability von *apo*-Strukturen abzuschätzen. Um diesen Aspekt zu untersuchen, wurden Bindetaschen für diejenigen Einträge des Datensatzes A mit *PocketPicker* berechnet, für die *apo*-Strukturen in der PDB (Berman *et al*., 2000) existieren. Die *apo*-Strukturen von sechs Kristallstrukturen wurden dazu mit ihren entsprechenden Komplexen strukturell aligniert, um die leere Tasche zu identifizieren, die mit der ligandenbindenden Tasche des Komplexes korrespondiert. Für die Überlagerung wurde der *align*-Befehl aus PyMOL (Version 1.0r2; DeLano, 2002) verwendet. Die ShapeDeskriptoren der betreffenden *apo*-Taschen wurden auf die SOM der Abbildung 37 a), c) projiziert. Die Ergebnisse zeigen, dass der größte Teil der *apo*-ShapeDeskriptoren dieselben oder benachbarte Neuronen besiedeln, wie ihre korrespondierenden *holo*-ShapeDeskriptoren (Tabelle 5).

Tabelle 5: Karten-Koordinaten der ShapeDeskriptoren ligandenbindender Taschen ausgewählter Komplexe (*holo*) und ihren entsprechenden *apo*-Taschen abgebildet auf die mit den Datensätzen A und B trainierten SOMs (vgl. Abbildung 37 a), c)). Die Koordinaten beziehen sich auf die trainierten SOMs, wobei (0,0) das obere linke Neuron (erste Reihe, erste Zeile) bezeichnet.

PDB-Eintrag holo/*apo*	(x,y) Position auf der mit Datensatz A trainierten SOM (*holo*)/(*apo*)	(x,y) Position auf der mit Datensatz B trainierten SOM (*holo*)/(*apo*)
1rbp/1brq	(3,5)/(3,5)	(3,2)/(3,5)
7cpa/5cpa	(4,4)/(5,4)	(4,4)/(5,4)
2ctc/2ctb	(4,3)/(5,4)	(3,4)/(5,4)
2h4n/2cba	(5,4)/(5,3)	(6,6)/(5,3)
2sim/2sil	(5,5)/(5,8)	(5,4)/(5,8)
2tmn/ll3f	(7,6)/(7,6)	(8,4)/(7,6)

Diese Eigenschaft bestätigt sich sowohl für die SOMs der Datensätze A und B und deutet an, dass die in dieser Arbeit entwickelte Methode zur Vorhersage der Druggability auch für *apo*-Strukturen anwendbar ist.

3.3.7 Weitere Anwendung der Druggability Analyse mit Selbstorganisierenden Karten

In dieser Arbeit wurde eine neue Methode zur Klassifikation der Druggability von Proteinbindetaschen unter Verwendung von Selbstorganisierenden Karten entwickelt. Eine weitergehende Untersuchung der Funktion der Ligandenbindestellen ist dabei durch genauere Analyse dieser Taschen mittels Künstlicher Neuronaler Netze (*ANN*) oder Support-Vektor-Maschinen (*SVM*; Burges, 1998; Cortes & Vapnik, 1995) denkbar. Die Verwendung dieser maschinellen Lernverfahren ist hierbei durch die Beschreibung der Bindetaschen durch Autokorrelationsdeskriptoren möglich. Dieses Konzept erlaubt darüber hinaus die Untersuchung weiterer Eigenschaften durch die Kodierung weiterer Charakteristika.

Neben der Zugänglichkeit für maschinelle Lernverfahren und Vorteilen für die Berechnungszeit von Ähnlichkeitssuchen kann die Abstraktion von Eigenschaften durch Korrelationsdeskriptoren jedoch einen Verlust an Information bedeuten. So enthält die Kodierung von Taschenvolumen durch ShapeDeskriptoren keine Aussage über deren Stereochemie oder Spiegelsymmetrie. Dies kann zu unzureichenden oder fehlerhaften Ähnlichkeitszuweisungen führen. Die Verwendung zusätzlicher geometrischer

Variablen eröffnet Lösungswege für diese Problematik (Pastor *et al*., 2000; Cruciani *et al*., 2000).

3.4 Konformationsanalyse von Aldose-Reduktase Bindetaschen mit ShapeDeskriptoren

Der in *PocketPicker* implementierte ShapeDeskriptor repräsentiert die unterschiedlichen Formen und Vergrabenheiten von Proteinbindetaschen durch einen Autokorrelations-deskriptor. Dies ermöglicht eine schnelle und alignmentfreie Ähnlichkeitssuche zuvor charakterisierter Bindestellen über Euklidische Abstandsbestimmungen. In diesem Abschnitt soll die Anwendung von *PocketPicker* ShapeDeskriptoren zum Vergleich von Induced-Fit Phänomenen in Aldose-Reduktase vorgestellt und diskutiert werden.

3.4.1 Auswahl des Datensatzes von Aldose-Reduktase Kristallstrukturen

Die Untersuchung der konformationellen Ähnlichkeit von Bindetaschen homologer Proteine wurde auf einem Datensatz von 13 Aldose-Reduktase Kristallstrukturen durchgeführt. Diese waren zuvor in einer Arbeit von Sotriffer und Mitarbeitern diskutiert worden (Sotriffer *et al*., 2004). Der Datensatz umfasst neun Strukturen humaner Aldose-Reduktase (PDB: 1ads, 1el3, 1iei, 1us0, 2acq, 2acr, 2acs, die Tyr48His-Mutante 2acu und die Cys298Ala/Trp219Tyr Doppelmutante 1az1), sowie vier Aldose-Reduktasen (PDB: 1ah0, 1ah3, 1ah4, 1eko) aus dem Schwein (Tabelle 6). Die hochauflösende (0,66 Å) Kristallstruktur von PDB 1us0 diente als Referenzpunkt. Die Proteine des Datensatzes zeigen eine Sequenzidentität von $\geq 85\%$ gegenüber dieser Referenz und besitzen eine Auflösung von $\leq 2{,}5$ Å.

Taschenvorhersagen mit *PocketPicker* wurden auf Aldose-Reduktase Strukturen mit gebundenem Co-Faktor NADPH oder $NADP^+$ durchgeführt. Weitere in den Komplexen enthaltene Liganden wurden vor Beginn der Berechnungen entfernt (Abbildung 41).

Tabelle 6: Verwendeter Datensatz aus 13 Aldose-Reduktase Kristallstrukturen.

PDB	Spezies	Ligand (PDB)	Auflösung [Å]
1ads	*Homo sapiens*	-	1,65
1ah0	*Sus scrofa*	Sorbinil (sbi)	2,3
1ah3	*Sus scrofa*	Tolrestat (tol)	2,3
1ah4	*Sus scrofa*	-	2
1az1	*Homo sapiens*	-	1,8
1eko	*Sus scrofa*	IDD384 (i84)	2,2
1el3	*Homo sapiens*	IDD384 (i84)	1,7
1iei	*Homo sapiens*	Zenarestat (zes)	2,5
1us0	*Homo sapiens*	IDD594 (ldt)	0,66
2acq	*Homo sapiens*	Glucose-6-Phosphat (g6p)	1,76
2acr	*Homo sapiens*	Cacodylat-Ion (cac)	1,76
2acs	*Homo sapiens*	Citrat (cit)	1,76
2acu	*Homo sapiens*	Citrat (cit)	1,76

Für eine einheitliche räumliche Ausrichtung der Proteine wurden alle Strukturen über den *align*-Befehl von PyMOL (DeLano, 2002) gegenüber der *apo*-Struktur PDB 1ads strukturell überlagert. Wasserstoffe wurden vor Beginn der Berechnungen mit PyMOL an die Strukturen angefügt.

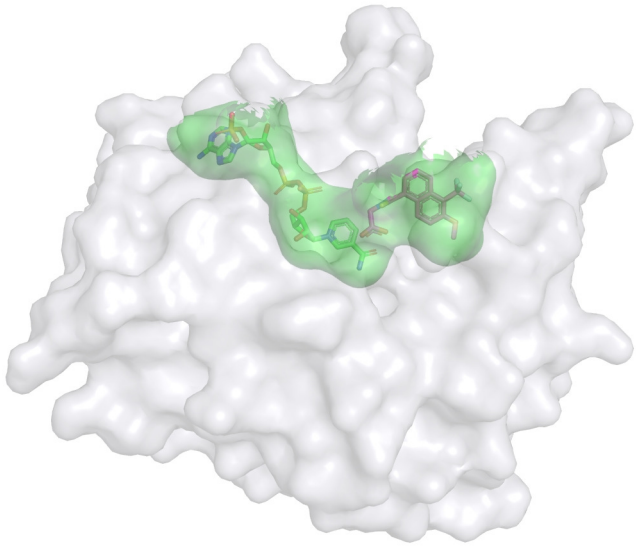


Abbildung 41: Aldose-Reduktase (PDB: 1ah3) mit gebundenem $NADP^+$ (grün) und Inhibitor Tolrestat (magenta). Der linke Teil der tunnelartigen Bindetasche bindet den Co-Faktor, während der rechte Teil verschiedene Liganden über Induced-Fit Verhalten binden kann.

3.4.2 Vergleich der automatischen Ähnlichkeitssuche mit Ergebnissen der manuellen Inspektionen

Vier verschiedene Bindetaschen-Konformationen wurden von Sotriffer und Mitarbeitern (Sotriffer *et al*., 2004) anhand visueller Untersuchungen unterschieden, nach den jeweils gebundenen Liganden benannt und in vier Klassen von Taschenformen eingeteilt: Die „IDD-594"-Konformation, die „holo"-Konformation (mit gebundenem Co-Faktor, aber ohne weiteren Liganden), die „Tolrestat"-Konformation und die „Zenarestat"-Konformation. An dieser Stelle sollen diese Bezeichnungen verwendet werden, um die verschiedenen Klassen struktureller Konformationen zu kennzeichnen, die durch Induced-Fit Verhalten während der Bindung der verschiedenen Liganden verursacht werden. In dieser Arbeit wurde eine automatische Charakterisierung der Bindetaschen von Aldose Reduktase anhand von zuvor für die Strukturen berechneten ShapeDeskriptoren vorgenommen. Ähnlichkeiten zwischen den verschiedenen Klassen von Taschenformen werden durch die Euklidische Distanz ihrer ShapeDeskriptoren ausgedrückt. Die Ergebnisse des automatisierten Taschenvergleichs für den verwendeten Datensatz sind in Tabelle 7 zusammengefasst. Sämtliche Vorhersagen mit *PocketPicker* haben das katalytische Zentrum der Aldose Reduktase als jeweils größte Tasche korrekt vorhergesagt.

Tabelle 7: Euklidische Distanzen zwischen den ShapeDeskriptoren der größten mit *PocketPicker* vorhergesagten Bindetaschen für einen Datensatz von 13 Aldose Reduktasestrukturen.

	1ah0	1ah3	1ah4	1az1	1eko	1el3	1iei	1us0	2acq	2acr	2acs	2acu
1ads	4065	3880	2861	1943	2309	1776	3592	3249	2448	2325	1536	3075
1ah0		1457	1637	3093	2746	2587	2459	2005	2281	2196	2967	1707
1ah3			1675	3216	2466	2372	1687	1852	2362	2360	3018	1605
1ah4				2396	2130	1405	2360	2094	1567	1353	2141	1009
1az1					2120	1673	2894	2217	1446	1449	1167	2452
1eko						1634	2036	1930	2073	2107	1541	2115
1el3							2356	2109	1354	1225	1428	1652
1iei								1414	2385	2484	2727	2091
1us0									1873	1927	2053	1937
2acq										549	1632	1335
2acr											1509	1344
2acs												2245

Der Komplex aus Aldose Reduktase und dem potenten Inhibitor (IC_{50} = 30nM) IDD594 (PDB: 1us0; Podjarny *et al*., 2004) bildet eine eigene konformationelle Klasse innerhalb

des betrachteten Datensatzes (Abbildung 42 a)). Eine strukturelle Ähnlichkeit zur Zenarestat-Konformation konnte anhand der berechneten ShapeDeskriptoren festgestellt werden (PDB: 1iei, Abbildung 42 b)): Von allen mit *PocketPicker* berechneten ShapeDeskriptoren zeigte der Beschreiber von 1iei die kleinste Euklidische Distanz zur IDD594-Konformation (Tabelle 7). Dies bestätigt die Ergebnisse der manuellen Inspektion, die den Bindemodus von Zenarestat und IDD594 als sehr ähnlich beschreiben (Sotriffer *et al.*, 2004).

Der Tolrestat-Komplex (PDB: 1ah3, Abbildung 42c)) kennzeichnet eine weitere Bindestellenkonformation, die sich deutlich von den anderen Taschengeometrien des Datensatzes unterscheidet. Dies wird erneut durch den ShapeDeskriptor reflektiert, der deutliche Euklidische Distanz zu anderen Beschreibern aufweist (Tabelle 7).

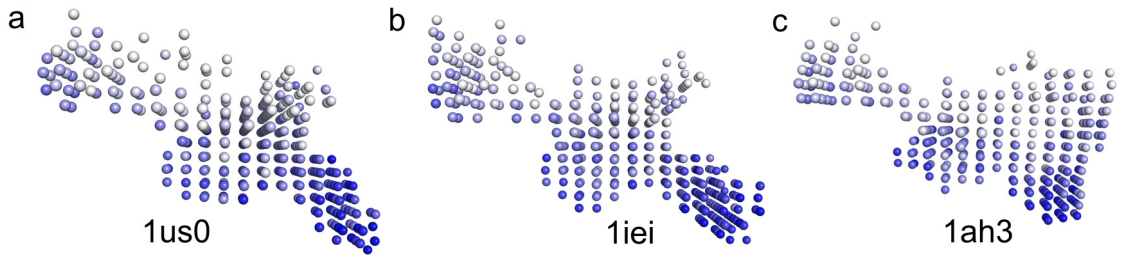


Abbildung 42: Taschenkonformationen der Aldose-Reduktase induziert durch (a) IDD594, (b) Zenarestat und (c) Tolrestat. Bindetaschen sind in *PocketPicker* Repräsentation dargestellt. Dunklere Kugeln kennzeichnen Taschenbereiche mit erhöhter Vergrabenheit.

Die Mehrheit der Konformationen wurde der *holo*-Konformation zugeordnet (1ADS, 1AH0, 1AH4, 1AZ1, 1EKO, 1EL3, 2ACQ, 2ACR, 2ACS, 2ACU), von denen drei Strukturen (1AH0, 1AH4, 1EKO) eine Untermenge mit kleinen Abweichungen zur Standard *holo*-Konformation aufweisen (Sotriffer *et al*., 2004). Dies wird wieder durch die ShapeDeskriptoren der drei Einträge (1AH0, 1AH4, 1EKO) reflektiert, die erhöhte Distanzen gegenüber der Standard *holo*-Konformation (1ADS) aufweisen (Tabelle 7).

Für eine detaillierte Untersuchung der verschiedenen Taschenkonformationen wurde ein hierarchisches Clustering mit dem Programmpaket PHYLIP (Phylogeny Interference Package; Felsenstein, 1989) durchgeführt. Durch Anwendung des in PHYLIP implementierten Neighbor-Joining-Algorithmus (Saitou & Nei, 1987) wurde ein ungewurzelter Baum für die ShapeDeskriptoren der extrahierten Taschen der Aldose-Reduktase erstellt (Abbildung 43).

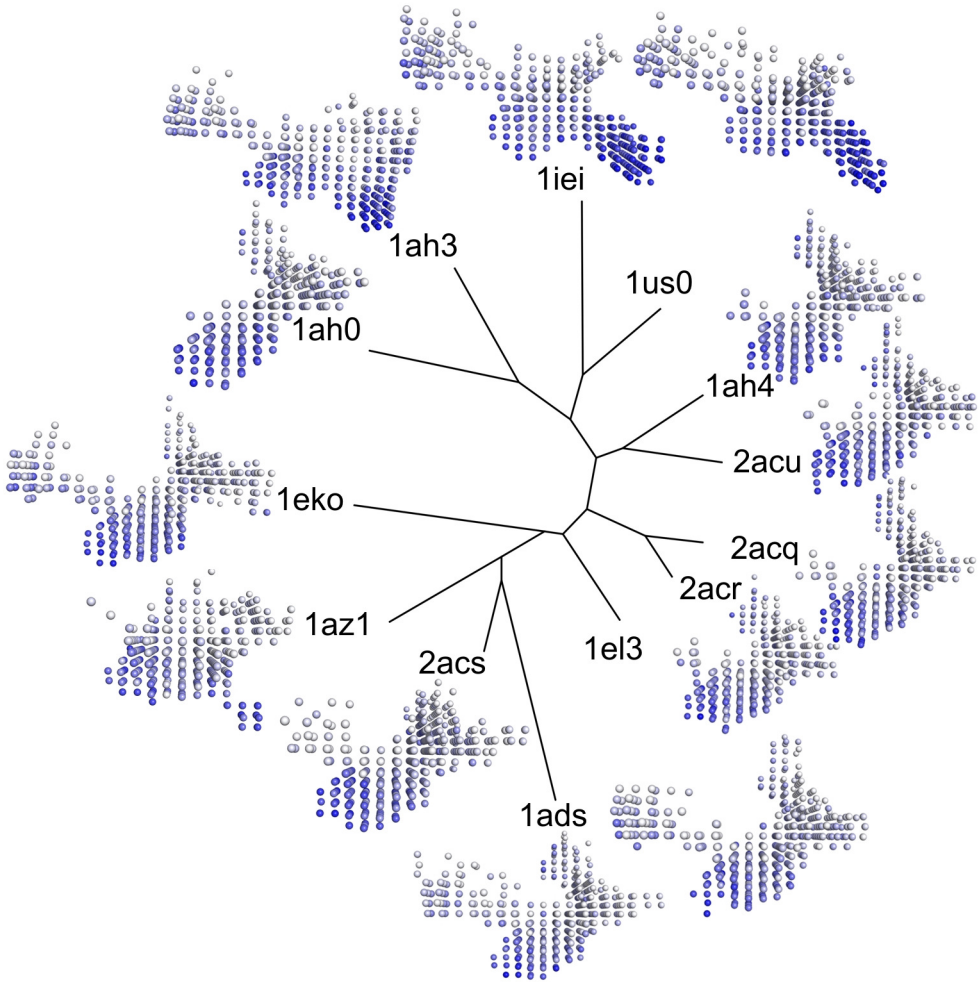


Abbildung 43: Hierarchisches Clustering der mit *PocketPicker* extrahierten ShapeDeskriptoren für 13 Bindetaschen homologer Aldose Reduktasen.

Ähnlichkeiten zwischen den ShapeDeskriptoren werden als Euklidische Abstände zwischen den Autokorrelationsdeskriptoren ausgedrückt und als Kantenlängen der Äste im Neighbor-Joining-Baum dargestellt. Die Taschenanalysen zeigten eine deutliche konformationelle Ähnlichkeit der ligandenbindenden Taschen von 1iei und 1us0, die folglich auch eine exponierte Stellung im Neighbor-Joining-Baum einnehmen. Charakteristisches Merkmal der IDD594- und der Zenarestat-Konformation ein Bereich mit erhöhter Vergrabenheit, der erst durch die Bindung der entsprechenden Liganden im Rezeptor induziert wird. Die Ausprägung einer solchen tief vergrabenen Subtasche bildet ein deutliches Unterscheidungsmerkmal gegenüber den übrigen Taschengeometrien des Datensatzes und ist sonst einzig in der Tolrestat-Konformation erkennbar (Abbildung 44).

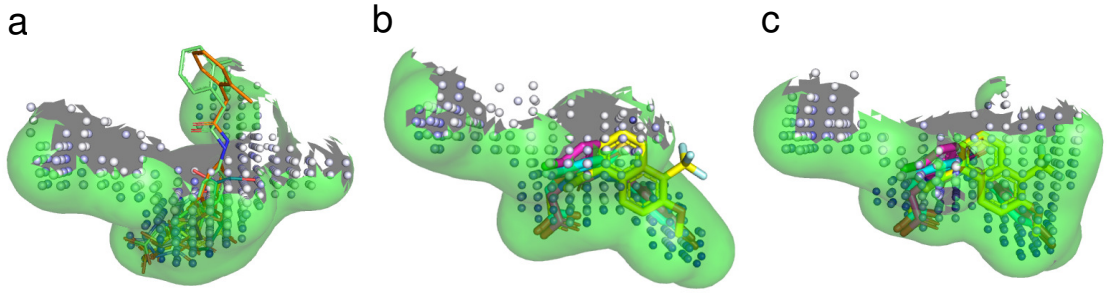


Abbildung 44: Visualisierung des Induced-Fit Bindeverhaltens von Aldose-Reduktase. Geglättete Taschenoberflächen wurden durch Einpassen von Gauß-Funktionen in die Kugelrepräsentationen berechnet. Abgeschnittene Oberflächenstücke beschreiben zum Solvens gerichtete Bereiche der Tasche. a) Repräsentation der *holo*-Konformation durch die extrahierte katalytische Tasche von PDB 1ads. Die überlagerten Liganden Sorbinil, IDD384, Glucose-6-Phosphat, Cacodylat-Ion und Citrat verursachen keinen nennenswerten Induced-Fit und besetzen ein Volumen, das ähnlich dem der unbesetzten Tasche ist. b) Taschenrepräsentation der IDD594-Konformation (PDB: 1us0) mit gebundenem Inhibitor IDD594 (türkis). Zenarestat (magenta; aus PDB: 1IEI) zeigt den gleichen Bindemodus, während Tolrestat (gelb; aus PDB: 1ah3) die Taschenoberfläche durchbricht. c) Die Tolrestat-Konformation (PDB: 1ah3) öffnet ebenfalls einen vergrabenen Bereich für die Bindung des Liganden. Dieser ist jedoch gegenüber der IDD594-Konformation verdreht.

3.4.3 Leistungsfähigkeit und Begrenzungen des ShapeDeskriptors für den Taschenvergleich

Der Vergleich von potentiellen Bindetaschen über den in *PocketPicker* realisierten ShapeDeskriptor ermöglicht eine effiziente Ähnlichkeitssuche auch für große Datenbanken, die auf rechenintensive strukturelle Alignments verzichtet. Der verwendete Autokorrelationsdeskriptor kodiert eine starke Abstraktion der Taschenform und verzichtet auf eine formgebende Beschreibung wie Oberflächenkurvatur oder Globularität des eingeschlossenen Volumens. Eine Tasche wird vielmehr als Summe der Distanzen zwischen den in *PocketPicker* berechneten Gittersonden beschrieben, die die Form der Bindestelle beschreiben. Zusätzlich werden hier die Vergrabenheiten der Sonden betrachtet, die ebenfalls im ShapeDeskriptor kodiert sind. Der für eine Tasche berechnete Deskriptor zeichnet daher ein Profil, das das Vorkommen von Domänen mit verschiedenen Vergrabenheiten innerhalb der Bindestelle dokumentiert. Dieses Profil ist nicht eindeutig für eine Tasche und ebenfalls nicht eindeutig für eine Konformation. So können die Korrelationsdeskriptoren von Taschen mit unterschiedlicher Konfiguration unter Umständen geringe Euklidische Distanzen zueinander besitzen, wenn sie Bereiche ähnlicher Vergrabenheiten beinhalten, die zudem vergleichbare Abstände zueinander aufweisen (Abbildung 45).

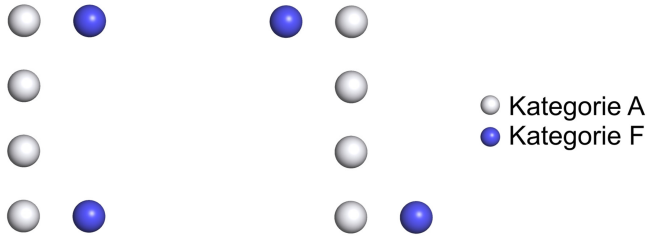


Abbildung 45: Gezeigt ist eine vereinfachte Darstellung zweier zwei-dimensionaler Taschen (Maschenweite des Gitters = 1 Å) mit gleichen Vergrabenheiten (je vier Sonden der Kategorie A, zwei Sonden der Kategorie F), die trotz verschiedener Konfiguration identische Deskriptoren erzeugen können. Betrachtet man Paarungen von Gittersonden und ihren Vergrabenheiten über Distanzen von < 3 Å, so werden für beide Taschen identische Deskriptoren berechnet. Bei einem Abstand von genau 3 Å wird für das linke Beispiel das Kategoriepaar F-F gefunden.

Dieser Effekt kann auch bei der Klassifizierung der induzierten Taschenkonformationen der Aldose Reduktasen beobachtet werden (Abbildung 43). So besitzt die Sorbinil-bindende Tasche (PDB: 1ah0) geringste berechnete Distanz (d = 1457) zur Tolrestat-Konformation (1ah3), obwohl sie eine andere Form beschreibt und tatsächlich der Klasse der *holo*-Konformation angehört. Die geringfügig größere nächst ähnliche Distanz (d = 1657) wird jedoch zu einem Deskriptor der *holo*-Konformation ausgegeben (PDB: 1ah4) und hätte somit eine korrekte Klassifizierung bedeutet.

Die Ähnlichkeit der ShapeDeskriptoren der Sorbinil- und Tolrestat-Konformation beruht ebenfalls auf dem Vorkommen von Gittersonden mit ähnlichen Vergrabenheiten und Distanzen (Abbildung 46).

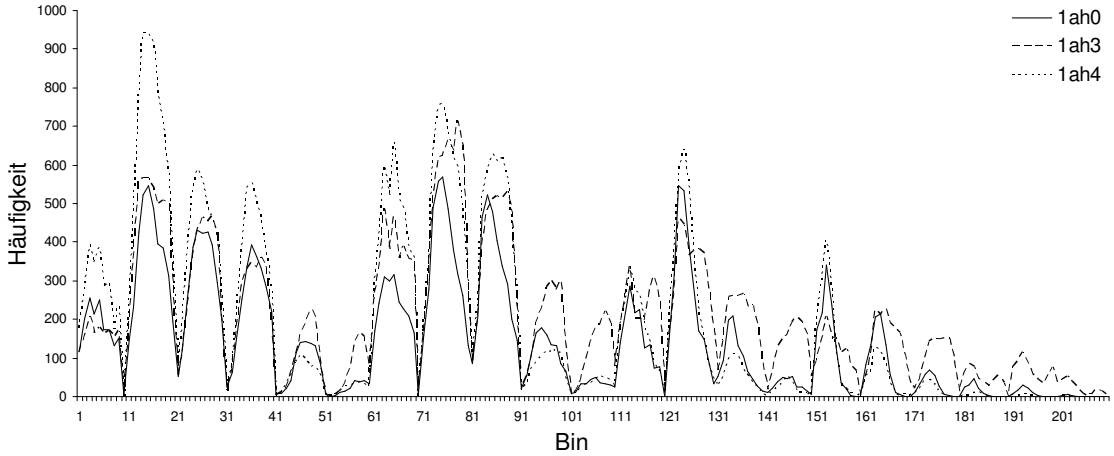


Abbildung 46: Histogramm der ShapeDeskriptoren der Sorbinil-bindenden Tasche (PDB: 1ah0, *holo*-Konformation) und Vergleich zur Tolrestat-Konformation (PDB: 1AH3) und dem nächst ähnlichen Deskriptor (PDB 1ah4, *holo*-Konformation). Größere Unterschiede für Bereiche niedrigerer Vergrabenheiten (Positionen 1-40) erzeugen erhöhte Distanzen zwischen den beiden Beschreibern der *holo*-Konformation.

Die Abbildung zeigt ähnliche Profile für die Vertreter der *holo*-Konformation (PDB; 1ah0, 1ah4) mit jedoch erheblichen Unterschieden in den ersten 40 Bins. Große Ähnlichkeiten für Distanzen zwischen tiefer vergrabenen Sonden (Bins 91-210) können den akkumulierten Fehler nicht ausreichend ausgleichen, so dass eine kleinere Distanz der Sorbinil-bindenden Tasche zur Tolrestat-Konformation ausgegeben wird.

Dieses Beispiel zeigt, dass die Ähnlichkeitssuche über Korrelationsdeskriptoren abhängig von der Genauigkeit der Vorhersageroutine von *PocketPicker* ist und wie präzise diese das Bindevolumen umrandet. So sind tiefer vergrabene Bereiche zweier ähnlicher Taschen in der Vorhersage meist besser eingegrenzt als die dem Lösungsmittel zugänglichen Abschnitte. Hier können Unterschiede in der Oberflächenkurvatur Unterschiede für das geometrische Clustering der Taschen bedeuten. Dies kann die hier vorliegenden erhöhten Differenzen für die ersten Positionen des ShapeDeskriptors zur Folge haben. Da die meisten Bindetaschen größere Lösungsmittel-zugängliche Bereiche besitzen, akkumulieren Unterschiede besonders stark für diese Abschnitte. Eine stärkere Gewichtung der hinteren Positionen des ShapeDeskriptors, die die seltener auftretenden tiefer vergrabenen Bereiche beschreibt, kann diesen Nachteil ausgleichen. Dieser Ansatz wurde in dieser Arbeit nicht verfolgt. Stattdessen wurde eine Methode entwickelt, die eine Beschreibung der Oberflächenkurvaturen von Bindetaschen über PocketShapelets verfolgt. Dieser Ansatz erlaubt eine detaillierte Charakterisierung von Taschenformen und erweitert die Beschreibung durch Korrelationsdeskriptoren.

3.5 Funktionsvorhersage der Proteinfunktion anhand von PocketPicker ShapeDeskriptoren

Der in *PocketPicker* implementierte ShapeDeskriptor ermöglicht den alignmentfreien Vergleich von zuvor extrahierten Bindetaschen. Dieser Ansatz erlaubt eine Funktionsvorhersage für Proteinbindetaschen über Ähnlichkeitsvergleiche mit den Einträgen einer vorberechneten Datenbank von Taschendeskriptoren und ihren annotierten Funktionen. Zu diesem Zweck wurden ShapeDeskriptoren für die Taschen der 1296 Komplexe des PDBbind *Refined Set* der Version 2005 (Wang *et al.*, 2004;

Wang *et al.*, 2005) vorbereitet und in einer Referenzdatenbank zusammengefasst. In diesem Abschnitt sollen Ergebnisse von Funktionsvorhersagen mit *PocketPicker* ShapeDeskriptoren vorgestellt werden.

3.5.1 Identifikation und Funktionsvorhersage einer neuen Bindetasche für APOBEC3C unter Verwendung von ShapeDeskriptoren

Die humanen APOBEC3 Proteine (engl. *apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3*) sind Proteine, die in somatischen Zellen an der Verhinderung der Replikation von Retroviren und Retrotransposons beteiligt sind. APOBEC3 Proteine werden von einer Wirtszelle spezifisch in retrovirale Virionen eingeschleust und wirken dort als Cytidin-Deaminasen. Die Schädigung des viralen Genoms gelingt durch Desaminierung von Cytosinresten zu Uracilresten, was entsprechend zu Hypermutationen von Guanin nach Adenin auf dem Gegenstrang führt. Dieser Mechanismus verhindert eine Vermehrung des Virus.

Im Verlauf der Evolution haben Retroviren Wege gefunden, um den Einflüssen von APOBEC3 zu entkommen. Dies gelingt etwa durch die akzessorischen Vif-Proteine des HI-Virus (HIV). Die Vif-Proteine (engl. *virus infectivity factor*) interagieren mit APOBEC3 und verhindern dessen Aktivität, indem sie diese der Ubiquitinierung und dem Abbau über das Proteasom zuführen (Conticello *et al*., 2003). Die viralen Strukturen, die APOBEC3 am Virion erkennt, sind noch nicht abschließend bekannt.

PocketPicker wurde für eine Charakterisierung der Bindetaschen von APOBEC3C verwendet. Da keine Kristallstruktur von APOBEC3C vorliegt, wurde ein Homologiemodell verwendet, das in einer vorhergehenden Arbeit erstellt wurde (Stauch *et al.*, 2009). Dieses basiert auf den Kristallstrukturen des humanen APOBEC2 und APOBEC3G, die etwa 32% bzw. 40% Sequenzidentität gegenüber APOBEC3C aufweisen. Als größte Tasche wurde eine Vertiefung in etwa 15 Å Abstand zum aktiven Zentrum mit *PocketPicker* gefunden. Für diese potentielle Bindestelle wurde eine Ähnlichkeitssuche anhand von ShapeDeskriptoren in einer vorberechneten Datenbank

mit 1296 Taschen des PDBbind *Refined Set* (Version 2005; Wang *et al.*, 2004; Wang *et al.*, 2005) durchgeführt. Die vier ähnlichsten gefundenen Taschen (mit geringster Euklidischer Distanz zum ShapeDeskriptor der Suchanfrage) bilden Bindestellen auf Proteinen mit DNA bindender Funktion oder RNase-Aktivität (Tabelle 8).

Tabelle 8: Beschreibung der ähnlichsten gefundenen Einträge der vorberechneten PDBbind Datenbank für die größte Tasche eines Homologiemodells von APOBEC3C. Die Ähnlichkeit wird als die Euklidische Distanz *d* der ShapeDeskriptoren ausgedrückt.

PDB	*d*	Beschreibung	Funktion
1r6n	435	humaner Papillomavirus Transaktivierungsdomäne	DNA bindend
1qhc	497	bovine Ribonuklease A	Nukleinsäure bindend, Nuklease
1jn4	563	bovine Ribonuklease A	Nukleinsäure bindend, Nuklease
1o0m	577	bovine Ribonuklease A	Nukleinsäure bindend, Nuklease

Die Mutation eines konservierten Argininrestes am Eingang der größten vorhergesagten Bindetasche führte dazu, dass APOBEC3C nicht mehr ins Virion verpackt werden konnte und damit inaktiv wurde. Die Fähigkeit zur Desaminierung der viralen DNA hingegen blieb bestehen (Stauch *et al*., 2009). Diese Beobachtung korreliert mit der Beobachtung einer veränderten Interaktion von APOBEC3C mit 7SL und 5,8S RNA. Es konnte nachgewiesen werden, dass Wildtyp APOBEC mit diesen kleinen RNAs interagiert, die Argininmutante jedoch nicht mehr (Stauch *et al*., 2009). Diese beiden RNAs sind Teil des Ribosoms und beeinflussen die Proteinsortierung. Es ist daher denkbar, dass die vorhergesagte Tasche eine Bindung an 7SL und 5,8S vermittelt und auf diesem Wege die Verpackung in den Viruspartikel steuert.

Das gezeigte Beispiel beschreibt eine erfolgreiche Anwendung der Vorhersageroutine von *PocketPicker* und des Ähnlichkeitsvergleichs über ShapeDeskriptoren. Die Funktionsvorhersage über den Vergleich mit ähnlichen Taschenformen lässt die Fähigkeit der Tasche zur Bindung an Nukleinsäuren vermuten. Diese These konnte durch anschließende Untersuchungen zur Interaktion mit kleinen RNAs bestärkt werden.

3.5.2 Ähnlichkeitsvergleich und Funktionsanalyse von potentiellen Bindetaschen der Glutamat Dehydrogenase 2 des Malariaerregers *Plasmodium falciparum*

Plasmodium falciparum (*P. falciparum*) ist ein einzelliger Parasit, der von der Stechmücke der Gattung *Anopheles* auf den Menschen übertragen wird und die Tropenkrankheit **Malaria** auslösen kann. Über den Blutkreislauf verteilt sich der Erreger, infiziert die roten Blutkörperchen und bildet dort Proteine, die eine Bindung des Erythrozyten an das Endothel der Blutgefäße bewirken. Die damit verbundenen Mikrozirkulationsstörungen, sowie Deformationen des Erythrozyten tragen zum schwereren Verlauf der durch *P. falciparum* ausgelösten Malaria tropica bei (Dondorp *et al*., 2000).

Malariaparasiten reagieren empfindlich auf oxidativen Stress und schützen sich mit antioxidativ wirkenden Enzymen vor reaktiven Sauerstoffspezies (Becker *et al*., 2004). Die Glutamat Dehydrogenasen (GDH) von *P. falciparum* dienen zusammen mit den Glucose-6-Phosphat Dehydrogenasen als wichtigster Lieferant von NADPH. Dieser Kofaktor wird benötigt, um eines der wichtigsten zellulären Antioxidantien, Gluthation Disulfid zu reduzieren, welches damit in der Lage ist, Sauerstoffradikale zu eliminieren. Die Hemmung der antioxidativ wirkenden Enzyme von *P. falciparum* bieten somit einen günstigen Angriffspunkt für die Entwicklung von Anti-Malaria Therapeutika (Becker *et al*., 2004).

Derzeit ist für *P. falciparum* nur die Struktur der Glutamat Dehydrogenase 1 (*Pf*GDH1) aufgeklärt (PDB: 2bma, Werner *et al*., 2005). Auf Grundlage dieser Struktur wurde ein Homologiemodell der Glutamat Dehydrogenase 2 von *P. falciparum* (*Pf*GDH2) im Arbeitskreis von Professorin Becker (Professur für Biochemie der Ernährung des Menschen, Justus-Liebig-Universität Gießen) erstellt. *PocketPicker* wurde zur Charakterisierung des katalytischen Zentrums und dem Taschenvergleich über ShapeDeskriptoren auf diesem Modell verwendet. Für eine möglichst genaue Beschreibung der enzymatischen Bindetasche mit *PocketPicker* wurde die Berechnung auf einen Ausschnitt des Homologiemodells beschränkt. So wurden sämtliche Atome mit einem Abstand von mehr als 12 Å zum Taschenzentrum aus der Struktur entfernt. Dies war notwendig, da die vergleichsweise lockere Packungsdichte der Proteinstruktur

zu großen tunnelartigen Taschen in der Berechnung durch *PocketPicker* führte (nicht gezeigt).

ShapeDeskriptoren wurden für die größten Taschen der vorbereiteten Strukturen von *Pf*GDH1 (PDB: 2bma) und *Pf*GDH2 (Homologiemodell) berechnet und eine Ähnlichkeitssuche mit 1296 Beschreibern des PDBbind *Refined Set* (Version 2005; Wang *et al.*, 2004; Wang *et al.*, 2005) durchgeführt. Unter den zehn Taschenformen mit größter berechneter Ähnlichkeit wurden für beide Eingaben je zwei Einträge gefunden, deren betrachtete Taschen mit Antimalariawirkstoffen komplexiert waren (Tabelle 9).

Tabelle 9: Vergleich von ShapeDeskriptoren von *Pf*GDH1 und *Pf*GDH2 mit Beschreibern des PDBbind *Refined Set*. Gezeigt ist eine Auswahl von Deskriptoren mit großer Ähnlichkeit zur Eingabe (Annotation durch PDB-Kennung oder GDH Beschreiber). Die Ähnlichkeit der Taschen wird als Euklidische Distanz *d* der ShapeDeskriptoren ausgedrückt, *p* definiert die Position unter den ähnlichsten gefundenen Deskriptoren.

***Pf*GDH1**	***Pf*GDH2**
*Pf*GDH2 (*p*=1, *d*=3463)	*Pf*GDH1 (*p*=1, *d*=3463)
1jqe (*p*=7, *d*=11608)	1jqe (*p*=7, *d*=11845)
1stc (*p*=8, *d*=12063)	2aou (*p*=8, *d*=12624)

Für beide GDH wurde an siebter Position eine Tasche aus dem PDB Eintrag 1jqe gefunden, die mit einem Antimalariawirkstoff komplexiert ist, der als Mepacrin um 1930 von Walter Kikuth bei der I.G. Farbenindustrie in Elberfeld entdeckt wurde und auch unter den Handelsnamen Atebrin® (Mauss & Mietzch, 1933) und Quinacrine bekannt ist. Für *Pf*GDH2 wurde an achter Position eine Tasche gefunden, die mit dem Antimalariamittel Amodiaquine komplexiert ist (PDB: 2aou). Mepacrin und Amodiaquin sind dem Arzneistoff Chloroquin ähnlich und besitzen das gleiche Grundgerüst (Abbildung 47). Alle drei Wirkstoffe besitzen Affinität gegenüber GDH (Jarzyna *et al*., 1997).

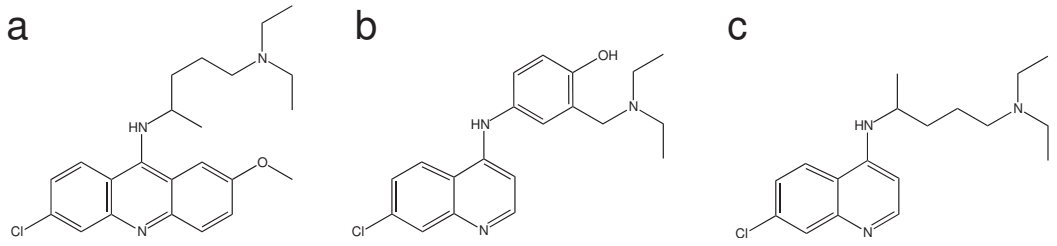


Abbildung 47: Strukturen von (a) Mecaprin, (b) Amodiaquin und (c) Chloroquin.

Für *Pf*GDH1 wurde an achter Position eine Tasche aus dem PDB Eintrag 1stc gefunden, an die das Naturprodukt Staurosporin gebunden ist. Dieses Alkaloid konnte 1977 aus *Streptomyces staurosporeus* isoliert werden (Omura *et al*., 1977) und wirkt als ATP-kompetitiver Inhibitor für Proteinkinasen. Der Wirkstoff bindet dabei mit hoher Affinität aber geringer Selektivität an viele verschiedene Kinasen (Karaman *et al*., 2008).

Die Deskriptoren der Taschen von *Pf*GDH1 (PDB: 2bma) und *Pf*GDH2 (Homologiemodell) wurden für den Ähnlichkeitsvergleich zur Datenbank der 1296 PDBbind Einträge hinzugefügt und gegenseitig als Einträge mit größter Ähnlichkeit bestimmt (p=1, d=3463). Dies unterstreicht die Robustheit der Methode ähnliche Taschenformen in großen Datenbanken aufzufinden. Weiterhin konnte in diesem Projekt gezeigt werden, dass die gefundenen gleichartigen Taschenformen die Bindung ähnlicher Liganden mit gleicher biologischer Funktion ermöglichen. Die Anreicherung der gefundenen Antimalariawirkstoffe in den Taschen der ähnlichsten gefundenen Einträge der *Pf*GDH unterstützt die Anwendbarkeit von *PocketPicker* ShapeDeskriptoren für die Funktionsanalyse von Proteinbindetaschen.

3.6 Strukturelles Alignment und Analyse physikochemischer Eigenschaften von Proteinbindetaschen mit PocketomePicker

In diesem Abschnitt soll der Vergleich von Bindetaschen und ihren Eigenschaften anhand des *PocketShapelets*-Algorithmus dargestellt werden, der in *PocketomePicker* implementiert ist. Die Überlagerung zweier mit *PocketPicker* extrahierten Taschen und die Bewertung (engl. *scoring*) der Ähnlichkeit ihrer Eigenschaften sollen nachfolgend erläutert und diskutiert werden.

3.6.1 Überlagerung von Bindetaschen mit PocketShapelets

Die Beschreibung des geometrischen Aufbaus von Proteinbindetaschen gelingt in *PocketomePicker* über die Charakterisierung der Oberflächenkurvatur einer möglichen Bindestelle durch *PocketShapelets*. Diese paraboloiden Flächenfunktionen

kennzeichnen markante Strukturen der Gauß'schen Oberfläche einer Tasche und ermöglichen eine maschinenlesbare Repräsentation der Taschenform über die Position, Richtung und Kurvatur der berechneten Paraboloide. Die Erkennung von strukturellen Ähnlichkeiten verschiedener Bindetaschen gelingt durch Identifikation von Cliquen auf den Assoziationsgraphen der *PocketShapelets*-Darstellungen (Bron & Kerbosch, 1973). Die Überlagerung der zu Grunde liegenden Bindetaschen in *PocketPicker*-Darstellung erfolgt anhand des Kabsch-Algorithmus (Kabsch, 1976).

Für eine Reduktion der Zahl der identifizierten Cliquen und des einhergehenden Rechenaufwandes werden in *PocketomePicker* nur Cliquen betrachtet, die aus mindestens sieben *PocketShapelets* bestehen. Ferner werden in den Cliquen nur gleichartige Paraboloide der Formen Pfropf ($SI \approx 1{,}0$) und Fläche ($SI \approx 0{,}0$) berücksichtigt. Diese Konventionen wurden nach Analyse der berechneten Überlagerungen durch manuelle Inspektionen festgelegt. Die Betrachtungen zeigten, dass die Fokussierung auf Ausstülpungen und flache Bereiche der Taschenoberfläche eine geeignete Überlagerung erlauben. Die minimale Größe der betrachteten Cliquen ist in der Software vom Benutzer jedoch frei wählbar.

Abhängig von der Größe einer Bindetasche und den gewählten Parametern können für einen Taschenvergleich viele zehntausend Cliquen gefunden werden. Für die Auswahl der Clique von *PocketShapelets*, die eine bestmögliche Überlagerung der Taschen ermöglicht, wurden verschiedene Ansätze verfolgt. Optimierungsstrategien, die eine Minimierung der physikochemischen Eigenschaften oder des geometrischen Abstandes der korrespondierenden *PocketShapelets* verfolgten, führten jedoch lediglich zu lokal guten Alignments. Die Minimierung bevorzugt kleinere Cliquen, die gemittelt über die Anzahl der betrachteten *PocketShapelets* geringere Differenzen aufweisen. Dies kann jedoch zu einem deutlich schlechteren globalen Alignment führen (Abbildung 48).

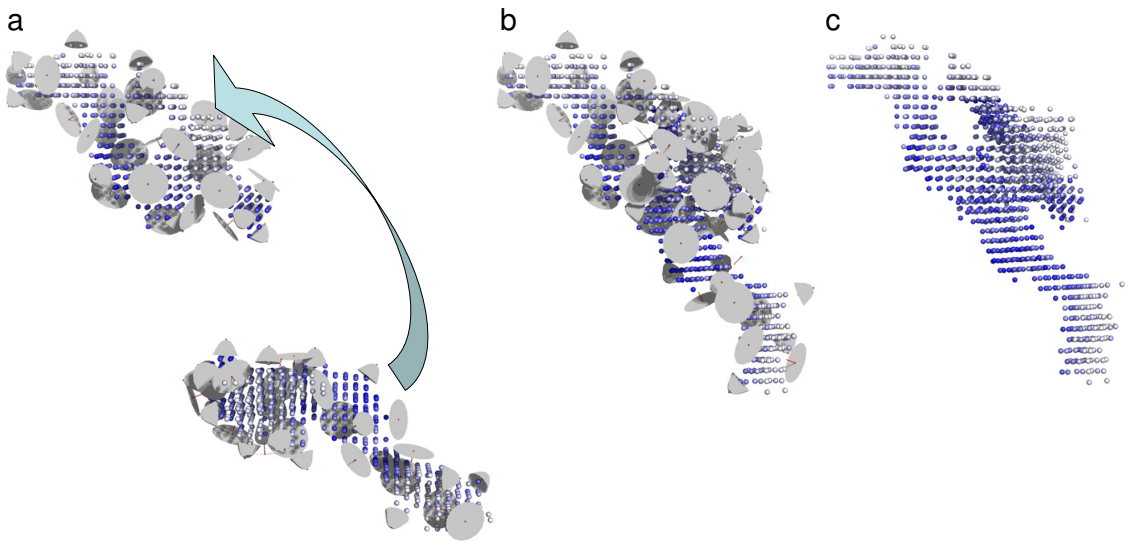


Abbildung 48: Problematik bei der Überlagerung von Taschen durch Minimierung der Abstände zwischen korrespondierenden *PocketShapelets*. a) Extrahierte Bindevolumen für die ligandenfreien Aldose Reduktasen PDB 1iei (oben) und PDB 1us0 (unten). Der Pfeil zeigt die erwünschte Überlagerung von 1us0 auf 1iei an. b, c) Das Alignment anhand kleinerer Cliquen mit geringen euklidischen Abständen zwischen den *PocketShapelets* erzeugt lokal gute Lösungen, ignoriert aber das globale Matching. (Verwendete Parameter: *PocketShapelet*-Radius: 3,0 Å; *PocketShapelet* Distanz-Schwellwert: 2,0 Å; Mindest-Cliquengröße: 7; Cliquenauswahl: Mittlere Distanz der korrespondierenden *PocketShapelets*).

Als ein alternativer Ansatz für die Auswahl einer Clique für das Alignment wird in *PocketomePicker* ein Verfahren verwendet, welches eine größtmögliche Überlagerung der Taschenpunkte für die *PocketShapelets* einer Clique verfolgt. So wird eine Tasche entsprechend jeder gefundenen Clique mit der Zieltasche (engl. *Target*) überlagert und der Grad der Überdeckung bestimmt. Dies gelingt durch Rotation und Translation der Suchtasche (engl. *Query*) in das künstliche Gitter der Zieltasche. Hier werden die Gauß'schen Funktionswerte für jeden Taschenpunkt der Query im Gitter des Targets interpoliert und aufsummiert. Die Transformation, die den größtmöglichen Wert erreicht, wird für die endgültige Überlagerung ausgewählt.

Die resultierende Translations-/Rotationsmatrix ermöglicht die Projektion der Query auf die Zieltasche. Darüber hinaus kann die Information aus dieser Matrix auch für die Rotation der Suchtasche und der in ihr enthaltenen Liganden verwendet werden. Der Vergleich dieser transformierten Strukturen mit dem Target dokumentiert die Leistungsfähigkeit der formbasierten Überlagerung mit *PocketomePicker* (Abbildung 49).

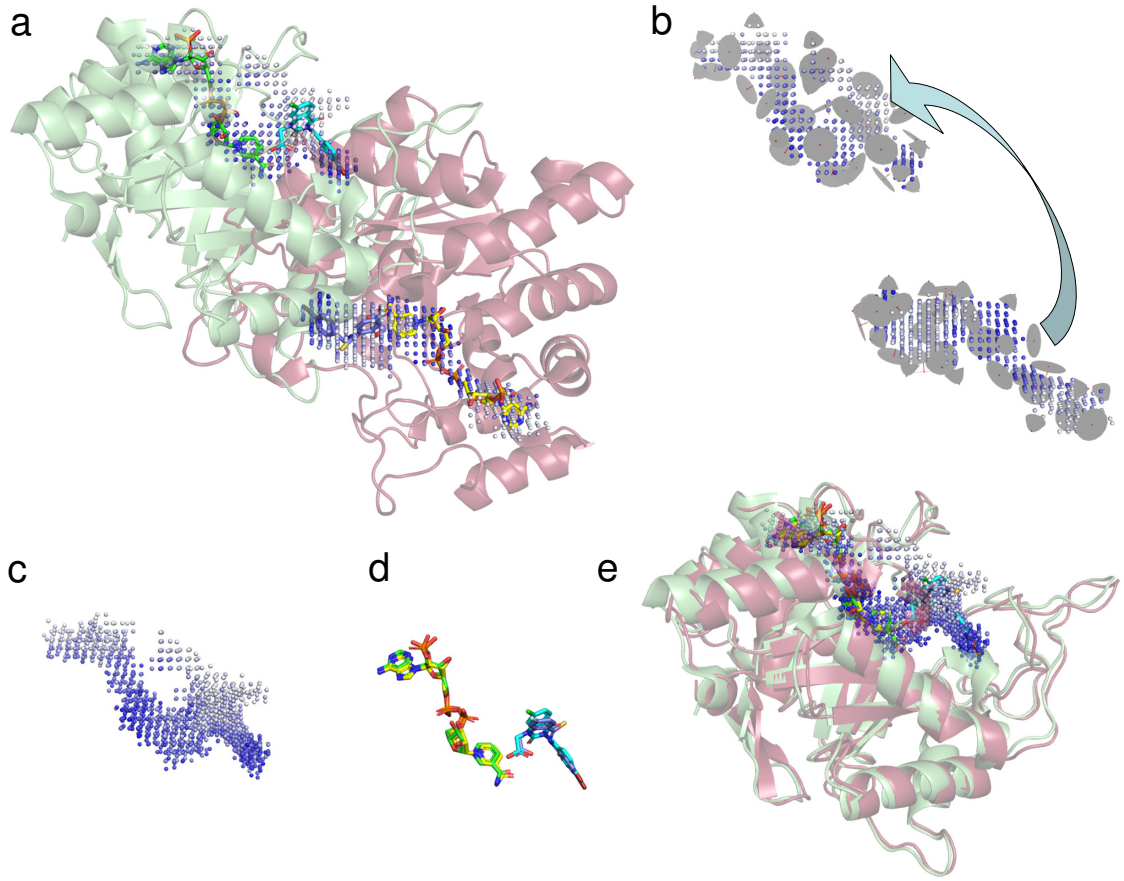


Abbildung 49: Alignment von Proteinbindetaschen in *PocketomePicker*. a) Cartoondarstellung zweier Aldose Reduktasen der PDB-Einträge 1iei (hellgrün mit Liganden Zenarestat (türkis, PDB: zes) und NADP (grün, PDB: NAP)) und 1us0 (hellrot mit Inhibitor IDD594 (blau, PDB: ldt) und NADPH (gelb, PDB: NDP)). Die Strukturen sind in ihren PDB-Koordinaten dargestellt und überschneiden sich. Gezeigt ist die jeweils größte Tasche der ligandenfreien Strukturen. b) Alignment der Taschen mit Hilfe der *PocketShapelets*. c) Überlagerung der Tasche von 1us0 mit der Bindetasche von 1iei. d, e) Überlagerung der Liganden bzw. Proteinstrukturen von 1us0 anhand der berechneten Rotations-/Translationsmatrix mit dem Target 1iei. (Verwendete Parameter: *PocketShapelet*-Radius: 3,0 Å; *PocketShapelet* Distanz-Schwellwert: 1,5 Å; Mindest-Cliquengröße: 7; Cliquenauswahl: Interpolation der Funktionswerte im Gitter des Targets)

Die Auflösung der Repräsentation der Oberflächenkurvatur durch *PocketShapelets* wird durch den Radius r_S der angepassten Paraboloide bestimmt. So bedeuten kleine Werte für r_S eine detaillierte Oberflächenzerlegung in viele *PocketShapelets*, was in großen Cliquen für den Formvergleich resultiert. Die Wahl von r_S bestimmt daher maßgeblich die Genauigkeit der Repräsentation als auch die Rechenzeit für den Ähnlichkeitsvergleich (Tabelle 10).

Tabelle 10: Einfluss des *PocketShapelet*-Radius r_S auf die Anzahl der berechneten Paraboloiden für die ligandenfreie Tasche der Aldose Reduktase PDB 1iei und die für das Alignment mit der Tasche von PDB 1us0 gefundenen Cliquen. (Verwendete Parameter: *PocketShapelet* Distanz-Schwellwert: 1,5; Mindest-Cliquengröße: 7)

verwendeter Radius r_S	Anzahl *PocketShapelets*	gefundene Cliquen
2,5 Å	60	222.510
3,0 Å	38	6.060
3,5 Å	28	49

Für die schnelle Überlagerung von Bindetaschen in großen Datenbanken verwendet *PocketomePicker* standardmäßig einen Radius r_S von 3,5 Å. Andere Werte für r_S können in der Software vom Benutzer als Argument übergeben werden.

Als Erweiterung des formbasierten Überlagerns extrahierter Bindetaschen mit *PocketShapelets* wurde eine Scoringfunktion entwickelt, die eine funktionelle Analyse der Bindevolumen unter Berücksichtigung ihrer physikochemischen Eigenschaften erlaubt. Die Leistungsfähigkeit dieses Ansatzes soll nachfolgend vorgestellt und diskutiert werden.

3.6.2 Funktionsanalyse von Proteinbindetaschen mit PocketShapelets

Wie im vorhergehenden Abschnitt gezeigt wurde, ermöglicht das Kabsch-Alignment auf den Assoziationsgraphen der *PocketShapelets* strukturelle Alignments von gleichartig aufgebauten Taschenformen ähnlicher Größe. Im Unterschied zum globalen Formvergleich zweier Bindetaschen über *PocketPicker* ShapeDeskriptoren ermöglicht die Methode *PocketShapelets* auch ein partielles Matching von Bindevolumen. Dies erlaubt das Alignment kleinerer Bindestellen oder Subtaschen mit gegenüber Teilen größerer Bindetaschen. Dieser technische Vorteil gegenüber globalen Ähnlichkeitsvergleichen auf Grundlage von ShapeDeskriptoren soll am Beispiel der Serinproteasen dargestellt werden. Ferner soll in diesem Abschnitt die Identifikation funktionell ähnlicher Bindetaschen durch die Betrachtung elektrostatischer Potentiale dargestellt werden.

Serinproteasen stellen die am besten untersuchte Klasse von Proteasen dar und ermöglichen die Spaltung von Proteinen und Peptiden. Die katalytische Spaltung gelingt dabei durch den nukleophilen Angriff der deprotonierten Hydroxylgruppe einer Serinseitenkette auf die Amidbindung eines gebundenen Substrats. Dieser Serinrest ist Bestandteil der katalytischen Triade (Ser-His-Trp) einer Serinprotease und namensgebend für die Enzymfamilie. Serinproteasen übernehmen vielfältige Aufgaben im menschlichen Organismus. So wirken Typsin und Chymotrypsin als Verdauungsenzyme, während Thrombin und Faktor Xa wichtige Rollen in der Blutgerinnungskaskade übernehmen.

Die katalytische Triade bildet eine für Serinproteasen charakteristische tiefe Tasche, in der die zu spaltende Seitenkette eines Substrats gebunden wird (Abbildung 50). Diese Tasche wird in der Literatur als **P_1-Tasche** bezeichnet (Kraut, 1977).

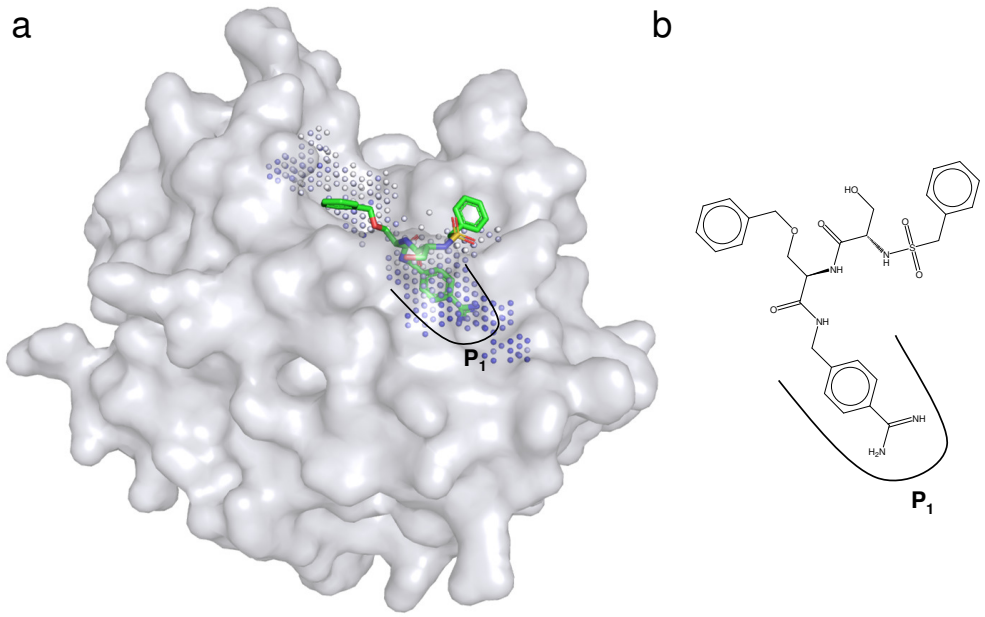


Abbildung 50: Aufbau der P_1-Tasche der Serinprotease Urokinase (PDB: 1vj9) in Seitenansicht. a) Urokinase mit gebundnem Inhibitor (PDB: 5in). Die dunkelblau gefärbten Kugeln zeigen die vergrabene P_1-Tasche an. b) Der Benzamidinrest des Inhibitors bindet in die P_1-Tasche des Enzyms.

In der Sequenz liegen die Aminosäuren der katalytischen Triade (Ser-His-Asp) weit auseinander, so dass sie die enymatisch aktive P_1-Tasche nur durch eine korrekte Faltung ausbilden können. Abweichungen in der Sequenz der verschiedenen Serinproteasen verursachen strukturelle Unterschiede, die in der Oberflächendarstellung

erkennbar sind (Abbildung 51). Zudem führt der halboffene Aufbau der Bindetasche, die die Bindung verschieden großer Polypeptide ermöglicht zu deutlichen Unterschieden in Form der mit *PocketPicker* vorhergesagten Taschen.

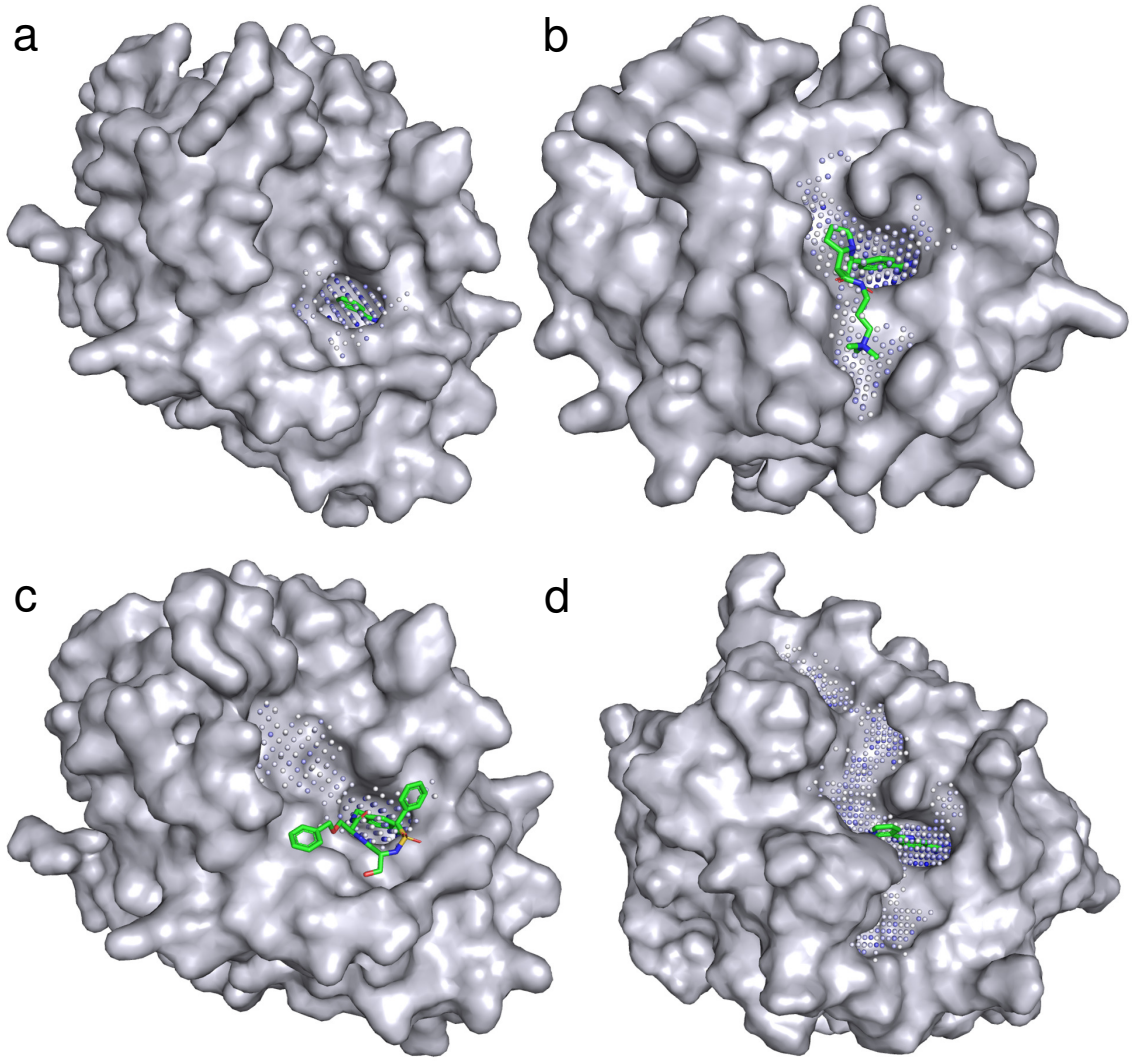


Abbildung 51: Taschenvorhersage für ausgewählte Serinproteasen mit *PocketPicker* in Draufsicht: a) Urokinase Plasminogen Aktivator (PDB: 1f5k) mit Benzamidin (PDB: bam), b) Blutgerinnungsfaktor Xa (PDB: 2bok) mit Inhibitor PDB: 784, c) Urokinase Plasminogen Aktivator (PDB: 1vj9) mit Inhibitor PDB: 5in, d) Thrombin (PDB: 1ghv) mit Inhibitor PDB: 120.

Detaillierte Ansichten der mit extrahierten Bindevolumen für verschiedene Serinproteasen sind in Abbildung 52 gezeigt. Die strukturellen Unterschiede der gezeigten Proteine führen zu deutlichen Größenunterschieden der mit *PocketPicker* extrahierten Taschen. Die Taschenvolumina der gezeigten Beispiele reichen von 207 $Å^3$ (PDB: 1f5k) bis zu einer Größe von 844 $Å^3$ (PDB: 1ghv). Dies erschwert den

Ähnlichkeitsvergleich über ShapeDeskriptoren erheblich. Aus diesem Grund soll an dieser Stelle die Leistungsfähigkeit des Taschenvergleichs über *PocketShapelets* am Beispiel der extrahierten Bindestellen von Serinproteasen gezeigt werden. Der Ähnlichkeitsvergleich gelingt durch die strukturelle Überlagerung der Taschen mit *PocketShapelets* und anschließendem Scoring auf Basis der elektrostatischen Potentiale, die an den Mittelpunkten der PocketShapelets gemessen wurden.

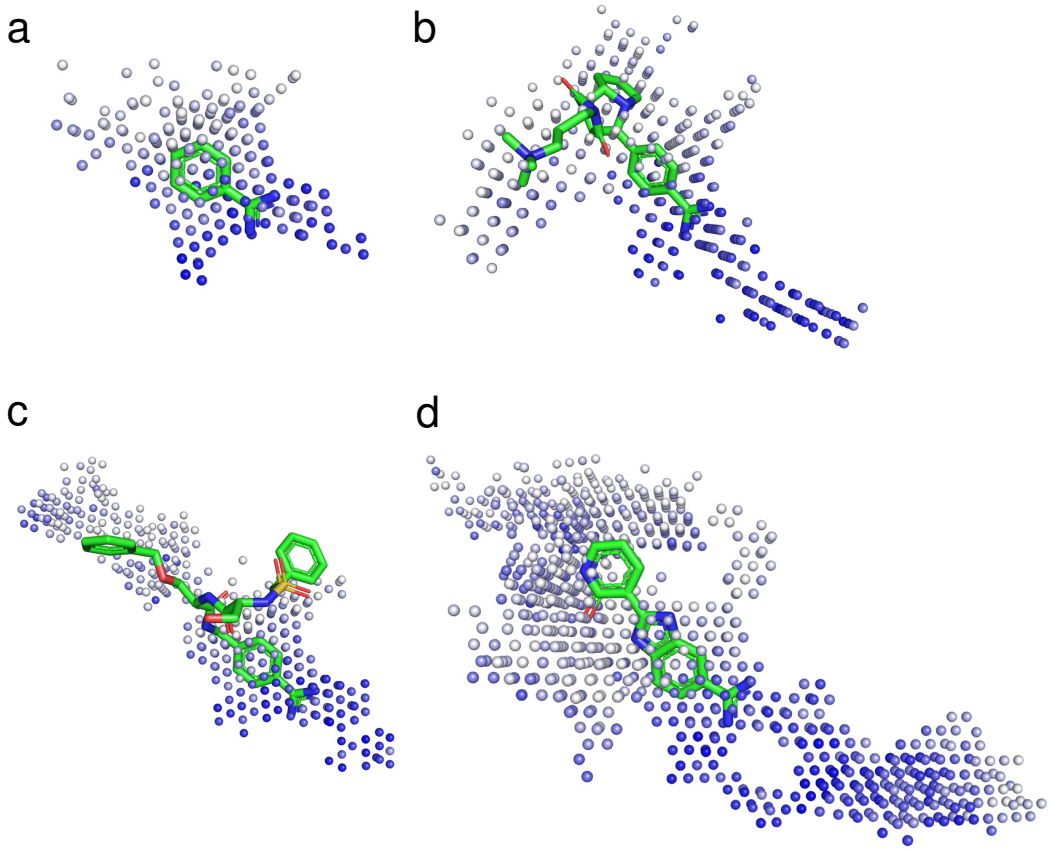


Abbildung 52: Extrahierte Taschenvolumen für verschiedene Serinproteasen mit *PocketPicker*: a) Urokinase Plasminogen Aktivator (PDB: 1f5k) mit Benzamidin (PDB: bam), b) Blutgerinnungsfaktor Xa (PDB: 2bok) mit Inhibitor PDB: 784, c) Urokinase Plasminogen Aktivator (PDB: 1vj9) mit Inhibitor PDB: 5in, d) Thrombin (PDB: 1ghv) mit Inhibitor PDB: 120.

Für die Suche nach Bindetaschen der Serinproteasen wurde die mit *PocketPicker* extrahierte Bindetasche des Blutgerinnungsfaktor Xa (PDB: 2bok) als Suchstruktur verwendet. Der Ähnlichkeitsvergleich wurde gegenüber den mit *PocketPicker* berechneten ligandenbindenden Taschen für die 1300 Proteinstrukturen des PDBbind *Refined Set* (Wang *et al.*, 2004; Wang *et al.*, 2005) durchgeführt. Folgende Parameter wurden für das Matching über *PocketShapelets* gewählt:

- r_{PS} = 3,5 Å (Radien der *PocketShapelets*)
- Δ_{SI} < 0,3 (Matching beschränkt auf gleichartige *PocketShapelets* der Typen „Fläche“ und „Pfropf“)
- $d_{PS} \leq$ 2,0 Å (maximaler Abstand der korrespondierenden *PocketShapelets* zweier Cliquen)
- cs_{min} = 6 (geforderte Mindestgröße der Clique von *PocketShapelets* zweier Taschen für eine Überlagerung)

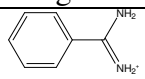
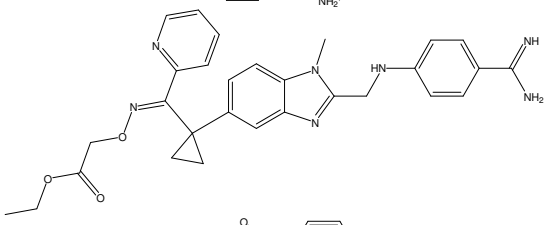
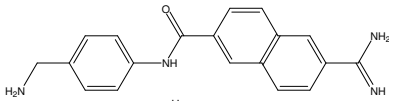
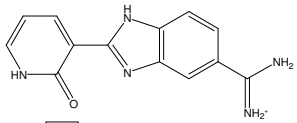
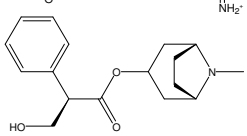
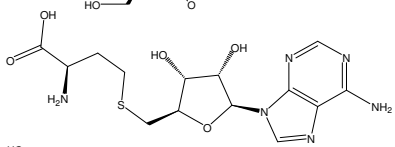
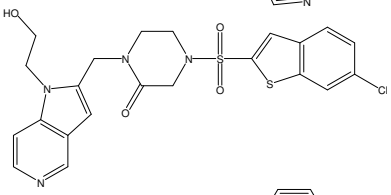
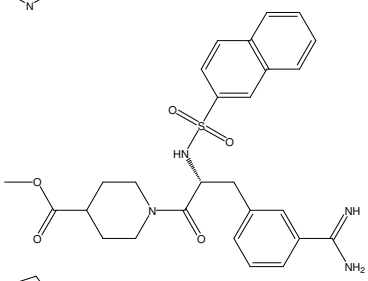
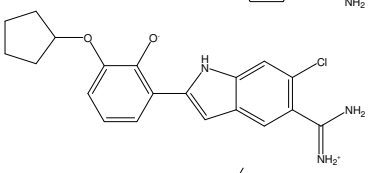
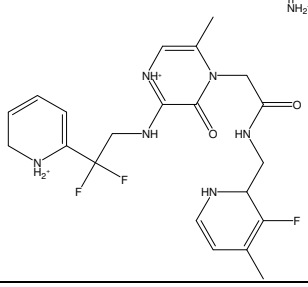
Für die Auswahl der besten Überlagerung zweier Taschen wurde ein rein formbasiertes Kriterium gewählt. So wird für ein Kabsch-Alignment der Suchtasche auf eine Tasche der Datenbank eine Überlagerung anhand der berechneten Transformationsmatrix berechnet. Unter allen Taschenalignments wird, diejenige Pose der Suchtasche ausgewählt, die die größtmögliche Überdeckung mit der Zielstruktur aufweist.
Für die Bewertung (engl. *Scoring*) der Überlagerung unter Berücksichtigung der elektronischen Eigenschaften der überlagerten Taschen wurde eine Scoringfunktion verwendet, die die durchschnittliche Differenz der an den *PocketShapelets* gemessenen elektrostatischen Potentiale betrachtet (Gleichung 13).

$$S = 1 - \frac{\sum_{i=1}^{n} \left| \phi_i - \phi_j \right|}{n}. \qquad (13)$$

Der maximale Score *S* beträgt 1. Die Differenz der elektrostatischen Potentiale ϕ zweier *PocketShapelets i*, *j* wird für alle *n PocketShapelets* berechnet, die einen Abstand von $\leq$ d_{PS} zueinander besitzen.

Die Ergebnisse der Ähnlichkeitssuche sind in Tabelle 11 zusammengefasst.

Tabelle 11: Auflistung der zehn ähnlichsten gefundenen Taschen aus dem PDBbind *Refined Set* für die Suchtasche aus Faktor Xa (PDB: 2bok). Gezeigt sind Scores *S*, E.C. Nummern und die Liganden der Taschen.

	S	Protein	E.C.	Ligand
1f5k	0,84	Urokinase	3.4.21.73	
1g2l	0,78	Faktor Xa	3.4.21.6	
1owh	0,77	Urokinase	3.4.21.73	
1ghv	0,77	Thrombin	3.4.21.5	
2arm	0,76	Phospholipase	3.1.1.4	
1nw7	0,75	DNA-Methyltrans-ferase	2.1.1.72	
1nfx	0,75	Faktor Xa	3.4.21.6	
1k1j	0,73	Trypsin	3.4.21.4	
1o2q	0,73	Beta-Trypsin	3.4.21.4	
1mu8	0,73	Thrombin	3.4.21.5	

Die Ergebnisse zeigen eine klare Privilegierung der Taschen von Serinproteasen mit 8 (13) Einträgen unter den zehn (20) ähnlichsten gefundenen Taschen. Dies kann anhand der E.C. (engl. *Enzyme Nomenclature*) Einträge verifiziert werden (E.C. 3.4.21.-: Hydrolasen (Peptidasen (Serin-Endopeptidasen))).
Unter den zehn ähnlichsten gefundenen Einträgen befinden zwei Proteine, die nicht der Familie der Serinproteasen angehören. Die Taschen einer Hydrolase (Phospholipase, PDB: 2arm) und einer Transferase (Methyltransferase, PDB: 1nw7) erreichen ähnlich gute Elektrostatik-Scores wie andere Serinproteasen. Dies kann im Fall der Methyltransferase dadurch erklärt werden, dass deren Bindetasche ähnliche Oberflächeneigenschaften aufweist, um den gezeigten Liganden S-Adenosyl-Homocystein zu binden. Dieser bindet mit seiner Cysteinseitenkette auch in die Bindetaschen von Hydrolasen (S-Adenosyl-Homocystein Hydrolase, E.C. 3.3.1.1, PDBs: 1a17, 1b3r, 1d4f, 1k0u, 1ky4, 1ky5, 1li4, 1v8b, 1xwf). Allerdings weist die Tasche der Methyltransferase eine geringere Vergrabenheit als die aktive Bindestelle der Serinproteasen. Dies wird in der Methode *PocketShapelets* nicht berücksichtigt und kann als eine Verbesserung des Ansatzes für eine gesteigerte Qualität der Ähnlichkeitsvorhersage diskutiert werden.

Eine weitere Ausnahme unter den zehn ähnlichsten Einträgen bildet die größte Tasche einer Phospholipase (PDB: 2arm). Dieses Protein gehört nicht zur Familie der Serinproteasen, die gefundene Bindestelle besitzt jedoch eine ähnliche Form und Vergrabenheit, wie sie die P_1-Tasche der Suchstruktur aufweist. Die Azabicyclo-Gruppe (ABC-Gruppe) des in der Struktur der Phospholipase (PDB: 2arm) enthaltenen Liganden Atropin (PDB: oin) weist eine andere Orientierung auf, als die gezeigten Serinproteaseinhibitoren (Tabelle 11). So besetzt die ABC-Gruppe nicht den Boden der becherartigen Tasche der Phospholipase, sondern zeigt von diesem weg. ABC-Funktionen von inhibitoren, die die P_2 Tasche von Serinproteasen besetzen sind in der Literatur beschrieben (Sattigeri *et al*., 2008). Es kann daher diskutiert werden, dass die rechnerisch bestimmte Ähnlichkeit der Phospholipasetasche (PDB: 2arm) gegenüber der Bindestelle der Suchstruktur (Faktor Xa, PDB: 2bok) durch die ähnliche Taschenform und die physikochemischen Oberflächeneigenschaften (ABC-Bindung) in P_2-Position hervorgerufen werden.

Das hier gewählte Beispiel der katalytisch aktiven Bindestelle von Serinproteasen stellt schwieriges Target für den Ähnlichkeitsvergleich dar. Die vergleichsweise offene Struktur der Bindetasche, die die Bindung verschiedener Polypeptide und Proteine ermöglichen soll, führt zu Taschen mit deutlich variierendem Volumen in der Vorhersage mit *PocketPicker* (Abbildung 52). Dieses Problem bildet das größte Hindernis beim Vergleich von Bindetaschen über ShapeDeskriptoren. Die Bindevolumen von deutlich unterschiedlicher Größe bedeuten steigende Euklidische Distanzen zwischen den resultierenden Korrelationsdeskriptoren. Die Ergebnisse bestätigen allerdings klar die Leistungsfähigkeit von *PocketShapelets* beim strukturellen Überlagen von Taschen mit unterschiedlichem Volumen und zum Teil verschiedener Form von Subtaschen und Anhängen durch die Implementierung einer Methode zum partiellen Matching (Abbildung 53).

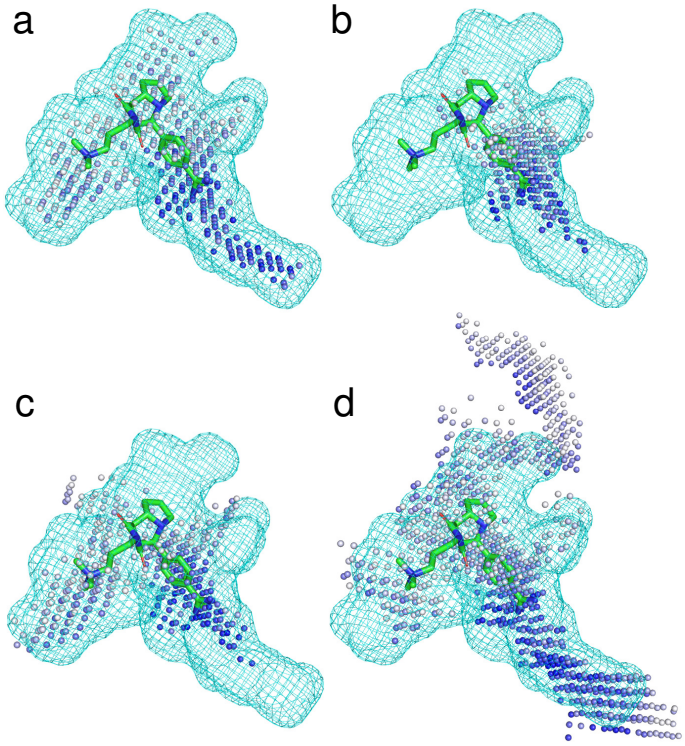


Abbildung 53: Volumenvergleiche extrahierter Bindetaschen von Serinproteasen in der Überlagerung zur Suchstruktur (PDB: 2bok). a) Berechnete Bindestelle der Suchstruktur von Faktor Xa (PDB 2bok). Der Benzamidinrest zeigt die Lage der P_1-Tasche an. Die Gitterdarstellung für die Taschenpunkte der Suchstruktur (PDB: 2bok) erlaubt den Volumenvergleich zu den Taschen von b) PDB: 1f5k, c) PDB: 1g2l und d) PDB: 1ghv.

Die in dieser Arbeit entwickelte Methode *PocketShapelets* erlaubt Vergleiche von funktionell ähnlichen Bindetaschen durch Betrachtung des strukturellen Aufbaus mit

anschließender Bewertung durch eine Scoringfunktion mit Focus auf physiko-chemischen Eigenschaften.

Für die Betrachtung struktureller Eigenschaften von Bindetaschen wurde in dieser Arbeit das Programm *PocketGraph* entwickelt. Nachfolgend soll die Analyse der topologischen Vielfalt von Bindetaschengeometrien durch diese Technik beschrieben werden.

3.7 Untersuchung von Bindetaschentopologien mit PocketGraph

Die in dieser Arbeit entwickelte Methode *PocketGraph* wurde zur Untersuchung der topologischen Vielfalt von ligandenbindenden Taschengeometrien verwendet. Hierzu wurden die Bindetaschen für eine Untermenge des PDBbind *Refined Set* (Wang *et al.*, 2004; Wang *et al.*, 2005) mit *PocketPicker* berechnet und reduzierte Graphdarstellungen mit *PocketGraph* extrahiert. Die Vorgehensweise und Ergebnisse sollen nachfolgend geschildert und diskutiert werden.

3.7.1 Auswahl der Parameter für das GNG-Training in PocketGraph

Die in dieser Arbeit entwickelte Methode zur Charakterisierung des geometrischen Aufbaus von Bindetaschen durch die Methode *PocketGraph* ermöglicht eine reduzierte Beschreibung der mit *PocketPicker* extrahierten Taschenvolumen durch linearisierte Graphrepräsentationen. Die Implementation der Methode durch einen GNG-Ansatz erlaubt die automatisierte Extraktion der Topologie.

Für die Wahl der Parameter wurden Richtwerte aus den Arbeiten von Fritzke (Fritzke, 1994; Fritzke, 1995) betrachtet und für die Anwendung des GNG auf Bindetaschen angepasst. Als kritisch für eine korrekte Wiedergabe der Taschentopologie erwiesen sich das maximale Höchstalter der Kanten eines Graphen α_{max} und der gewählte Mittlere Quadratische Quantisierungsfehler (MSQE) als Abbruchkriterium für die Approximation. Beide Werte wurden anhand visueller Analysen der berechneten Graphen festgelegt. Das zulässige Höchstalter der Kanten wurde empirisch mit α_{max} =

16 gewählt. Dies bewirkt die Herausnahme von Kanten, die nicht regelmäßig erneuert werden. Auf diese Weise werden Kanten entfernt, die zwei Subtaschen miteinander verbinden können, was eine falsche Wiedergabe der Topologie bedeuten würde. Die Wahl des MSQE für die Approximation der Tasche entscheidend für die Auflösung mit der eine Tasche in *PocketGraph* dargestellt wird. Als Abbruchkriterium für das GNG-Training wurde nach Voruntersuchungen ein MSQE von standardmäßig 5 Å gewählt. Der Einfluss dieses Parameters auf die Güte der topologischen Beschreibung einer Tasche ist in Abbildung 54 dargestellt.

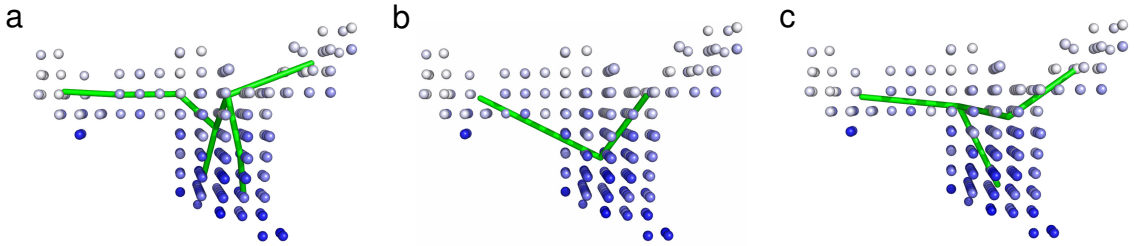


Abbildung 54: Vergleich der Topologiebeschreibung mit *PocketGraph* für die größte berechnete Bindestelle von Calmodulin (PDB: 1qiw). a) Die Approximation der Tasche mit einem Mittleren Quadratischen Quantisierungsfehler (MSQE) von 3 Å bewirkt eine Verästelung des Graphen in der unteren Subtasche. b) Die Berechnung mit einem MSQE von 8 Å bedeutet eine zu grobe Beschreibung des Bindevolumens. c) Die Approximation der Tasche mit einem MSQE von 5 Å als Abbruchkriterium für das GNG-Training ermöglicht die korrekte Beschreibung der Topologie der gezeigten Tasche mit den drei Subtaschen.

Als weitere Parameter wurden folgende Werte für das Training der GNG-Netzwerke in *PocketGraph* festgelegt:

- Einfügung eines neuen Neurons alle $\lambda = 600$ Schritte.
- Sprungweite $\varepsilon_b = 0{,}05$ (als Teil des euklidischen Abstandes) des Siegerneurons zum Eingangssignal.
- Sprungweite $\varepsilon_n = 6\times10^{-4}$ der topologischen Nachbarn des Siegerneurons zum Eingangssignal.
- Skalierungsfaktor $\alpha = 0{,}5$ für die Fehlervariablen der Neurone mit größtem Fehler beim Einfügen eines neuen Neurons.
- Skalierungsfaktor $\beta = 5\times10^{-4}$ für die Fehlervariablen der übrigen Neurone.

Die in *PocketGraph* berechneten Taschengraphen werden als SDF-Dateien exportiert, was eine Visualisierung etwa in *PyMOL* (DeLano, 2002) ermöglicht. Weiter Beispiele für die Topologieerkennung mit *PocketGraph* sind in Abbildung 55 dargestellt.

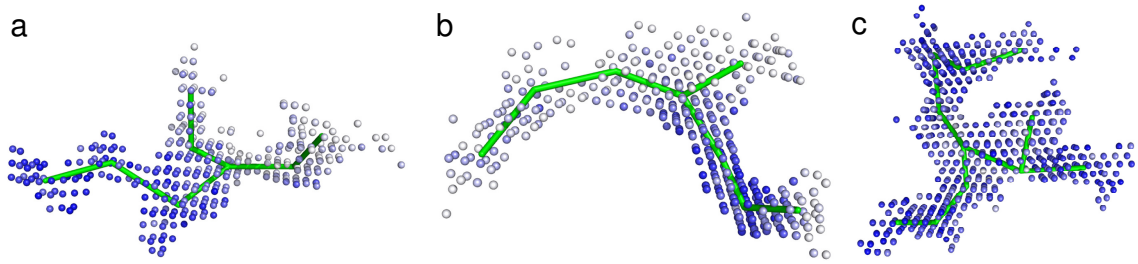


Abbildung 55: Beschreibung des strukturellen Aufbaus von Bindetaschen für die größten Taschen von a) Faktor Xa (PDB: 2boh), b) Matrix-Metallo-Proteinase 8 (PDB: 1jao), c) Oligo Peptid Bindeprotein (PDB: 1b05).

Nachfolgend soll die Anwendung von *PocketGraph* zur Charakterisierung des strukturellen Aufbaus von ligandenbindenden Taschen auf einem repräsentativen Datensatz dargestellt und diskutiert werden.

3.7.2 Analyse der Formenvielfalt und des strukturellen Aufbaus von Proteinbindetaschen

Zur Untersuchung des geometrischen Aufbaus von Proteinbindetaschen wurde ein Datensatz betrachtet, der aus dem PDBbind *Refined Set* (Wang *et al.*, 2004; Wang *et al.*, 2005) abgeleitet wurde und 623 Einträge umfasst. Die betrachtete Datensammlung entspricht dem **Datensatz A**, der in dieser Arbeit für die Analyse der Druggability von Proteinbindetaschen erstellt wurde (Kapitel 3.3.1). Diese Auswahl, die auch nachfolgend als *Datensatz A* bezeichnet werden soll, ermöglicht strukturelle Analysen von 623 ligandenbindenden Taschen, die aus der PDBbind Datenbank (Wang *et al.*, 2004; Wang *et al.*, 2005) abgeleitet sind.

Für die Extraktion der linearen PocketGraphen für Datensatz A wurden die Standardeinstellungen ($\lambda = 600$, $\varepsilon_b = 0{,}05$, $\varepsilon_n = 0{,}0006$, $\alpha = 0{,}5$, $\beta = 0{,}0005$) und ein MSQE von 5 Å als Abbruchkriterium für das GNG-Training gewählt (Kapitel 3.7.1). Die berechneten PocketGraphen wurden als SDF-Dateien exportiert, was eine computergestützte Analyse der Graphentopologie ermöglicht. Für den alignmentfreien Vergleich der extrahierten 623 PocketGraphen wurden topologische Korrelationsdeskriptoren berechnet. Diese kodieren die Häufigkeiten aller Pfadlängen eines Graphen über Distanzen von null bis neun Knoten. Die strukturelle Ähnlichkeit zweier Taschen

lässt sich nun als die Pearson Korrelation *r* der Korrelationsdeskriptoren der korrespondierenden PocketGraphen ausdrücken. Hierbei zeigt eine $r = 1$ eine perfekte positive Korrelation an, die eine optimale Trennung strukturell verschiedener Taschentopologien als disjunkte Cluster anhand der zugehörigen Korrelationsdeskriptoren ermöglicht.

Zur Visualisierung der strukturellen Vielfalt der Geometrien ligandenbindender Taschen wurde die Software Cytoscape (www.cytoscape.org, Shannon *et al.*, 2003) verwendet. PocketGraphen deren topologische Deskriptoren mit einem Koeffizienten korreliert sind, der einen zuvor gewählten Schwellenwert *r* übersteigen, werden in Cytoscape durch Knoten repräsentiert, die durch eine Kante verbunden sind. So sind bei einem Schwellenwert von $r = 0$ sämtliche Individuen in einem vollständig verbundenem Graphen zusammengefasst, während größer werdende Schwellenwerte eine feinere Trennung erzeugen, die graphentheoretisch einem Wald von Graphen entsprechen. Der gewählte Korrelationskoeffizient *r* als Schwellenwert reguliert die Auflösung mit der die PocketGraphen in Clustern zusammengefasst werden. Diese Einteilung lässt sich an der Anzahl der Kanten eines Waldes (bzw. eines vollständig verbundenen Graphen für $r = 0$) einer Darstellung in Cytoscape ablesen (Tabelle 12).

Tabelle 12: Einfluss des gewählten Korrelationskoeffizienten *r* als Schwellenwert für das Clustering der für Datensatz A erzeugten PocketGraphen (MSQE = 5 Å). Kleinere Koeffizienten *r* produzieren Netzwerke mit wenigen stark verbundenen Clustern und hohen Populationsdichten.

$r \geq$	Anzahl Kanten
0	109206
0,1	108238
0,2	106423
0,3	102536
0,4	97390
0,5	91930
0,6	87433
0,7	79206
0,8	66442
0,9	54856
0,98	18087
0,99	13755
1,0	6586

Die mit Cytoscape generierten Netzwerke geben Aufschluss über die Vielfalt der Taschengeometrien und die Relationen einzelner Cluster untereinander. Abbildung 56 zeigt eine grobe Klassifikation der extrahierten PocketGraph-Topologien für Datensatz A.

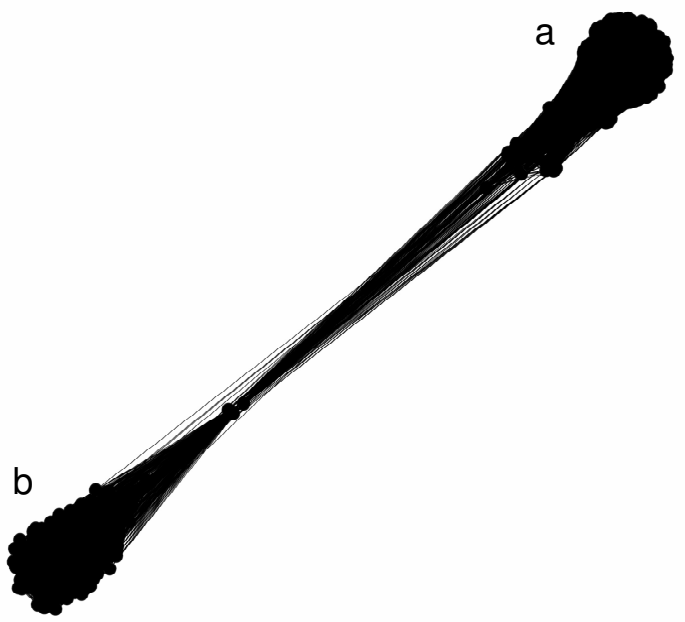


Abbildung 56: Analyse der strukturellen Ähnlichkeit von Taschengeometrien anhand von topologischen Korrelationsdeskriptoren für 623 PocketGraphen (Datensatz A, MSQE = 5 Å, *r* = 0,05). Die Abbildung zeigt die Einordnung von komplexen PocketGraphen (> 7 Neurone, (a)) und einfachen PocketGraphen (≤ 7 Neurone, (b)).

Die Anwendung eines niedrigen Korrelationsmaßes $r = 0{,}05$ als Schwellenwert erzeugt eine grobe Klassifikation als einen verbundenen Graphen, in dem zwei Domänen mit 209 (Abbildung 56 a)) bzw. 410 (Abbildung 56 b)) Individuen erkennbar sind. Diese erste Analyse zeigt, dass es etwa doppelt so viele Taschentopologien mit einfacherem Aufbau (≤ 7 Neurone) als komplexer gestaltete Bindestellen (> 7 Neurone) gibt.

Für eine genauere Betrachtung des geometrischen Aufbaus von ligandenbindenden Taschen und ihrer strukturellen Vielfalt wurde ein weiteres Clustering mit einem Pearson Korrelationskoeffizient von $r = 0{,}98$ als Diskriminante für die 623 Korrelationsdeskriptoren (Datensatz A) durchgeführt (Abbildung 57).

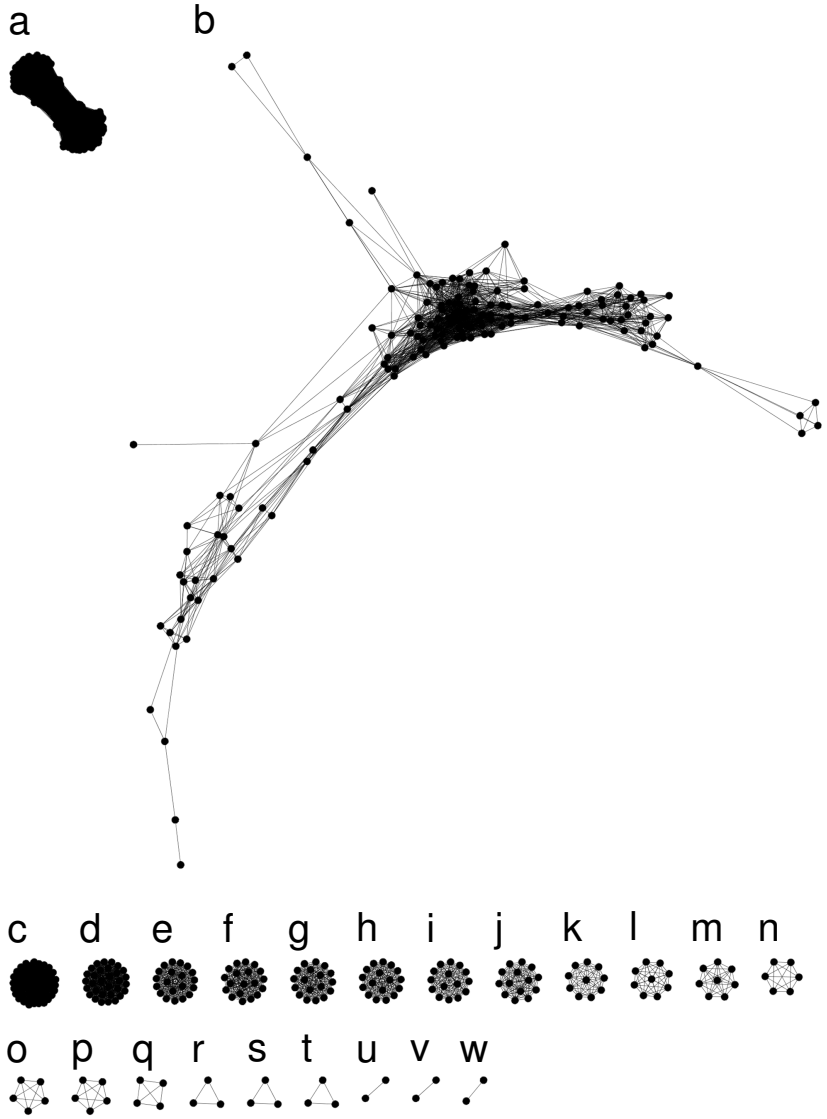


Abbildung 57: Analyse der strukturellen Ähnlichkeit von Taschengeometrien anhand von topologischen Korrelationsdeskriptoren für 623 PocketGraphen (Datensatz A, MSQE = 5 Å, *r* = 0,98). Die Anwendung eines stringenten Korrelationsmaßes (*r* = 0,98) erzeugt eine feine Einteilung der PocketGraphen in Cluster a) bis w), die absteigend nach Größe sortiert sind. Individuen sind durch Kanten verbunden, wenn die topologischen Korrelationsdeskriptoren ihrer PocketGraphen mit mindestens *r* = 0,98 korreliert sind. Singletons sind nicht gezeigt.

Ein höher gewählter Korrelationskoeffizient von $r = 0{,}98$ erzeugt eine deutlich feinere Einteilung der topologischen Deskriptoren in 23 disjunkte Cluster. Die Zuordnung der PocketGraphen zu den einzelnen Clustern a) bis w), sowie deren Populationsdichte ist in Abbildung 58 dargestellt. Für eine vereinfachte Darstellung sind die PocketGraphen als zweidimensionale Strichzeichnungen dargestellt, die lediglich die Größe und Topologie der PocketGraphen repräsentieren. Längen und Winkel der Kanten, die die Neurone des GNG verbinden, sind in dieser Darstellung nicht berücksichtigt.

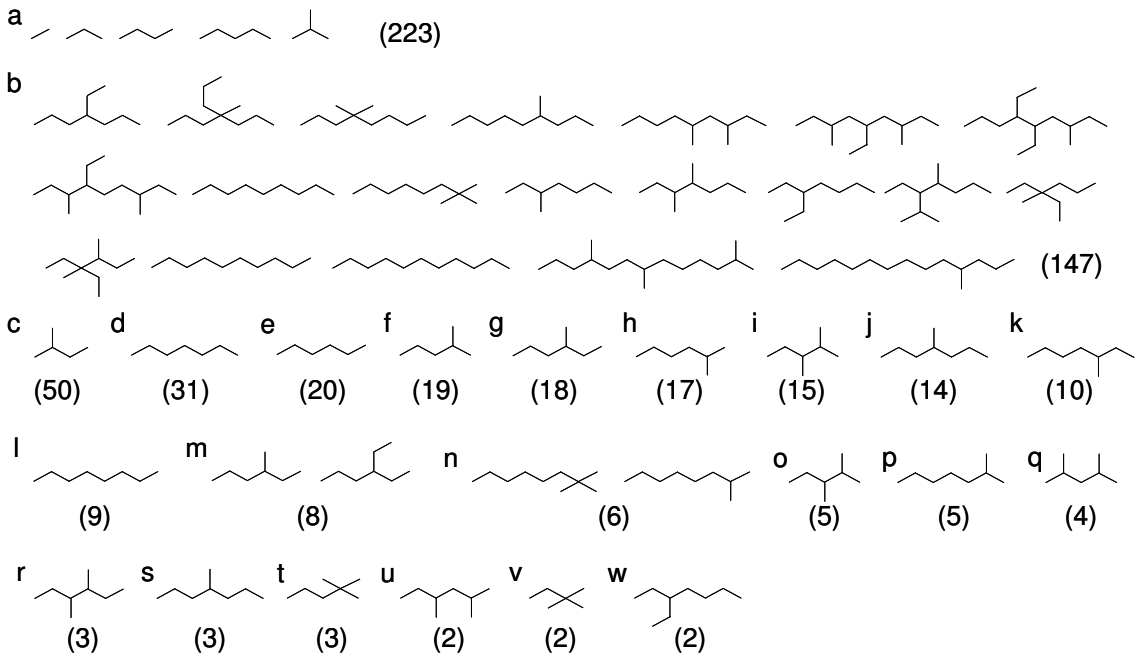


Abbildung 58: Darstellung der topologischen Vielfalt für die ligandenbindenden Taschen des Datensatzes A. Gezeigt sind zweidimensionale Repräsentationen der PocketGraphen aus dem Clustering der Korrelationsdeskriptoren mit Pearson Korrelationskoeffizient *r* = 0,98. Die Bezeichner a) bis w) entsprechen den Cluster in Abbildung 57, die Zahlen in Klammern geben die Populationsdichte der Cluster an.

Die stringentere Klassifikation mit Pearson Korrelationskoeffizient $r = 0{,}98$ sortiert die verschiedenen Taschentopologien in 23 disjunkte Cluster ein, wobei 19 der 23 Gruppen lediglich ein Gerüst enthalten. Die Populationsdichte der einzelnen Cluster ist in Abbildung 59 als Histogramm dargestellt. Die Ergebnisse zeigen, dass mehr als ein Drittel der untersuchten Ligandenbindetaschen einen strukturell einfachen Aufbau aufweisen und im größten Cluster vereinigt werden. Der zweitgrößte Cluster umfasst die größten und toplogisch anspruchvollsten Taschengeometrien des betrachteten Datensatzes. Dieser Cluster besitzt ferner die größte Variabilität aller 23 Gruppierungen, die durch die gesteigerte Größe der enthaltenen PocketGraphen bedingt ist. So verändern strukturelle Unterschiede in größeren Strukturen das Aussehen eines resultierenden topologischen Deskriptors weniger stark als im Fall kleinerer PocketGraphen. Zudem ist die Beschreibung eines Graphen durch einen topologischen Deskriptor nicht eindeutig. So können Symmterieeffekte in größer werdenden Strukturen zu sehr ähnlichen Deskriptoren für PocketGraphen mit leicht unterschiedlichem Aufbau führen.

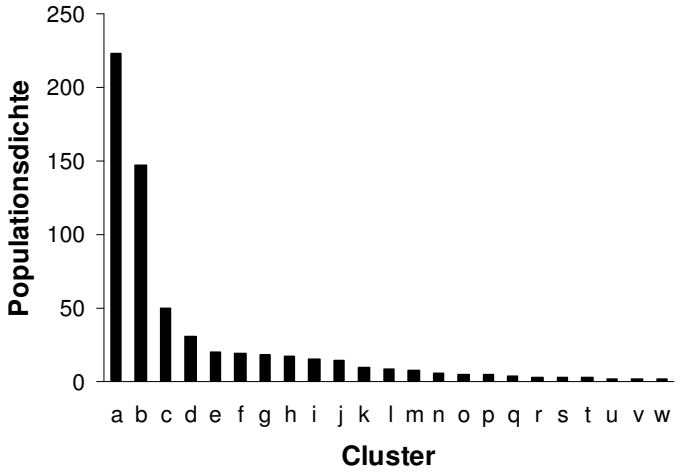


Abbildung 59: Populationsdichte der einzelnen Cluster für die strukturelle Klassifikation der PocketGraphen (MSQE = 5 Å) für Datensatz A und Pearson Korrelation *r* = 0,98 als Schwellenwert.

Der betrachtete Datensatz betrachtet die ligandenbindenden Taschen von 623 Komplexen, die aus dem PDBbind *Refined Set* (Wang *et al.*, 2004; Wang *et al.*, 2005) abgeleitet wurden. Diese Sammlung umfasst eine Teilmenge der Proteinkomplexe aus der PDB Datenbank (Bermann *et al.*, 2000) für die Affinitätsdaten bekannt sind. Darüber hinaus müssen die Einträge des PDBbind *Refined Set* eine Reihe von Qualitätsmerkmale erfüllen, etwa eine Auflösung der Kristallstruktur von mindestens 2,5 Å. Die in dieser Studie betrachteten PocketGraphen ermöglichen somit eine Beschreibung des strukturellen Aufbaus von Bindestellen aus einer repräsentativen Auswahl von Protein-Ligandenkomplexen von hoher Qualität.

Die Untersuchung des topologischen Aufbaus durch *PocketGraph* beschreibt die strukturelle Vielfalt ligandenbindender Taschenformen. Die Ergebnisse geben Aufschluss über vorherrschende strukturelle Muster und Komplexität des geometrischen Aufbaus von Bindestellen. Während etwa ein Drittel der betrachteten Taschen durch einfache PocketGraphen mit wenigen Neuronen approximiert werden können, beschreiben die Strukturgraphen mehrerer Cluster einen gestreckten Aufbau mit einer oder zwei abzweigenden Subtaschen. Diese Beobachtung zeigt einerseits die Variabilität wiederkehrender Muster, führt aber gleichzeitig zu der Feststellung, dass die Formenvielfalt der Taschengeometrien offenbar begrenzt ist.

Der begrenzten topologischen Variabilität des „Pocketoms“ steht somit eine übermächtige Vielfalt an Proteinfunktionen mit den Spezifitäten und Selektivitäten der funktionstragenden Bindetaschen gegenüber. Dies unterstützt die Vermutung, dass diese Phänomene der Ligandenbindung nicht allein über die Struktur von Bindetaschen vermittelt werden können. Eine Analyse der Taschentopologien in exponierten Knoten im Cluster (b) mit größter struktureller Variabilität (Clustering mit r = 0,98; siehe Abbildung 58) unterstreicht, dass eine Klassifikation von Funktion oder Spezifität von Bindetaschen nicht anhand struktureller Merkmale möglich ist (Abbildung 60).

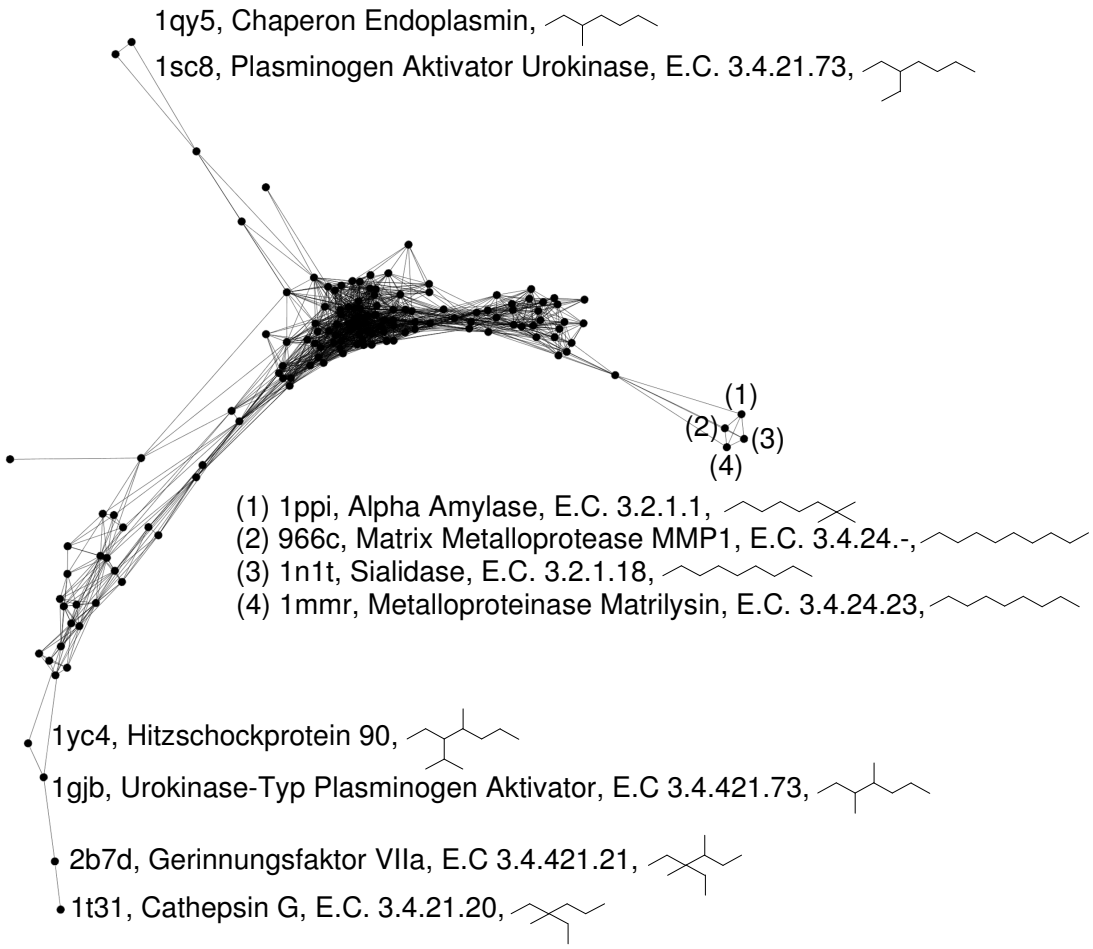


Abbildung 60: Beschreibung der Funktion und E.C. Klassifikation (bei Enzymen) für exponierte Knoten des Clusters mit größter struktureller Variation der PocketGraphen (Clustering mit r = 0,98).

Eine vollständige Analyse der Funktionen der für das Clustering betrachteten Taschen des Datensatzes A ist im Anhang dargestellt (Tabelle A4). Eine eindeutige Zuordnung ähnlicher Proteinfunktionen der ligandenbindenden Taschen zu den disjunkten Clustern

ist nicht zu erkennen. Lediglich Bindestellen, deren Funktion direkt abhängig von der Struktur topologisch anspruchsvoller Liganden ist, können in diesem Ansatz korrekt erkannt werden. So werden 24 von 25 Einträgen des Datensatzes, deren Funktion als „Peptidtransport“ oder „Peptidbindeprotein“ annotiert sind dem Cluster mit größter topologischer Variabilität (Cluster (b)) zugewiesen (Tabelle A4).

Für die Abschätzung der Funktion von Proteinen ist die Topologie der ligandenbindenden Taschen ein notwendiges, jedoch kein hinreichendes Kriterium. Dennoch erlaubt die hier vorgestellte Technik einen Einblick in die Vielfalt des geometrischen Aufbaus von Taschenformen und die Häufigkeit mit der diese auftreten. Eine genauere Analyse der funktionsbestimmenden Merkmale von Bindestellen kann durch Berücksichtigung von physikochemischen Eigenschaften der Taschen vorgenommen werden. Dies kann etwa durch die Projektion der vorherrschenden Eigenschaften auf die Knoten der PocketGraphen realisiert werden. Der folgende Abschnitt soll weitere Anwendungsmöglichkeiten der Software *PocketGraph* vorstellen.

3.7.3 Anwendung von PocketGraphen zur Repräsentation von Bindestellen

Im Unterschied zu ligandenbasierten Ansätzen, die sich meist auf eine Darstellung von Molekülen durch ihre Strukturformeln stützt, gibt es für rezeptorbasierte Verfahren keine vereinheitlichte Beschreibung der betrachteten Bindevolumen. Während manche Methoden eine minimalistische Repräsentation von Ligandenbindestellen durch deren geometrische Zentren verfolgen, gibt es eine Vielzahl unterschiedlicher Ansätze, die komplexere Beschreibungen nutzen. So werden Formen und Eigenschaften von Bindetaschen über gitterartige Raster (Caron *et al*., 2009), Gruppen virtueller Sphären (Laurie & Jackson, 2005; Weisel *et al*., 2007), Oberflächenausschnitte (Laskowsi *et al.*, 1995; Hendlich *et al*., 1997) oder Bindehüllen (An *et al*., 2004) charakterisiert. In dieser Arbeit wurden mit *PocketPicker* und *PocketShapelets* Konzepte entwickelt, die eine technisch aufwändige Beschreibung von Bindestellen über virtuelle Sphären oder hyperbole Paraboloide implementieren. Diese Techniken ermöglichen detaillierte

Analysen von Bindetaschenformen über Autokorrelationsdeskriptoren, sowie strukturelle Überlagerungen durch Kabsch Alignments (Kabsch, 1976).

Zwischen der Beschreibung einer Tasche durch ihr geometrisches Zentrum und den deutlich anspruchsvolleren Darstellungen besteht eine konzeptionelle Lücke. *PocketGraph* füllt diesen Raum und bietet eine schlichte Beschreibung der Topologie einer Bindestelle über Strukturzeichnungen, die durch die Verwendung von GNG Netzwerken automatisch generiert werden. Die reduzierte Charakterisierung von Bindetaschen durch Graphen aus Punkten und Kanten weist starke Ähnlichkeit mit der Darstellung von kleinen Molekülen durch Strukturformeln auf. Tatsächlich erlaubt diese Repräsentation von Bindevolumen durch *PocketGraphs* die Anwendung von Programmen, die zur Beschreibung von Liganden verwendet werden. Der hier vorgestellte Ansatz macht die Welt der Bindetaschen zugänglich für die Methoden aus der Welt der ligandenbasierten Techniken. Dies erlaubt eine starke Beschleunigung von Operationen auf den bisher komplexen Beschreibungen von Bindetaschen durch computergestützte Verfahren und ligandenbasierte Methoden. So ermöglicht die Graphdarstellung von Bindestellen durch *PocketGraph* Anwendungsmöglichkeiten für künftige Arbeiten, so etwa Ähnlichkeitsvergleiche von Taschen über topologische Korrelationsvektormethoden für ligandenbasiertes Virtuelles Screening (Schneider *et al.*, 1999; Fechner *et al.*, 2003, Tanrikulu *et al.*, 2007).

4 Ausblick

In dieser Arbeit wurden neue Methoden entwickelt, die sich mit dem Vergleich von Proteinbindetaschen, ihren Formen und Eigenschaften beschäftigen. Für die Ähnlichkeitssuche über Autokorrelationsdeskriptoren sowie für die Abschätzung der Druggability einer Tasche erwies sich die Betrachtung der Vergrabenheit als ein wichtiges Konzept für eine erfolgreiche Klassifikation. Die Betrachtung der Vergrabenheit findet bisher keine Anwendung strukturelle Taschenalignments mit *PocketShapelets*. Eine Anwendung dieses Prinzips verspricht eine Unterscheidung von Taschen mit ähnlicher Struktur und ähnlichen Oberflächeneigenschaften, die sich jedoch in ihrer Zugänglichkeit unterscheiden.

Des Weiteren bietet die Verwendung des Konzepts Wachsender Neuronaler Netze eine Reihe von Anwendungsmöglichkeiten zukünftiger Arbeiten. So kann die Darstellung der Topologie von Bindetaschen durch PocketGraphen etwa durch die Projektion physikochemischer Eigenschaften oder Vergrabenheitswerten auf die Knoten der Graphen erweitert werden. Die Repräsentation von Bindevolumen durch topologische Strukturgraphen verspricht deutliche Rechenbeschleunigungen für den Vergleich von Bindetaschen mit Laufzeiten, die ähnlich derer ligandenbasierter Ansätze sind.

Die Approximation von Bindevolumen durch unüberwacht lernende Neuronale Netzwerke bietet einen interessanten Ansatzpunkt für ein neues Verfahren zum Protein-Liganden Docking. Die Kodierung der Heteroatome eines Liganden als Neurone eines nicht Wachsenden Neuronalen Netzwerkes mit einer Netzwerktopologie entsprechend den Atombindungen des zu dockenden Moleküls ermöglicht ein selbsttätiges Docking der Struktur. Dieses Docking benötigt allerdings eine möglichst präzise Beschreibung des Bindevolumens, sowie eine Methode zur Korrektur von Bindungswinkeln und Bindungslängen. Die Bewertung der Dockingpose kann durch eine Scoringfunktion erfolgen, die die Ausrichtung der Wechselwirkungszentren von Ligand von Rezeptor zueinander betrachtet.

Alle in dieser Arbeit vorgestellten Verfahren basieren auf der Identifikation und der Beschreibung von Bindetaschen durch das Programm *PocketPicker*. Der Erfolg der neu entwickelten Methoden ist somit direkt abhängig von der Qualität der Vorhersage durch *PocketPicker*. Dieses Programm verwirklicht ein rein geometrisches Konzept zur

Identifikation von Bindetaschen, das jedoch den Erfolg der meisten etablierten Vorhersagemethoden übertrifft. Eine weitere Steigerung der Leistungsfähigkeit von *PocketPicker* ist durch die Erweiterung des geometrischen Ansatzes um die Betrachtung physikochemischer Eigenschaften, etwa der Lipophilie möglich.

5 Literaturverzeichnis

A

Ajay, A., Walters, W. P., Murcko, M. A., Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.*, 1998, **41**, 3314–3324.

An, J., Totrov, M., Abagyan, R., Comprehensive Identification of "Druggable" Protein Ligand Binding Sites. *Genome Inf.*, **15**, 2004, 31-41.

Atiyah, M., Sutcliffe, P., The geometry of point particles. *Royal Soc. London Proc. Ser. A Math. Phys. Eng. Sci.*, **215**, 2002, 1089-1115.

B

Bäck, T., Schwefel, H.P. An Overview of Evolutionary Algorithms for Parameter Optimization., *Evolut. Comp.*, **1**, 1993, 1-23.

Baker, N., Holst, M., Wang, F., Adaptive Multilevel Finite Element Solution of the Poisson-Boltzmann Equation II. Refinement at Solvent-Accessible Surfaces in Biomolecular Systems, *J. Comput. Chem.*, 2000, **21**, 1343-1352.

Balducci, R., Pearlman, R.S., Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 822-831.

Becker, K., Tilley, L., Vennerstrom, J.L., Roberts, D., Rogerson, S., Ginsburg, H., Oxidative stress in malaria parasite-infected erythrocytes: host-parasite interactions, *Int. J. Parasitol.*, 2004, **34**, 163-189.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., The Protein Data Bank, *Nucl. Acids Res.*, **28**, 2000, 235-242.

Binkowski, T., Naghibzadeh, S., Liang, J., $CAST_p$: Computed Atlas of Surface Topography of proteins. *Nucl. Acids Res.*, 2003, **31**, 3352-3355.

Blinn, J.F., A Generalization of Algebraic Surface Drawing, *ACM Transactions on Graphics*, 1982, **1**, 235-256.

Böhm, H.J., The computer program LUDI: A new method for the de novo design of enzyme inhibitors, *J. Comput-Aided Mol. Des.*, 1991, **6**, 61-78.

Böhm, H.J., Klebe, G., Kubinyi, H., *Wirkstoffdesign*, Spektrum Verlag, Heidelberg, **2002**, 96 ff.

Böhm, H.J., Schneider, G., *Protein-Ligand Interactions*, Wiley-VCH, Weinheim, **2003**, 4f.

Brady, G.P., Stouten, P.F.W., Fast prediction and visualiuzation of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, 2000, **14**, 383-401.

Brint, A.T., Willet, P., Algorithms for the Identification of Three-Dimensional Maximum Common Substructures, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 152-158.

Bron, C., Kerbosch, J., Algorithm 457: Finding all cliques of an undirected graph. *Communications of the ACM*, 1973, **16**, 575-577.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, *J. Comp. Chem.*, 1983, **4**, 187-217.

Broto, P., Moreau, G., Vandyke, C., Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.*, 1984, **19**, 66-70.

Burges, C.J.C., A tutorial on support vector machines for pattern recognition. *Knowl. Discov. Data Min.*, 1998, **2**, 121-167.

C

Cai, X., Langtangen, H.P., Moe, H., On the performance of the Python programming language for serial and parallel scientific computations. *J. Sci. Prog.*, 2005, **13**, 31-56.

Caron, G., Nurisso, A., Ermondi. G., How to Extend the Use of Grid-Based Interaction Energy Maps from Chemistry to Biotopics. *ChemMedChem*, 2009, **4**, 29-36.

Cheatham, T.E., Cieplak, P., Kollman, P.A., A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, 1999, **16**, 845-862.

Colman, P.M., Structure-based drug design, *Curr. Opin. Struct. Biol.*, 1994, **4**, 868-74.

Collins, F.S., Finishing the euchromatic sequence of the human genome. *Nature*, 2004, **431**, 931-945.

Conolly, M.L., Analytical molecular surface calculation, *J. Appl. Crystallogr.*, 1983, **16**, 548-558.

Conticello, S.G., Harris, R.S., Neuberger, M.S., The Vif Protein of HIV Triggers Degradation of the Human Antiretroviral DNA Deaminase APOBEC3G, *Curr. Biol.*, 2003, **13**, 2009-2013.

Convey, J.H., Sloane, N.J.A., *Sphere Packings, Lattices and Groups*, 2nd ed., Springer-Verlag, New York, **1988**, 25.

Corey, R., Pauling, L., Molecular models of amino acids, peptides and proteins, *Rev. Sci. Instr.*, 1953, **24**, 621.

Cortes, C., Vapnik, V., Support-vector networks, *J. Mach. Learn.*, 1995, **20**, 273-297.

Cruciani, G., Pastor, M., Guba,W., VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci*., 2000, **11**, 29–39.

D

DeLano, W.L., The PyMOL Molecular Graphics System (2002) DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org.

Dingwall, C., Ernberg, I., Gait, M.J., Green, S.M., Heaphy, S., Karn, J., Lowe, A.D., Mohinder, S., Skinner, M.A., Valerio, R., Human immunodeficiency virus 1 tat protein binds trans-activation-response region (TAR) RNA *in vitro*, *Proc. Natl. Acad. Sci*., 1989, **86**, 6925-6929.

Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A., PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. *Nucl. Acids Res.*, **32**, 2004, W665-W667.

Dondorp, A.M., Kager, P.A., Vreeken, J., White, N.J., Abnormal Blood Flow and Red Blood Cell Deformability in Severe Malaria, *Parasitol. Today*, 2000, **16**, 228-232.

Drews, J., Reyser, S., The role of innovation in drug development. *Nature Biotechnol.*, 1997, **15**, 1318-1319.

Duncan, B.S., Olson, A.J., Approximation and characterization of molecular surfaces, *Biopolymers*, **33**, 319-229.

Dunn, F., Parberry, I., 3D Math primer for Graphics and Game Development, Wordware Publishing, Plano, **2002**, 331.

Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G., Reoptimization of MDL keys for use in drug discovery, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273-1280.

E

Ehrlich, P., Zur Kenntnis der Antitoxinwirkung, *Fortschr. d. Med.*, 1987, **2**, 41-43.

Ehrlich, P., Über den jetzigen Stand der Chemotherapie, *Ber. Dt. Chem. Ges.*, 1909, **42**, 17-47.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M., The Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies, *Comp. Chem.*, 1995, **16**, 273-284.

Exner, T.E., Keil, M., Brickmann, J., Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *J. Comp. Chem.*, 2002, **23**, 1176-1187.

F

Fechner, U., Franke, L., Renner, S., Schneider, P., Schneider, G., Comparison of correlation vector methods for ligand-based similarity searching, *J. Comp-Aid. Mol. Des.*, 2003, **17**, 687-698.

Feng, L., Soon, S.H. An effective 3D seed fill algorithm. *Comp. Graph.*, 1998, **22**, 641-644.

Felsenstein, J. PHYLIP – Phylogeny Interference Package (Version 3.2), *Cladistics*, 1989, **5**, 164-166.

Fischer, E., Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dt. chem Ges.*, 1894, **27**, 2985-2993.

Fischer, H., Polikarpov, I., Craievich, A.F., Average protein density is a molecular-weight-dependent function. *Prot. Sci.*, 2004, **13**, 2825-2828.

Fourment, M., Gillings, M.R., A comparison of common programming languages used in bioinformatics, *BMC Bioinf.*, 2008, **9**:82.

Fritzke, B., Growing cell structures - a selforganizing network for unsupervised and supervised learning, *Neural Networks,* 1994, **7**, 1441-1460.

Fritzke, B., A Growing Neural Gas Network Learns Topologies. *Neural Networks*, in Tesauro, G., Touretzky, D.S., Leen, T.K., editors, *Advances in Neural Information Processing Systems*, **1995, 7**, MIT Press, Cambridge MA, 625-632.

Fuhrmann, G.F., Allgemeine Toxikologie für Chemiker – Einführung in die Theoretische Toxikologie, 1994, Teubner, Stuttgart, 104-105.

Fujita, T., Iwasa, J., Hansch, C., A New Substituent Constant, pi, Derived From Partition Coefficients. *J. Am. Chem. Soc.*, 1964, **86**, 5175-5180.

G

Gardiner, E.J., Willet, P., Artymiuk, P.J., Graph-Theoretic Techniques for Macromolecular Docking. *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 273-279.

Gasteiger, J., Engel, T., *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, **2003**, 31ff.

Gerhards, L., Lindenberg, W., Clique detection for nondirected graphs: Two new algorithms, *Computing*, 1979, **21**, 295-322.

Ghose, A.K., Crippen, G.M., Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.*, 1986, **7**, 565-577.

Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J., Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem.*, 1998, **102**, 3762-3772.

Gibbs, J.W., On the equilibrium of heterogeneous substances, *Trans. Conn. Acad. Arts Sci.*, 1878, **3**, 343-524.

Glaser, F., Rosenberg, Y., Kessel, A., Tal, P., Ben-Tal, N., The ConSurf-HSSP Datenbank: the mapping of evolutionary conservation among homologs onto PDB Structures. *Proteins*, 2005, **58**, 610-617.

Goodford, P.J., A computational procedure for determining energetically favourable binding-sites on biologically important macromolecules. *J. Med. Chem.*, 1985, **28**, 849-857.

Grant, J.A., Pickup, B.T., A Gaussian Description of Molecular Shape, *J. Phys. Chem.*, 1995, **99**, 3503-3510.

H

Hansch, C., Leo, A.J., *Substituent Constants for Correlation Analysis in Chemistry and Biology*, John Viley, New York, **1979**.

Hajduk, P.J., Huth, J.R., Tse, C., Predicting protein druggability. *Drug Discov. Today*, 2005a, **10**, 1675-1682.

Hajduk, P.J., Huth, J.R., Fesik, S.W., Druggability Indices for protein targets derived from NMR-based screening data. *J. Med. Chem.*, 2005b, **48**, 2518-2525.

Heiden, W., Moeckel, G., Brickmann, J., A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comput. Aided Mol. Des.*, 1993, **7**, 503-514.

Hendlich, M., Rippmann, F., Barnickel, G., LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model.*, 1997, **15**, 359-363.

Hofbauer, C., Lohninger, H., Aszódi, A., SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 837-847.

Hopcroft, J.E., Motwani, R., Ullman, J.D., *Introduction to Automata Theory, Languages, and Computation*, Addison Wesley, Boston, **2001**, 179.

Hopkins, A.L., Groom, C.R., The druggable genome. *Nat. Rev. Drug Discov.*, 2002, **1**, 727-730.

Huang, B., Schröder, M., LIGSITE^{CSC}: predicting ligand binding sites using the Conolly surface and degree of conservation. *BMC Struct. Biol.*, 2006, **6**, 19-29.

I

Imming, P., Sinning, C., Meyer, A., Drugs, their targets and the nature and number of drug targets, *Nature Rev. Drug Disccov.*, 2006, **5**, 821-834.

J

Jarzyna, R., Lenarcik, E., Bryla, J., Chloroquine is a potent inhibitor of glutamate dehydrogenase in liver and kidney-cortex of rabbit, *Pharmacol. Res.*, 1997, **35**, 79-84.

K

Kabsch, W., A solution of the best rotation to relate two sets of vectors, *Acta Crystallogr*., 1976, **32**, 922-923.

Karaman, M.W., Herrgard, S., Treiber, D.K., Gallant, P., Atteridge, C.E., Campbell, B.T., Chan, K.W., Ciceri, P., Davis, M.I., Edeen, P.T., Faroni, R., Floyd, M., Hunt, J.P., Lockhart, D.J., Milanov, Z.V., Morisson, M.J., Pallares, G., Patel, H.K., Pritchard, S., Wodicka, L.M., Zarrinkar, P.P., A quantitative analysis of kinase inhibitor selectivity, *Nat. Biotechnol.*, 2008, **26**, 127-132.

Klapper, I., Hagstrom, R., Fine, R., Sharp, K., Honig, B., Focusing of Electric Fields in the Active Site of Cu-Zn Superoxide Dismutase: Effects of Ionic Strength and Amino-Acid Modification., *Proteins*, 1986, **1**, 47-59.

Klebe, G., Recent developments in structure-based drug design. *J. Mol. Med.*, 2000, **78**, 269-281.

Knuth, D.E., *The Art of Computer Programming*, 3rd ed., Vol. 2, Addison-Wesley-Verlag, Upper Saddle River, NJ, **1998**, 10ff.

Kohonen, T., *Self-Organization and Associative Memory*, 3rd ed., Springer-Verlag, Berlin, **1989**.

Koshland, D.E., Application of a Theory of Enzyme Specificity to Protein Synthesis, *Proc. Natl. Acad. Sci.*, 1958, **44**, 98-104.

Koshland, D.E., The Key-Lock Theory and the Induced fit Theory, *Angew. Chem. Int. Ed. Engl.*, 1994, **33**, 2375-2378.

Kraut, J., Serine Proteases: Structure and Mechanism of Catalysis, *Annu. Rev. Biochem.*, 1977, **46**, 331-358.

Kubinyi, H., High throughput in drug discovery, *Drug Discov. Today*, 2002, **7**, 707-709.

Kwok, T., Zabler, D., Urman, S., Rohde, M., Hartig, R., Wessler, S., Misselwitz, R., Berger, J., Sewald, N., König, W., Backert, S., *Helicobacter* exploits integrin for type IV secretion and kinase activation, *Nature*, 2007, **449**, 862-866.

L

Langley, J.N, Dickinson, W.L., On the Local Paralysis of Periphal Ganglia, and on the Connexion of Different Classes of Nerve Fibres with Them, *Proc. R. Soc. Lond.*, 1889, **46**, 423-431.

Langley, J.N., On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari, *J. Physiol.*, 1905, **33**, 374-413.

Laskowski, R.A., SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, 1995, **13**, 323-330.

Laurie, A.T., Jackson, R.M., Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 2005, **21**, 1908-1916.

Leach, A.R., *Molecular Modelling*, 2nd ed., Pearson-Verlag, Prentice Hall, **2001**, 190**.**

Lee, B., Richards, F.M., The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, 1971, **55**, 179-400.

Levitt, D.G., Banaszak, L.J., POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.*, 1992, **10**, 229-234.

Liang, J., Edelsbrunner, H., Woodward, C., Anatomy of protein pockets and cavities: Measurements of binding site geometry and implications for ligand design. *Protein Sci.*, 1998, **7**, 1884-1897.

Liang, J., Dill, K.A., Are proteins well packed?, *Biophys. J.*, 2001, **81**, 751-766.

Lin Y.H., Chang H.C., Lin Y.L., A Study on Tools and Algorithms for 3-D Protein Structures Alignment and Comparison, *International Computer Symposium,* Dec 15-17, 2004, Taipei, Taiwan.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.*, 1997, **23**, 3–25.

Lorensen, W.E., Cline, H.E., Marching Cubes: A high resolution 3D surface construction algorithm, *Comput. Graph.*, 1987, **21**, 163-169.

Löwer, M., Weydig, C., Metzler, D., Reuter, A., Starzinski-Powitz, A., Wessler, S., Schneider, G., Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA, *PLoS ONE*, 2008; **3**, e3510.

Löwer M., Weydig, C., Wessler, S., Tanrikulu, Y., Weisel, M., Schneider, G., Structure-based design of *Helicobacter pylori* HtrA protease, 2009, *manuscript in preparation*.

M

MacKerell, A.D.J., Brooks, B., Brooks III, C.L., Nilsson, L., Roux, B., Won, Y., Karplus, M. Charmm: The energy function and its parametrization with an overview of the program, in Schleyer, P.v.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer III, H.F., Schreiner, P.R., editors, The Encyclopedia of Computational Chemistry, 1998, **1**, John Wiley & Sons, Chichester, UK, 271-277.

Martinetz, T.M., Schulten, K.J., A "neural-gas" network learns topologies. In *Artificial Neural Networks*, 1991, 397-402.

Mauss, H., Mietzch, F., Atebrin, ein neues Heilmittel gegen Malaria, *Klin. Wochenschr.*, 1933, **12**, 1278-1278.

Moon, J.B., Howe, W.J., Computer design on bioactive molecules: A method for receptor-based de novo ligand design. *Proteins: Struct. Funct. Genet.*, 1991, **11**, 314-328.

Moreau, G., Broto, P., Autocorrelation of molecular structures: application to SAR studies, *Nouv. J. Chim*., 1980, **4**, 757-764.

Morris, J.R., Deaven, D.M., Ho, K.M., Genetic algorithm energy minimization for point charges on a sphere. *Phys. Rev.*, **B53**, 1995, 1740-1743.

N

Nietert, M., Virtuelles Screening nach RNA-Liganden: Zum Umgang mit einer flexiblen Zielstruktur, Dissertation, **2008**, Goethe Universität, Frankfurt am Main.

Nietert, M.M., Weisel, M., Proschak, E., Kestner, E., Gohlke, H., Schneider, G., Pocket Surface Dynamic Ligand Binding Pockets for Structure-Based Virtual Screening, 2008, *manuscript in preparation*.

O

Omura, S., Iwai, Y., Hirano, A., Nakagawa, A., Awaya, J., Tsuchiya, H., Takahashi, Y., Masuma, R., A new alkaloid AM-2282 of Streptomyces origin taxonomy, fermentation, isolation and preliminary characterization, *J. Antibiot.*, 1977, **30**, 275-282.

P

Pastor, M., Cruciani, G., McLay, I., Pickett, S., Clementi, S., GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem*., 2000, **43**, 3233–3243.

Podjarny, A., Cachau, R.E., Schneider, T., Van Zandt, M., Joachimiak, A., Subatomic and atomic crystallographic studies of aldose reductase: implications for inhibitor binding, *Cell. Mol. Life Sci.*, 2004, **61**, 763-773.

Proschak, E., Rupp, M., Derksen, S., Schneider, G., Shapelets: Possibilities and limitations of shape-based virtual screening. *J. Comp. Chem.*, **29**, 2007, 108-114.

Proschak, E., Virtuelles Screening und *De Novo* Design von PPARα Agonisten mit Oberflächen-Deskriptoren, Dissertation, **2008**, Goethe Universität, Frankfurt am Main, 17ff.

Q

Qasba, P. K., Involvement of sugars in protein-protein interactions. *Carbohydrate Polymers*, 1999, **41**, 293–309.

Quillin, M.L., Matthews, B.W., Accurate calculation of the density of proteins, *Acta Crystallogr.*, 2000, **56**, 791-794.

R

Rekker, R.F., de Kort, H.M., The Hydrophobic Fragmental Constant; An Extension to a 1000 Data Point Set. *Eur. J. Med. Chem.*, 1979, **14**, 479-488.

Reif, F., *Statistische Physik und Theorie der Wärme*, 2nd edition, de Gruyter, Berlin, **1985**, 454.

Renner, S., Schwab, C.H., Gasteiger, J., Schneider, G., Impact of conformational flexibility on three-dimensional similarity searching using correlation vectors. *J. Chem. Inf. Model.*, 2006, **46**, 2324-2332.

Rottler, J., Maggs, A.C., Local Molecular Dynamics with Coulombic Interactions. *Phys. Rev. Lett.*, **93**, 2004, 170201.

S

Sadowski, J., Kubinyi, H., A scoring scheme for discriminating between drugs and nondrugs, *J. Med. Chem.*, 1998, **41**, 3325–3329.

Saff, E.B., Kuijlaars, A.B.J., Distributing Many Points on a Sphere. *Math. Intelligencer*, **19**, 1997, 511-549.

Saitou, N., Nei, M., The Neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 1987, **4**, 406-425.

Sattigeri, J.A., Andappan, M.M.S., Kishore, D., Thangathirupathy, S., Sundaram, S., Singh, S., Sharma, S., Davis, J.A., Churgh, A., Bansal, V., Discovery of conformationally rigid 3-azabicyclo[3.1.0]hexane-derived dipeptidyl peptidase-IV inhibitors, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 4087-4091.

Schmitt, S., Kuhn, D., Klebe, G., A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology, *J. Mol. Biol.*, 2002, **323**, 387-406.

Schneider, G., Wrede, P., Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.*, 1998, **70**, 175-222.

Schneider, G., Neidhart, W., Giller, T., Schmid, G., „Scaffold Hopping“ by topological pharmacophore search: a contribution to virtual screening, *Angew. Chemie Int. Ed.*, 1999, **38**, 2894-2896.

Schneider, P., Schneider, G., Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.*, 2003, **22**, 713–718.

Schneider, G., Schneider, P., in: Kubinyi, H., Müller, G., editors, Chemogenomics in Drug Discovery, **2004**, Wiley-VCH, Weinheim, 341–376.

Schneider, G., Baringhaus, K.H.. *Molecular Design*, Wiley-VCH, Weinheim, **2008**, 3.

Schroeder, W.J., Zarge, J.A., Lorensen, W.E., Decimation of triangle meshes. *Comput. Graph.*, 1992, **26**, 65-70.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schikowski, B., Ideker, T., Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, *Genome Res*., 2003, **13**, 2498-504.

Sotriffer, C.A., Krämer, O., Klebe, G., Probing flexibility and „induced-fit“ phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations, *Proteins*, 2004, **56**, 52-66.

Stahl, M., Taroni-Osterroth, C., Schneider, G., Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.*, 2000, **13**, 83-88.

Stauch, B., Hofmann, H., Weisel, M., Cichutek, K., Münk, C., Schneider, G., Model Structure of APOBEC3C Supports a Critical Role of Protein Dimerization for Cell-Intrinsic Antiviral Activity, *Proc. Natl. Acad. Sci. USA*, 2009, *in revision*.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E.L., The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.,* 2003, **43**, 493-500.

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., Willighagen, E.L., Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.,* 2006, **12**, 2111-2120.

T

Tammes, P.M.L., On the origin of number and arrangement of the places of exit on the surface of pollen-grains. *Rec. Trav. Bot. Neerl.*, 1930, **27**, 1-84.

Tanford, C., The hydrophobic effect and the organization of living matter. *Science*, 1978, **200**, 1012-1018.

Tanrikulu, Y., Nietert, M., Scheffer, U., Proschak, E., Grabowski, K., Schneider, P., Weidlich, M., Karas, M., Göbel, M., Schneider, G., Scaffold Hopping by “Fuzzy”

Pharmacophores and its Application to RNA Targets, *ChemBioChem*, 2007, **8**, 1932-1936.

Tarjan, R., Depth-First Search and Linear Graph Algorithms. *SIAM J. Comput.*, 1972, **1**, 146-160.

Taylor, D.E., Plasmid-mediated tetracycline resistance in Campylobacter jejuni: expression in Escherichia coli and identification of homology with streptococcal class M determinant, *J. Bacteriol.*, 1986, **165**, 1037-1039.

Thurston, W.P., Shapes of polyhedra and triangulations of the sphere. *Geom. Topol. Monographs*, 1998, **1**, 511-549.

Tsai, J., Taylor, R., Chothia, C., Gerstein, M., The packing density in proteins: Standard radii and volumes, *J. Mol. Biol.*, 1999, **290**, 253-266.

UV

Viswanadhan, V.N., Ghose, A.K., Revankar, G.R., Robins, R.K., Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occuring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 163-172.

W

Wang, R., Fang, X., Lu, Y., Wang, S., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med Chem.*, 2004, **47**, 2977-2980.

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., Wang, S., The PDBbind database: methodologies and updates. *J. Med. Chem.*, 2005, **48**, 235-242.

Weisel, M., Entwicklung einer Rezeptor-basierten Pharmacophorfunction in PyMOL, Diplomarbeit, **2006**, Goethe Universität Frankfurt am Main.

Weisel, M., Proschak, E., Schneider, G., PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.*, 2007, **1**:7.

Weisel, M., Proschak, E., Kriegl, J.M., Schneider, G., Form follows function: Shape analysis of protein cavities for receptor-based drug design, *Proteomics*, 2009, **9**, 451-459.

Werner, C., Stubbs, M.T., Krauth-Siegel, R.L., Klebe, G., The crystal structure of Plasmodium falciparum glutamate dehydrogenase, a putative target for novel antimalarial drugs, *J. Mol. Biol.*, 2005, **349**, 597-607.

Weydig, C., Starzinski-Powitz, A., Carra, G., Löwer, J., Wessler, S., CagA-independent disruption of adherence junction complexes involves E-cadherin shedding and implies multiple steps in *Helicobacter pylori* pathogenicity, *Exp. Cell Res.*, 2007, **313**, 3459-3471.

Wolfram, S., Statistical Mechanics of cellular automata, *Rev. Mod. Phys.*, 1983, **55**, 601-644.

Wolfram, S., Cellular automata as models of complexity, *Nature*, 1984, **311**, 419-424.

Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C., Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *J. Mol. Biol.*, 1999, **285**, 1735-1747.

XYZ

Zachmann, C.D., Heiden, W., Schlenkrich, M., Brickmann, J., Topological analysis of complex molecular surfaces. *J. Comput. Chem.*, 1992, **13**, 76-84.

6 Danksagung

An dieser Stelle möchte ich mich bei den vielen Freunden und Kollegen bedanken, die mich bei der Anfertigung dieser arbeitet begleitet und unterstützt haben.

Mein besonderer Dank geht an Prof. Dr. Gisbert Schneider für die Betreuung dieser Arbeit als Doktorvater und die vielen motivierenden Gespräche, die immer wieder Anstoß zu spannenden wissenschaftlichen Ideen gaben. Ich möchte mich für eine Zeit bedanken, in der ich mit viel Enthusiasmus in einem professionellen Umfeld vieles lernen konnte.

Dr. Jan Kriegl danke ich herzlich für die Betreuung dieser Arbeit von industrieller Seite und die vielen wertvollen Anregungen, die sehr zum Gelingen dieser Arbeit beigetragen haben. Ich bedanke mich für die freundliche Aufnahme bei meinen Besuchen in Biberach und für die Einblicke in den Forschungsalltag bei Boehringer Ingelheim Pharma.

Ich möchte mich bei Prof. Dr. Joachim Engels für die Übernahme der Zweitkorrektur dieser Arbeit herzlich bedanken.

Dr. Ewgenij Proschak danke ich für die Überlassung des *Shapelets*-Algorithmus ohne den diese Arbeit nicht möglich gewesen wäre. Ich bedanke mich außerdem für viele wichtige Ratschläge bei der Entwicklung neuer Algorithmen und die Einführung in Java.

Bei Yusuf Tanrikulu möchte ich für die Bereitstellung von *LIQUID* und die Unterstützung bei vielen technischen Problemen bedanken.

Martin Löwer danke ich für die Zusammenarbeit bei der Weiterentwicklung und Anwendung von *ReverseLIQUID*. Über die Ergebnisse des strukturbasierten Virtuellen Screenings habe ich mich sehr gefreut.

Bei Benjamin Stauch möchte ich mich für erfolgreiche Anwendung von *PocketPicker* für Funktionsvorhersagen im APOBEC-Projekt bedanken.

Kathleen Zocher danke ich für die Zusammenarbeit bei spannenden Untersuchungen der Taschen der Glutamat Dehydrogenase.

Bei Dr. Manuel Nietert möchte ich mich für die Analysen vieler hunderttausend RNA-Bindetaschen bedanken.

Dr. Paul Czodrowski danke ich für seine Hilfestellungen bei der Verwendung von PDB2PQR.

Ganz besonderer Dank geht an die vielen netten Kollegen des modlab Teams für den regen Austausch, viele spannende Seminare und eine tolle Zeit in einer ausgesprochen angenehmen Atmosphäre. Vielen Dank an Swetlana Derksen, Norbert Dichter, Uli Fechner, Lutz Franke, Tim Geppert, Alireza Givehchi, Tina Grabowski, Volker Hähnke, Markus Hartenfeller, Jan Hiß, Bettina Hofmann, Natalie Jäger, Alexander Klenner, Björn Krüger, Felix Reisen, Carmen Rödl, Matthias Rupp, Brigitte Scheidemantel-Geiß, Michael Schmuker, Petra Schneider, Andreas Schüller, Tim Werner und Joanna Wisniewski.

Ich möchte mich weiterhin bei den Mitarbeitern der Abteilungen Chemieinformatik und Modelling der Firma Boehringer Ingelheim am Standort für die freundliche Aufnahme und zahlreiche fachliche Anregungen bedanken.

Der Firma Boehringer Ingelheim Pharma GmbH & Co. KG danke ich für die Finanzierung dieser Arbeit.

Ich danke der Frankfurt International Research Graduate School for Translational Biomedicine (F.I.R.S.T) für die Abschlussfinanzierung dieser Arbeit.

Mein Dank geht an die Unterstützung durch das Beilstein-Institut zur Förderung der Chemischen Wissenschaften und den Sonderforschungsbereich 579 „RNA-Liganden-Wechselwirkungen“.

Ich möchte mich bei meinen Freunden bedanken, die mich seit vielen Jahren während des Studiums und der Promotion begleitet haben.

Abschließend möchte ich bei meiner Familie für die liebevolle Unterstützung in allen Lebenslagen bedanken. Ganz besonders gilt mein Dank meiner Schwester Friederike Weisel und meinen Eltern Inge Weisel und Karl-Heinrich Weisel.

7 Anhang

I Handbuch PocketomePicker

Die in dieser Arbeit entwickelten Methoden sind in einer Software *PocketomePicker* zusammengefasst worden. Die einzelnen Methoden stehen dem Benutzer als Kommando-zeilenprogramme zur Verfügung, die eine serielle, verteilte Berechnung großer Datenbanken auf Großrechnern erlauben. Ergebnisse der Berechnungen werden in Ausgabedateien gespeichert. Automatisch generierte Python-Skripte ermöglichen zudem eine Visualisierung der Ergebnisse in PyMOL (DeLano, 2002) für jede einzelne Eingabe einer seriellen Berechnung.

Darüber hinaus wurde eine grafische Benutzeroberfläche implementiert, die einen intuitiven Zugang und eine einfache Handhabung der Programmpakete ermöglicht. Ergebnisse einer Berechnung werden hier in einem eingebettetem JyMOL Molekülbetrachter angezeigt, die über alle wesentlichen Funktionsmerkmale von PyMOL verfügt.

Die implementierten Programme wurden für die Betriebssysteme Linux und Windows entwickelt. Eine detaillierte Beschreibung zur Installation und Handhabung von *PocketomePicker* ist nachfolgend abgedruckt. Die Software wurde in Zusammenarbeit mit der Firma **Boehringer Ingelheim Pharma GmbH & Co. KG** entwickelt. Die Verwendung der Software ist auf beschränkt auf das **Molecular Design Laboratory** (modlab®) der Beilstein-Stiftungsprofessur Chemieinformatik an der Goethe-Universität Frankfurt am Main, sowie auf die Firma Boehringer Ingelheim Pharma GmbH & Co. KG.

Auf den folgenden Seiten ist das Handbuch zur Software *PocketomePicker* in englischer Sprache abgedruckt. Es umfasst folgende Abschnitte:

- Requirements – Voraussetzungen (installierte Programme, Betriebssysteme)
- Installation – Hinweise zur Installation von *PocketomePicker*
- Usage – Anleitung zum Gebrauch der Software, Anwendungsbeispiele
- Appendix – Anhang mit Informationen über Dateitypen der Ein- und Ausgabedateien

PocketomePicker

version 1.1

Manual

April 2009

Table of Contents

Requirements

The following programs must be installed on the machine (server) running PocketomePicker:

- Java 3D™ (Sun Microsystems)[1], version 1.3 or higher
- Python™ (Python Software Foundation)[2], version 2.3 or higher

Output files are generated as Python files which can be visualized using:

- PyMOL™ (DeLano Scientific)[3], all versions, subscription required for commercial use

The following programs are included in PocketomePicker:

- Java™ (Sun Microsystems)[1], build 1.5.0_10-b03
- PocketPicker[4], Pocket Prediction
- Reduce[5], Hydrogen Adding (Kinemage), v3.13
- PDB2PQR[6], APBS, v1.3.0
- JyMOL[7], Java variant of PyMOL, v0.63 (win), v0.71 (unix)

Installation

The software is provided on a CD, precalculated results of PDBbind[8,9] Refined Set are supplied on DVD.

Installation Linux:

Copy the contents of folder `PocketomePicker/Linux/` from the software CD to the desired location on your machine. Make sure that the provided reduce hydrogen adding program (`/red/reduce.3.1.3.080428.linuxi386`) is executable and works in your environment. Otherwise you need to compile the program from source and update `/red/reduce.sh` indicating the new file name.

Installation Windows:

Copy the contents of folder `PocketomePicker/Windows/` from the software CD to the desired location on your machine.

Remark:

Make sure Python is installed and available from command line as `python`.

PDBbind Database:

Protein `*.pdb` files, `*.sdf` ligands and precalculated pocket and shapelets results of the PDBbind[8,9] Refined Set monomeric structures (v2007) are provided with the software. Results are given for smoothed and unsmoothed pocket surfaces. Copy the `PDBbind07monomers/` folder to the desired location. Place the `results/` and/or `smoothresults/` folder into the `PDBbind07monomers/` folder.

Usage

General Remarks

It is highly recommended to stick to the PocketomePicker folder structure and file naming (see Appendix). Consistent folder and file naming is required as pocket or shapelet files are imported and exported during computations. Changing output file names might cause program crashes.

Note that user defined PDB identifiers must not contain blanks TABs or string separators.

It is recommended to use the `*.sh` or `*.bat` files provided with the software to start the programs. Indicated in this manual are the Java commands to launch the command line and GUI versions of PocketomePicker. If you plan to launch the software with the Java commands make sure to use the provided Java version to ensure comparability of results:

```
./jre/bin/java -jar <VM arguments, program name, arguments>
```

PocketomePicker (command line)

PocketomePicker enables serialized prediction of protein binding sites (serial PocketPicker) as well as calculation of shapelets for selected binding pockets.

The program is given as a java archive (*.jar file) and is called from the command line with syntax:

`java -jar -Xmx1G PocketomePicker.jar` *`arguments`*

(`-Xmx1G` reserves 1 gigabyte of memory for calculation)

mandatory *arguments*:

```
-inputfolder <String>   [default = '']
```

optional *arguments*:

```
-outputfolder <String>  [default = '<inputfolder>/results']
-pdbinputfile <String>  [default = '<inputfolder>/serialfiles.txt']
-biggestPocketShapelets
-noshapelets
-noproperties
-nohydrogenadding
-noLigandRemoval
-keepTempFiles
-surfaceSmoothing
```

The ordering of arguments on the command line is arbitrary.

Online Help :

Online help is called with arguments `-h` or `-help`:

```
java -jar PocketomePicker.jar -h
java -jar PocketomePicker.jar -help
```

Argument definitions

`-inputfolder`

Specifies the location where input *.pdb files are stored. The *.pdb files must be stored in separate folders rooting in `inputfolder`. Subfolders must be assigned the same name as the respective *.pdb file:

```
inputfolder/1abc/1abc.pdb
inputfolder/1xyz/1xyz.pdb
```

Note: The inputfolder must contain `serialfiles.txt` specifying which PDBs should be calculated (format: 1PDB-ID per line, e.g. 1abc), if no argument `-pdbinputfile` is given.

`-outputfolder`
Defines the location for output files. Default is `inputfolder/results`.

`-pdbinputfile`
Defines path to file listing PDB-IDs for calculation (format: 1 PDB-ID per line, e.g. 1abc). The inputfolder must contain a file `serialfiles.txt` if no argument `-pdbinputfile` is given.

`-biggestPocketShapelets`
Shapelets are computed only for the biggest pocket of each protein if this argument is set.

`-noshapelets`
No shapelet calculation is performed if this argument is given. Pocket prediction with PocketPicker only.

`-noproperties`
Shapelets are calculated without electrostatic or lipophilic property assignment

`-nohydrogenadding`
Reduce[5] hydrogen adding is disabled using this argument. Please note that your input pdb files have to be protonated if you choose this option. This is required for PocketPicker predictions and for property assignments.

`-noLigandRemoval`
Ligands (defined as `HETATM` in PDB notation) will not automatically be removed if this argument is set.

`-keepTempFiles`
The following temporary files will not be removed after calculation if you choose this option:

`*.<pdb-ID>_processed.pdb`	Input file after ligand removal (if not disabled)
`*.<pdb-ID>_processed_H.pdb`	Input file with hydrogens (Reduce[5])
`*.<pdb-ID>_PQR.pdb`	Input file with assigned properties (PQR[6])

`-surfaceSmoothing`
Shapelets are calculated on a smoothed pocket surface. This approach makes use of a gaussian function fitting method resulting in an increased runtime

Example usage

```
java -jar -Xmx1G PocketomePicker.jar -inputfolder C:\Data\PDBfiles\ -outputfolder C:\results\ -pdbinputfile C:\Data\allPDB_IDs.txt
```

This command starts pocket prediction and shapelet calculation (with property assignments) using inputfiles stored in subfolders rooting in `C:\Data\PDBfiles\` for all PDB-IDs listed in `C:\Data\allPDB_IDs.txt`. Results will be saved to `C:\results\` after program termination.

Output files

Output files are generated for each input <PDB-ID> and saved to the respective subfolder in the predefined `outputfolder` (default = `inputfolder/results`).

A detailed survey of PocketomePicker output files is given in the Appendix - *Output File Types*.

ShapeDescriptorComparator (command line)

ShapeDescriptorComparator enables command line based Euclidean distance calculations of PocketPicker ShapeDescriptors.

The program is given as a java archive (*.jar file) and is called from the command line with syntax:

`java -jar -Xmx1G ShapeDescriptorComparator.jar` *`arguments`*

(`-Xmx1G` reserves 1 gigabyte of memory for calculation)

General Remark:

Note that the crosswise calculation of Euclidean distances in 210 dimensional space can lead to considerably long runtimes in large databases!

mandatory *arguments*:

```
-inputfolder <String>          [default = '']
-databasefolder <String>       [default = '']
```

optional *arguments*:

```
-inputdescriptors <String>    [default = inputfolder/shapes2compare.txt]
-databasedescriptors <String> [default = databasefolder/shapes2compare.txt]
```

The ordering of arguments on the command line is arbitrary.

Online Help :

Online help is called with arguments `-h` or `-help`:

```
java -jar ShapeDescriptorComparator.jar -h
java -jar ShapeDescriptorComparator.jar -help
```

Argument definitions

`-inputfolder`

Specifies root folder of input pocket shape descriptors. `Inputfolder` must contain a file shapes2compare.txt stating which descriptors and which pockets should be used for distance calculation. Descriptors must be stored in subfolders according to PocketPicker folder structure and naming.

`-databasefolder`

Specifies root folder of shape descriptors for comparison.

ShapeDescriptorComparator uses `defaultshapes2compare.txt` (automatically generated after serial pocket predictions) as a reference for database pocket 1 descriptors. Optional: Place a file called shapes2compare.txt in databasefolder specifying descriptors and pocket for calculation. Descriptors must be stored in subfolders according to PocketPicker folder structure and naming.

`-inputdescriptors`

Specifies full path to alternate query input file. File Format: PDB-ID<TAB>pocket-nr.

`-databasedescriptors`

Specifies full path to alternate database input file. File Format: PDB-ID<TAB>pocket-nr.

For file formats see also Appendix – *Input File Types ShapeDescriptorComparator*.

Example usage

```
java -jar -Xmx1G ShapeDescriptorComparator.jar -inputfolder C:\Queries\results\ -databasefolder C:\Targets\results\
```

This command launches ShapeDescriptorComparator performing Euclidean distance calculations of queries and pockets indicated in:
`inputfolder\shapes2compare.txt` to targets with pockets specified in:
`databasefolder\defaultshapes2compare.txt` (if no shapes2compare.txt is present in `databasefolder`).

Output Files

Euclidean distances of the ten nearest target shape descriptors are stored to `inputfolder\ distance_matrix.html`.

PocketShapeletsMatching (command line)

PocketShapeletsMatching enables crosswise pocket shapelet matching of precalculated *.shapelets files. Results are displayed in a output *.html file listing the ten most similar pocket shapes to each input pocket (represented by a set of shapelets).

Defintion:
A query remains unchanged while target shapelets will be transformed in matching.

General Remark:
Shapelet matching is restricted to pockets having at least 5 shapelets. It is highly recommended to stick to the Shapelets Folder Structure contained in the PocketPicker Folder Structure (see Appendix - *Shapelets Root Folder*).

The program is given as a java archive (*.jar file) and is called from the command line with syntax:
`java -jar -Xmx1G PocketShapeletsMatching.jar` *`arguments`*
(`-Xmx1G` reserves 1 gigabyte of memory for calculation)

mandatory *arguments*:

```
-queryfolder <String>   [default = '']
-targetfolder <String>  [default = '']
```

optional *arguments*:

```
-outputfolder <String>          [default = '<queryfolder>/matching']
-outputfolder <String>          [default = folder of queryfile/matching]
-targetfiles <String>           [default = <targetfolder>targetfiles.txt]
-queryfiles <String>            [default = <queryfolder>queryfiles.txt]
-shapescoring                   [default = shapelet property matching]
-nopockettranslation            [default: transform querypocket]
-translatetargetpdb <String>    [default: no *.pdb file translation]
-translatetargetshapelets       [default: no *.shapelets file translation]
-nosizeconstraint               [default: target < 3*query pocket size]
-distancecutoff <double>        [default = 3.5]
-curvaturecutoff <double>       [default = 0.3]
-cliquesize <int>               [default = 7]
```

```
-matrixsize <int>              [default = 10]
```

The ordering of arguments on the command line is arbitrary.

Online Help :

Online help is called with arguments `-h` or `-help`:

```
java -jar PocketShapeletsMatching.jar -h
java -jar PocketShapeletsMatching.jar -help
```

Argument definitions

mandatory *arguments*:

`-queryfolder`

Specifies the full path to the root folder where query *.shapelets files are stored. The *.shapelets files must be stored in separate folders rooting in `queryfolder`. Subfolders must be named according to the respective *.shapelets file name:
`inputfolder/1abc/1abc.pdb`

`Queryfolder` MUST contain `queryfiles.txt` (if not `-queryfiles` is given) listing which *.shapelets files should be used for matching, e.g.
`1abc_unsmoothedShapelets.shapelets`
(see Appendix – *PocketShapeletsMatching Input Files*)

Remark: It is recommended to stick to the PocketPicker Folder Structure. In this way `queryfolder` always ends with '*results*'.

`-targetfolder`

Specifies the full path to the root folder where target *.shapelets files are stored. The *.shapelets files must be stored in separate folders rooting in `targetfolder`. Subfolders must be named according to the respective *.shapelets file name:
`targetfolder/1abc/1xyz.pdb`

`Targetfolder` MUST contain `targetfiles.txt` (if not `-targetfiles` is given) listing which *.shapelets files should be used for matching, e.g.
`1abc_unsmoothedShapelets.shapelets`
(see Appendix – *PocketShapeletsMatching Input Files*)

optional *arguments*:

`-outputfolder`

Defines alternate full path to output folder (default is `queryfolder/matching`)

`-queryfiles`
: Defines alternate file listing input query shapelets files (default is `queryfolder/queryfiles.txt`). For file format see Appendix – *PocketShapeletsMatching Input Files*.

`-targetfiles`
: Defines alternate file listing input target shapelets files (default is `targetfolder/targetfiles.txt`). For file format see Appendix – *PocketShapeletsMatching Input Files*.

`-shapescoring`
: Enables shapelet shape scoring – shapelet properties are not regarded (use this option for input shapelets without property parameters). Default is shapelet property scoring.

`-nopockettranslation`
: No output files of transformed target pockets will be created (default is: create transformed pocket *.txt and *.py files stored to `outputfolder`)

`-translatetargetpdb`
: Specifies the root folder of the target pdb files (see Appendix – *PDB Files Root Folder*). This option enables transformation of respective actual target pdb file with respect to transformation matrix of the actual target pocket. Resulting pdb file will be stored to `outputfolder`. Default is no pdb file transformation.

`-translatetargetshapelets`
: Enables transformation of actual target shapelets. Resulting *.shapelets file will be stored to `outputfolder`. Default is no target shapelets translation.

`-nosizeconstraint`
: Enables shapelet matching to all pockets (without size constraint). On default shapelet matching is restricted to pockets smaller than three times the size of the query. Using this argument can produce significantly larger runtimes.

`-distancecutoff`
: Specifies maximum distance of shapelets to be matched in association graph. Default is 3.5 (scale is angstroem).

`-curvaturecutoff`
: Specifies maximum curvature distance of shapelets to be matched. Default is 0.3 (scale is angstroem).

`-cliquesize`
: Specifies minimum number of pocket shapelets used for pocket translation. Default is 7.

`-matrixsize`
: Specifies the number of nearest pocket neighbors (pockets with highest score towards query) to be displayed in scores_matrix.html. Default is 10.

Example usage

```
java -jar -Xmx1G PocketShapeletsMatching.jar -queryfolder
C:\Queries\results\ -targetfolder C:\Targets\results\
-translatetargetshapelets
```

This command starts pocket shapelets matching for query shapelets files indicated in `queryfolder\queryfiles.txt`. Input *.shapelet files must be stored in subfolders rooting in `queryfolder` (files must be placed according to standard Shapelets folder structure, see Appendix - *Shapelets Root Folder* Structure), e.g. `queryfolder\1abc\1abc_unsmoothedShapelets.shapelets`). Results will be saved to `Folder of act query ID\matching\target ID\` after program termination. *.txt and *.py files will also be created in this example.

Output Files

Crosswise shapelet scores are stored to `queryfolder/scores_matrix.html`. Listed are the ten nearest target shapelets with highest scores for every query.

A number of different output files are generated depending on the arguments passed. A detailed survey of PocketShapeletsMatching output files is given in the Appendix - Output File Types.

PocketomePicker_gui (graphical user interface)

The PocketomePicker GUI provides a user interface that includes PocketPicker calculations, pocket shapelet decomposition and JyMOL visualization.

General remark:

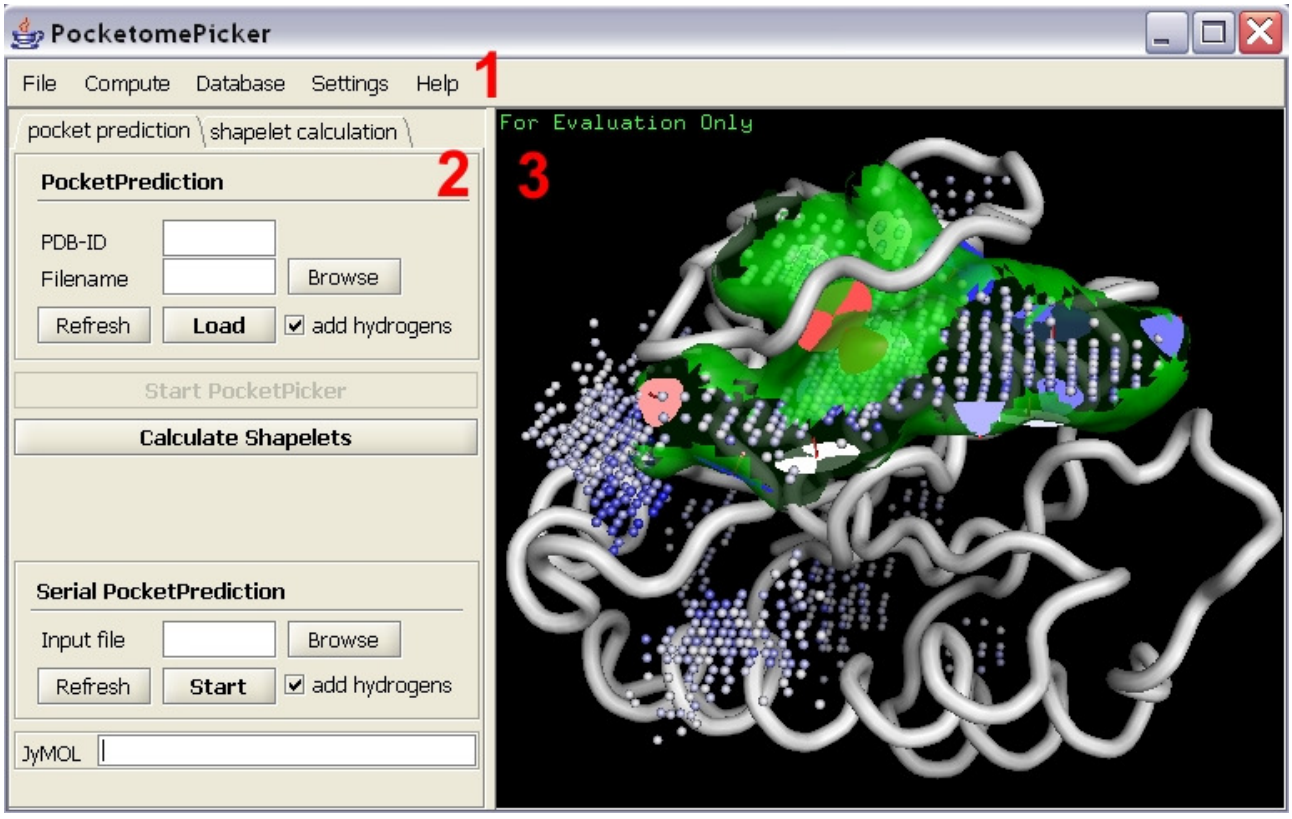
Please note PyMOL scripts for predicted pockets and calculated shapelet solutions will be created during PocketomePicker calculations. The *.py output files will be saved to the results folder.

General layout

(1) A **menubar** offers general functions such as pocket or shapelets reload or online download of pdb files.

(2) The **control panel** is divided into two tabs for pocket and shapelet calculation. A command line allows simple display manipulations with basic PyMOL commands.

(3) The **JyMOL display panel** is used for visualization of pocket points, shapelets and pdb files.



Detailed control panel instructions

Menubar:

The following functions are available using menu bar navigation:

- File menu:
  - Load PDB file — loads pdb file from file system into JyMOL visualization panel. (Note: For single pocket prediction use file browser in pocket prediction panel.)
  - Load PocketPoints — reload of PocketPicker output *.txt file
  - Load Shapelets — reload of <PDB ID>*.shapelets file.
- Compute menu:
  - ShapeDescriptor Comparison — serialized Euclidean distance calculator for PocketPicker pocket autocorrelation vectors (find a detailed instruction in Appendix - *PocketPicker ShapeDescriptor Comparison*)
  - PocketShapeletsMatching
    - single
    - serial
- Database menu:
  - PDB Download — serialized download of pdb files. Enables saving into separate folders.
- Settings menu:
  - Serial PockerPicker — Save location of serial files input root folder and path to results folder.
- Help menu:
  - About — information

Control Panel:

- PocketPicker Tab:

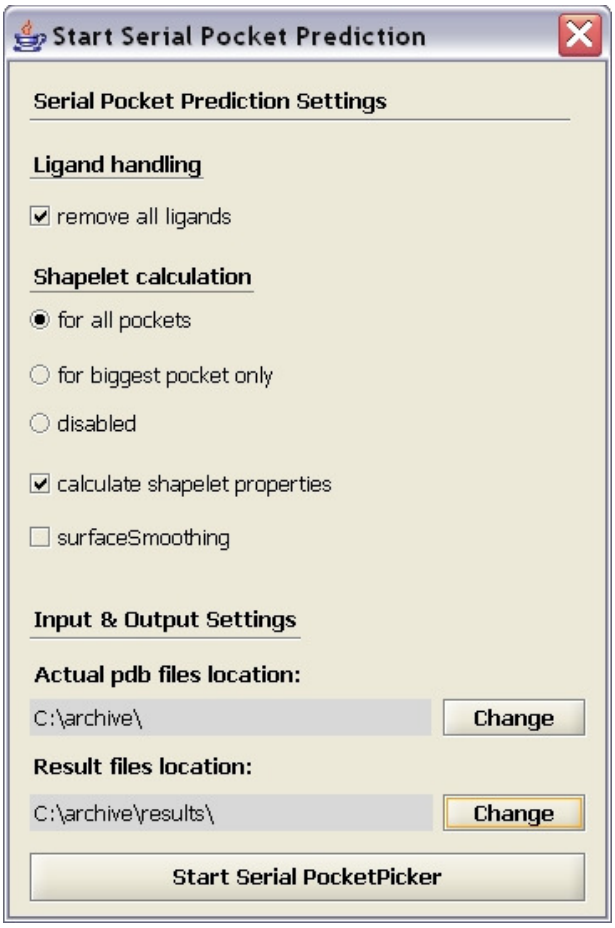
(1) Single file PocketPicker (top):

- Specify a four digit pdb id for online pdb download or use the Browse button to select a pdb from the file system.
- Use the Load button to load the requested file into the GUI. Reduce hydrogen adding is employed if the add hydrogens checkbox is selected.

- A popup menu lists all ligands and chains contained in the pdb and offers a routine for ligand and chain removal.
- A `Start PocketPickerbutton` appears after the pdb has loaded. Resulting pockets will be displayed in the JyMOL panel.

(2) Serial PocketPicker (middle):

- Use the `Browse button` to select a *.txt file holding the pdb identifiers for serial calculations from the file system (file format: one 4-digit pdb id per line, see Appendix - *Serial Pocket Prediction File Format*). Reduce hydrogen adding is employed if the `add hydrogens checkbox` is selected.
- The `Start button` opens the Serial PocketPicker settings dialog opens. The user can change settings for ligand handling, shapelet calculation and surface smoothing.



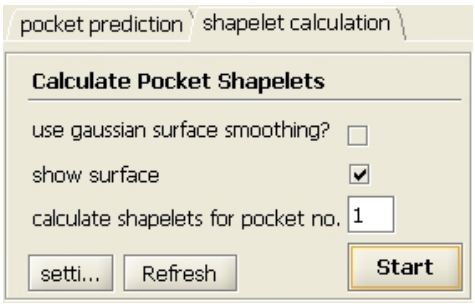
- Press the `Start Serial PocketPicker button` to start serial pocket prediction. Note that this procedure can lead to large runtimes depending on the number and sizes of the selected pdb files and the settings specified.
- Paths to pdb files root folder and result files location (see Appendix – *PDB Files Root Folder*, *PDB Files Result Folder*) have to be specified. Note that PocketomePicker will remember the settings of the last run.

⚠ Resulting pocket files and input pdb files will not be displayed!

Reload via `File --> load PDB file, ...`

(3) Command line (bottom):
- The command line supports a number of simple PyMOL commands to manipulate JyMOL representation and visualization (see Appendix – *PyMOL to JyMOL Command Parser*).

- Shapelets Tab:
  The shapelets tab is used for shapelet calculation following a single file PocketPicker calculation.
  - Specify a pocket number (1 = biggest) and press the `Start button` for shapelet calculation.
  - Use Refresh button to delete shapelets and surface.



JyMOL display panel:

The `JyMOL display panel` is used for visualization of *PocketomePicker* calculations. Zooming, Clipping and rotation controls are identical to PyMOL handling. Supported PyMOL commands (representations, actions; see Appendix – *PyMOL to JyMOL Command Parser*) can be executed in the JyMOL command line (`pocket prediction panel`).

References

[1] Sun Microsystems, Inc., Santa Clara, CA, USA.

[2] Python Software Foundation, Hampton, NH, USA.

[3] DeLano, W.L., The PyMOL Molecular Graphics System. *DeLano Scientific*, 2002, San Carlos, CA, USA.

[4] Weisel, M., Proschak, E., Schneider, G., Analysis of Ligand Binding-Sites with Shape Descriptors. *Chem. Cent. J.*, 2007, **1**:7.

[5] Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C., Asparagine and Glutamine; Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *J. Mol. Biol.*, 1999, **285**, 1735-1747.

[6] Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A., PDB2PQR: an automated pipeline for the setup, execution and analysis of Poisson-Boltzmann electrostatics calculations. *Nucl. Acids Res.*, 2004, **32**, W665-W667.

[7] DeLano, W.L., JyMOL: PyMOL Graphics for Java Developers. *DeLanoScientific*, 2007, San Carlos, CA, USA.

[8] Wang R, Fang X, Lu Y, Wang S (2004). „The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with known Three-Dimensional Structures.” *J. Med. Chem.* **47**, 2977-2980.

[9] Wang R, Fang X, Lu Y, Yang CY, Wang S (2005). “The PDBbind Database: Methodologies and updates.” *J. Med. Chem.* **48**, 4111-4119.

Appendix

Output File Types

PocketPicker output files

The following file types are automatically created by the PocketPicker pocket prediction routine:

`JPocPic_outText_<PDB-ID>.txt` TAB separated txt file
Format: (Pocket-nr. | x | y | z | buriedness-index)

JPocPic_output_`<PDB-ID>`.py PyMOL[3] script
Execute this script with `File→Run→<script name>` from the PyMOL menu bar.
This file visualizes pocket volume and buriedness predicted with PocketPicker[4].

JPocPic_ShapeDesc210_`<PDB-ID>`.txt TAB separated txt file
This file holds PocketPicker[4] shape descriptors (210-dimensional autocorrelation vectors) for each pocket predicted.
Format: One descriptor per line, first line refers to pocket number 1 (biggest pocket).

JPocPic_ShapeDesc210_`<PDB-ID>`_SCALED.txt TAB separated txt file
This file holds scaled shape descriptors for each pocket predicted. Shape descriptors are scaled to length 1.
Format: One descriptor per line, first line refers to pocket number 1 (biggest pocket).

defaultshapes2compare.txt TAB separated txt file
Format: PDB-ID<TAB>1
This file is generated to assist Euclidean ShapeDescriptor comparison. It can be used as a default file listing the all input IDs and pocket nr '1' for Euclidean distance calculations.

Shapelet Output Files

Subsequent file types are automatically generated by the PocketomePicker pocket shapelets routine:

*.shapelets TAB separated shapelets file
This file holds information of the shapelets calculated for the respective protein.
Format: (Shapelet-nr. | xyz-coords of shapelet center | xyz of vector normal | curvature | electrostatic potential | lipophilic potential of shapelet center)

*Shapelets.py PyMOL[3] script

Execute this script with `File→Run→<script name>` from the PyMOL menu bar.
This file visualizes shapelets in PyMOL. Shapelets are colored with respect to calculated properties (if selected).

Shapelet Matching Output Files

scores_matrix.html lists the ten target shapelets with maximum score towards the actual query. This file is stored to queryfolder/scores_matrix.html.

The following files are generated by PocketShapeletsMatching (depending on the arguments passed) and stored to queryfolder/<query ID>/matching/<query ID pocket number>:

*<target pocket number>_translated.pdb PDB file
Target PDB file translated according to the translation matrix of the actual target pocket.

*<target pocket number>_translatedPocket.py PyMOL[3] script
Python script for visualization of translated target pocket. Execute this script with `File→Run→<script name>` from the PyMOL menu bar.

*<target pocket number>_translatedPocket.txt TAB separated txt file
Format: (Pocket-nr. | x | y | z | buriedness-index)

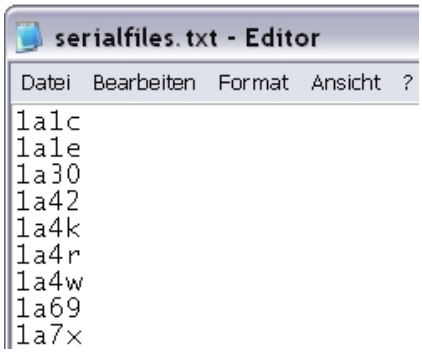
*<target pocket number>*Shapelets_translated.py PyMOL[3] script
Execute this script with `File→Run→<script name>` from the PyMOL menu bar.
This file visualizes the translated shapelets of the actual pocket in PyMOL.

*<target pocket number>*Shapelets_translated.shapelets TAB separated txt file
This file holds information of the translated shapelets of the actual target pocket.
Format: (Shapelet-nr. | xyz-coords of shapelet center | xyz of vector normal | curvature | electrostatic potential | lipophilic potential of shapelet center)

Input File formats

Serial Pocket Prediction Input File

In this example serialfiles.txt holds all 4-digit pdb identifiers selected for serial pocket prediction. Note that the respective pdb files have to be stored according to the *PocketPicker Folder Structure*.



ShapeDescriptorComparator Input Files

The input files for ShapeDescriptorComparator have to meet the following format (one query/target per line):
`PDB-ID<TAB>pocket-nr`

Please use multiple line entries for comparisons of multiple queries or targets.

PocketShapeletsMatching Input Files

In this example `queryfiles.txt` and `targetfiles.txt` are the files indicating which shapelets of which pocket should be used for matching.

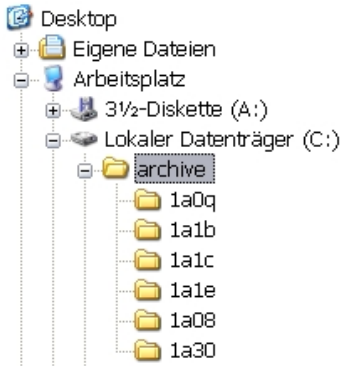
`queryfiles.txt` lists one *.shapelets filename per line, e.g.
`1abc_unsmoothedShapelets.shapelets`
Optional `queryfiles` format to indicate desired pocket (default = pocket 1):
`<*.shapelets filename>TAB<pocket for shapelet comparison>`, e.g.
`1abc_unsmoothedShapelets.shapelets<TAB>2`

`targetfiles.txt` has to match the `queryfiles` format with one difference:
In the Optional format also negative integers are allowed → '-3' would start a matching of the three biggest target pockets to the actual query.

PocketPicker Folder structure

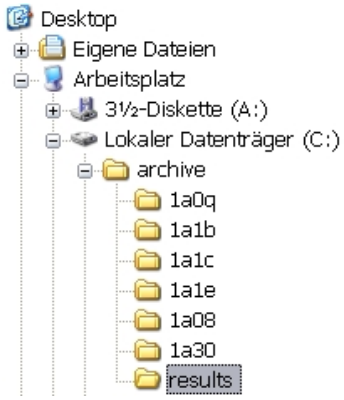
PDB Files Root Folder

In this example `C:\archive\` is the root folder. The respective pdb files have to be placed in separate subfolders with the respective name, e.g. `1a0q\1a0q.pdb`



PDB Files Result Folder

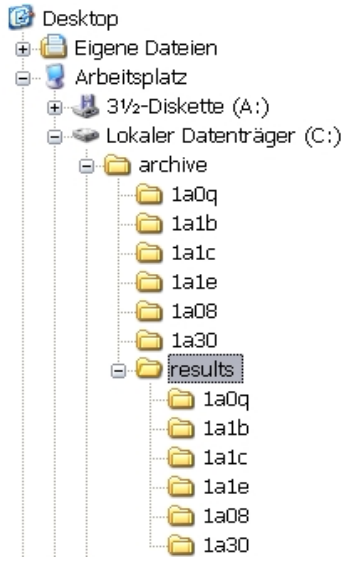
A `results folder` can be selected from the file system. `Results folders` always end with *results*.



Shapelets Root Folder

In this example `C:\archive\results\` is the shapelets root folder. The respective *.shapelets files have to be stored in separate subfolders (this is done automatically by PocketPicker shapelet calculation) with the respective name, e.g. `1a0q\1a0q_unsmoothedShaplets.shaplets`.

Remark:
Do not remove the `JPocPic_outText_<PDB-ID>.txt` file from this folder, for it might be used for target pocket translation.



PyMOL to JyMOL Command Parser

Supported PyMOL commands:

The following PyMOL commands (including arguments) can be used in *PocketomePicker*. Please note that command abbreviations are not supported.

```
align, angle, center, clip, color, create, delete, dihedral,
disable, distance, draw, enable, get_angle, get_dihedral,
get_distance, get_view, gradient, hide, isolevel, isomesh,
isosurface, label, load, orient, origin, originAt,
reinitialize, select, selectList, set, set_view, show, zoom
```

A reference manual for the above PyMOL commands can be found at:
http://pymol.sourceforge.net/newman/ref/toc.html

PocketPicker ShapeDescriptor Comparison

Serial Euclidean distance calculation:

This method employs Euclidean distance calculations of a set of query PocketPicker ShapeDescriptors compared to a precalculated database of ShapeDescriptors.

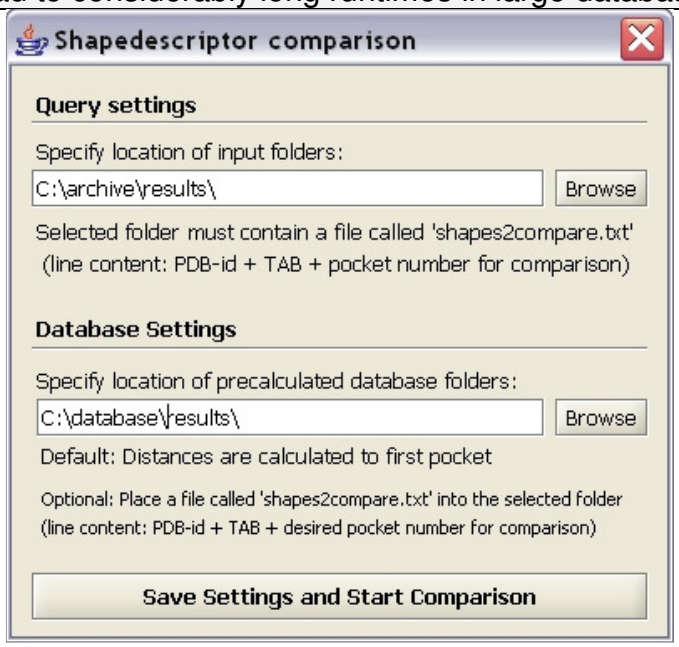
PocketPicker ShapeDescriptors are autocorrelation vectors encoding pocket size and buriedness into 210 dimensional tab separated descriptors of type `JPocPic_ShapeDesc210_<PDB ID>.txt`. Every line represents the ShapeDescriptor for one pocket.

Folders specifying the location of the query ShapeDescriptor location as well as the database Descriptors have to be set prior to computation. Furthermore, the query folder must contain a file called `shapes2compare.txt` listing PDB-IDs and pocketnumber of query files selected for comparison (file format: `PDB-ID <TAB> pocketnumber` – one line per PDB-ID).

Euclidean distances are calculated to biggest pocket (pocket 1) of every file in database folder on default. Optional: place a file called `shapes2compare.txt` into the database folder specifying database PDB-IDs and pocket numbers desired for comparison (file format, see above).

⚠ It is strongly recommended to use the standard PocketPicker result folders, folder structure and ShapeDescriptor file names as they are reconstructed from the PDB-IDs specified in the `shapes2compare.txt` files.

⚠ Note that the crosswise calculation of Euclidean distances in 210 dimensional space can lead to considerably long runtimes in large databases!



Absolute Euclidean distances of the ten nearest neighbors for every Shapedescriptor are listed in a color coded html Euclidean distance matrix as output.

Euclidean distance matrix

10gs	11gs (2562.6927244599574)	1bp0 (2840.825584227233)	1ppm (2857.6873516884243)	1gvw (2989.660683087631)	1gvu (3154.98003
11gs	1gvw (1848.876145121679)	5er2 (1962.7949459890099)	1gvu (2035.340512051976)	1epo (2039.6423706130445)	1zap (2350.66437
16pk	1utl (336.8486306933724)	1o2u (392.91093138267354)	1o2o (422.2440052860431)	1o30 (435.5513746964874)	1c83 (441.154168
1a07	1fzq (172.9942195566083)	1m48 (201.61845153655952)	1a1b (205.9538783320188)	1lkk (212.38408603282875)	1oba (219.972725
1a08	1a07 (248.81117338254728)	1gzc (259.5707995904008)	1gz9 (270.22398117117586)	1fzq (271.66523517005265)	1oxn (318.981190

II Datensatz zur Evaluation der Qualität der Taschenvorhersage von PocketPicker

Tabelle A1: Sammlung von 48 Komplexen und ihren entsprechenden *apo*-Strukturen (dargestellt durch ihre PDB-Abkürzungen) zur Evaluation der Taschenvorhersage mit *PocketPicker*.

Komplex	*apo*-Struktur	Protein Beschreibung	Referenzligand[1]	Weitere Liganden[2]
1bid	3tms	Thymidylate-Synthase	ump	cbx
1cdo	8adh	Alkohol-Dehydrogenase	nad	zn
1dwd	1hxf	Alpha-Thrombin	mid	Ketten i, l
1fbp	2fbp	Fructose 1,6-bisphosphatase	amp	f6p, mg
1gca	1gcg	Glucose/Galactose-Bindeprotein	gal	ca
1hew	1hel	Hühnereiweiß-Lysozym	nag	-
1hyt	1npc	Thermolysin	bzs	dms, ca, zn
1inc	1esa	Elastase	icl	ca, so4
1rbp	1brq	Retinol Bindeprotein	rtl	-
1rob	8rat	Ribonuclease A	c2p	-
1stp	1swb	Streptavidin	btn	-
1ulb	1ula	Purin-Nucleoside-Phosphorylase	gun	so4
2ifb	1ifb	Fettsäure-Bindeprotein	plm	-
3ptb	3ptn	Beta-Trypsin	ben	ca
2ypi	1ypi	Triose-Phosphat-Isomerase	pga	-
4dfr	5dfr	Dihydrofolat-Reduktase	mtx	ca, cl
4phv	3phv	HIV-1-Protease	vac	-
5cna	2ctv	Concanavalin A	mma	ca, cl, mn
7cpa	5cpa	Carboxypeptidase A	fvf	zn
1a6w	1a6u	B1-8 FV Fragment	nip	-
1acj	1qif	Acetylcholinesterase	tha	-
1apu	3app	Penicillopepsin	[iva-val-val-sta-oet]	man
1blh	1djb	Beta-Lactamase	fos	-
1byb	1bya	Beta-Amylase	glc	so4
1hfc	1cge	Fibroblast-Collagenase	hap	ca, zn
1ida	1hsi	HIV-2-Protease	[qnd-val-hpb-ppl-py2]	-
1igj	1a4j	Immunoglobulin	dgx	Kette y
1imb	1ime	Inositol-Monophosphatase	lip	gd
1ivd	1nna	Hydrolase	st1	fuc, nag, man, ca
1mrg	1ahc	Alpha-Momorcharin	adn	-
1mtw	2tga	Trypsin	dx9	ca
1okm	4ca2	Carboanhydrase II	sab	hg, zn
1pdz	1pdy	Enolase	pga	ace, mn
1phd	1phc	Camphor-5-Monoxygenase	pim	hem
1pso	1psn	Pepsin 3a	[iva-val-val-sta-ala-sta]	-

1qpe	3lck	Lck-Kinase	pp2	ptr, so4
1rne	1bbs	Renin	c60	nag
1snc	1stn	Staphylokokken-Nuklease	ptp	ca
1srf	1pts	Streptavidin	mtb	-
2ctc	2ctb	Carboxypeptidase A	lof	zn
2h4n	2cba	Carboanhydrase II	azm	zn
2pk4	1krn	Plasminogen	aca	-
2sim	2sil	Sialidase	dan	-
2tmn	1l3f	Thermolysin	[pho-leu-nh2]	ca, zn
3gch	1chg	Gamma-Chymotrypsin	cin	-
3mth	6ins	Methylparaben-Insulin	mpb	cl, zn
5p2p	3p2p	Phospholipase A	dhg	ca
6rsa	7rat	Ribonuklease A	uvc	dod

[1]Betrachteter Ligand zur Beschreibung der jeweiligen Bindetasche. Eckige Klammern zeigen Liganden an, die aus mehreren Fragmenten bestehen.
[2]Weitere Liganden und Ionen wurden vor Beginn der Berechnungen aus der Struktur entfernt.

III Klassifikationsgüte der Selbstorganisierenden Karten zur Vorhersage der Druggability

Tabelle A2: Klassifikationsgüte der für die Datensätze A und B berechneten binären SOMs. Gegeben sind die absoluten Zahlen der ShapeDeskriptoren (als Summe der leeren und ligandenbindenden Taschen), positiv korrekte ligandenbindende Taschen (*P*), negativ korrekt klassifizierte leere Taschen (*N*), falsch klassifizierte leere Taschen (*O*), falsch klassifizierte ligandenbindende Taschen (*U*), Sensitivität (*s = P/P+U*) und Matthews Korrelationskoeffizienten (*cc*).

	# Deskriptoren	*P*	*N*	*O*	*U*	*s*	*p*	*cc*
Datensatz A	13859	419	13123	94	223	0,65	0,82	0,72
Datensatz B	2257	80	2127	25	25	0,76	0,76	0,76

Tabelle A3: Klassifikationsgüte der binären SOMs, die für ligandenbindende Taschen mit vorhergesagter hoher Drug-Likeness (*DL*) und niedriger Drug-Likeness (*NDL*). *P* = positiv korrekt, *N* = negativ korrekt, *O* = falsch positiv (engl. *overpredicted*), *U* = falsch negativ (engl. *underpredicted*), *s* = Sensitivität, *p* = Präzision, *cc* = Matthews Korrelationskoeffizient.

	# Deskriptoren	*P*	*N*	*O*	*U*	*s*	*p*	*cc*
Datensatz B (*DL*)	638	25	607	3	3	0,89	0,89	0,89
Datensatz B (*NDL*)	579	20	549	4	6	0,8	0,83	0,79
Datensatz A (*DL*)	638	24	608	2	4	0,86	0,92	0,88
Datensatz A (*NDL*)	579	23	547	6	3	0,88	0,79	0,83

IV Beschreibung der Proteinfunktionen für die Taschen aus Datensatz A

Tabelle A4: Übersicht über die Verteilung der Bindetaschen und der Funktionen ihrer jeweiligen Proteine für das Clustering der PocketGraphen (MSQE = 5 Å, *r* = 0,98) aus Datensatz A.

PDB	Cluster	Beschreibung	Name	E.C.
1od8	a	Hydrolase	Xylanase a	3.2.1.8
1wdn	a	Bindeprotein	Glutamin Bindeprotein	-.-.-
1ofz	a	Lectin	Fucose-specifisches Lectin	-.-.-
5yas	a	Lyase	Hydroxynitril Lyase	4.1.2.39
1qbn	a	Hydrolase	Trypsin	3.4.21.4
1li3	a	Hydrolase	Lysozyme	3.2.1.17
1y2f	a	Zellzyklus	Zellteilungsprotein Zipa	-.-.-
1n51	a	Hydrolase	Proaminopeptidase	3.4.11.9
1o2k	a	Hydrolase	Beta Trypsin	3.4.21.4
1ce5	a	Hydrolase	Trypsin	3.4.21.4
1ws1	a	Hydrolase	Peptidedeformylase 1	3.5.1.88
2cji	a	Hydrolase	Faktor Xa	3.4.21.6
1li6	a	Hydrolase	Lysozyme	3.2.1.17
1v2q	a	Hydrolase	Trypsin	3.4.21.4
2fw6	a	Lyase	Ribonucleotidmutase	-.-.-
2b4l	a	Transportprotein	Glycine betaine-Bindeprotein	-.-.-
1w8l	a	Isomerase	Peptidyl-prolyl Cis-trans Isomerase A	5.2.1.8
1jn2	a	Zuckerbindeprotein	Concanavalin A	-.-.-
1v2r	a	Hydrolase	Trypsin	3.4.21.4
1fki	a	Cis-trans Isomerase	Fk506 Bindeprotein	-.-.-
2hdr	a	Hydrolase	Beta-lactamase	3.5.2.6
1o0h	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1bzj	a	Hydrolase	Tyrosine Phosphatase	3.1.3.48
1d7i	a	Isomerase	Fk506 Bindeprotein	5.2.1.8
1rgl	a	Hydrolase	Ribonuclease t1	3.1.27.3
2dri	a	Zuckertransportprotein	D-Ribose-Bindeprotein	-.-.-
1oss	a	Hydrolase	Trypsin	3.4.21.4
1gj6	a	Hydrolase	Beta Trypsin	3.4.21.4
2fai	a	Hormonrezeptor	Östrogenrezeptor	-.-.-
2i3i	a	Apoptose	Baculoviral Protein 7	-.-.-
1xk5	a	Transportprotein	Snurportin-1	-.-.-
1o33	a	Hydrolase	Beta Trypsin	3.4.21.4
2afw	a	Transferase	Glutaminyl-peptide Cyclotransferase	2.3.2.5
1k1m	a	Hydrolase	Trypsin	3.4.21.4
1apb	a	Bindeprotein	L-arabinose-Bindeprotein	-.-.-
1x8r	a	Transferase	Carboxyvinyltransferase	2.5.1.19
1fzq	a	Signalprotein	Adp-RibosylationFaktorprotein 3	-.-.-
1w4p	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1oyq	a	Hydrolase	Trypsin	3.4.21.4
1gny	a	Kohlenhydrat Bindeprot.	Xylanase 10c	3.2.1.8
1d7j	a	Isomerase	Fk506 Bindeprotein	5.2.1.8
1uto	a	Hydrolase	Trypsinogen	3.4.21.4
1c87	a	Hydrolase	Tyrosine Phosphatase 1B	3.1.3.48
1w4o	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1gi1	a	Hydrolase	Beta Trypsin	3.4.21.4

1o2s	a	Hydrolase	Beta Trypsin	3.4.21.4
2f5t	a	Transkription	Transkriptionsregulator trmb	-.-.-
1fao	a	Signalprotein	Dualadapter von Phosphotyrosin	-.-.-
8abp	a	Bindeprotein	L-arabinose-Bindeprotein	-.-.-
1c84	a	Hydrolase	Tyrosine Phosphatase 1B	3.1.3.48
1w4q	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
2fwp	a	Lyase	Ribonucleotidmutase	-.-.-
1axz	a	Lectin	Lectin	-.-.-
1z95	a	Transkriptionsregulation	Androgenrezeptor	-.-.-
1o30	a	Hydrolase	Beta Trypsin	3.4.21.4
1laf	a	Aminosäuretransport	Lys/Arg/Orn Bindeprotein	-.-.-
1jn4	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1o3i	a	Hydrolase	Beta Trypsin	3.4.21.4
1ctt	a	Hydrolase	Cytidine deaminase	3.5.4.5
1ecq	a	Lyase	Glucarat Dehydratase	4.2.1.40
1kdk	a	Transportprotein	Sexualhormon Bindeglobulin	-.-.-
1tnh	a	Hydrolase	Trypsin	3.4.21.4
1uou	a	Transferase	Thymidin Phosphorylase	2.4.2.4
1br6	a	Hydrolase	Ricin	3.2.2.22
1g7f	a	Hydrolase	Tyrosinphosphatase	3.1.3.48
1g7g	a	Hydrolase	Tyrosinphosphatase	3.1.3.48
1o8b	a	Isomerase	Ribose 5-phosphat isomerase	5.3.1.6
1tnk	a	Hydrolase	Trypsin	3.4.21.4
2ca8	a	Transferase	Glutathione s-transferase 28 kda	2.5.1.18
1gi8	a	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
7abp	a	Bindeprotein	L-arabinose-Bindeprotein	-.-.-
1j14	a	Hydrolase	Trypsin II	3.4.21.4
1j16	a	Hydrolase	Trypsin II	3.4.21.4
1utl	a	Hydrolase	Trypsin I	3.4.21.4
1zpa	a	Hydrolase	Pol Polyprotein	3.4.23.16
2j4i	a	Hydrolase	Faktor Xa	3.4.21.6
1alw	a	Calciumbindung	Calpain	3.4.22.17
3kiv	a	Kringle	Apolipoprotein	-.-.-
1fkg	a	Cis-trans Isomerase	Fk506 Bindeprotein	-.-.-
1cbx	a	Hydrolase/Peptidase	Carboxypeptidase a	3.4.17.1
1lah	a	Aminosäuretransport	Lys/Arg/Orn Bindeprotein	-.-.-
2aac	a	Transkriptionsfaktor	Arac	-.-.-
1gzc	a	Lectin	Erythrina Lectin	-.-.-
1lgw	a	Hydrolase	Lysozyme	3.2.1.17
1dzk	a	Duftstoffbindung	Odorant-Bindeprotein	-.-.-
1rnt	a	Endoribonuclease	Ribonuclease T1 Isozyme	3.1.27.3
1o0f	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
6rnt	a	Endoribonuclease	Ribonuclease T1	3.1.27.3
1v0l	a	Hydrolase	Endo-1,4-beta-Xylanase a	3.2.1.8
1c5s	a	Hydrolase	Trypsin	3.4.21.4
5abp	a	Bindeprotein	L-arabinose-Bindeprotein	-.-.-
1bzc	a	Hydrolase	Tyrosine Phosphatase	3.1.3.48
1kav	a	Hydrolase	Tyrosinphosphatase	3.1.3.48
1tx7	a	Hydrolase	Trypsinogen	3.4.21.4
1fkh	a	Cis-trans Isomerase	Fk506 Bindeprotein	-.-.-
1qy1	a	Transportprotein	Harnprotein	-.-.-
1wcq	a	Hydrolase	Sialidase	3.2.1.18
1oko	a	Lectin	Pa-i Galactophiles Lectin,	-.-.-

1o3h	a	Hydrolase	Beta Trypsin	3.4.21.4
1zp8	a	Hydrolase	Pol Polyprotein	3.4.23.16
1trd	a	Oxidoreductase	Triosephosphat Isomerase	5.3.1.1
1xzx	a	Hormonrezeptor	Thyroid Hormonerezeptor beta-1	-.-.-
1jak	a	Hydrolase	Beta-n-Acetylhexosaminidase	3.2.1.52
1gi6	a	Hydrolase	Beta Trypsin	3.4.21.4
1k1j	a	Hydrolase	Trypsin	3.4.21.4
1ctu	a	Hydrolase	Cytidin Deaminase	3.5.4.5
1qbo	a	Hydrolase	Trypsin	3.4.21.4
2b1v	a	Hormonerezeptor	Östrogenrezeptor	-.-.-
1li2	a	Hydrolase	Lysozyme	3.2.1.17
1o0n	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1w8m	a	Isomerase	Peptidyl-prolyl Cis-trans Isomerase	5.2.1.8
2h4g	a	Hydrolase	Tyrosine-protein Phosphatase	3.1.3.48
1mai	a	Signaltransduktion	Phospholipase C	3.1.4.11
1pb9	a	Ligand Bindeprotein	N-methyl-d-Aspartat Rezeptor	-.-.-
1p1o	a	Membranprotein	Glutamatrezeptor 2	-.-.-
1vjc	a	Transferase	Phosphoglycerate Kinase	2.7.2.3
1txr	a	Hydrolase	Leucyl Aminopeptidase	3.4.11.10
1qca	a	Acyltransferase	Chloramphenicol Acetyltransferase	2.3.1.28
1a99	a	Bindeprotein	Putrescine-Bindeprotein	-.-.-
1o2z	a	Hydrolase	Beta Trypsin	3.4.21.4
1afl	a	Hydrolase	Ribonuclease a	3.1.27.5
1iy7	a	Hydrolase	Carboxypeptidase a	3.4.17.1
2ctc	a	Peptidase	Carboxypeptidase a	3.4.17.1
1k1l	a	Hydrolase	Trypsin	3.4.21.4
1nf8	a	Hydrolase	Phenazine Biosyntheseprotein phzd	3.3.2.1
6abp	a	Bindeprotein	L-Arabinose-Bindeprotein	-.-.-
1l83	a	Hydrolase(o-glycosyl)	T4 lysozyme	3.2.1.17
1v2w	a	Hydrolase	Trypsin	3.4.21.4
1drj	a	Zuckertransportprotein	D-ribose-Bindeprotein	-.-.-
1uwf	a	Zelladhäsion	Fimh protein	-.-.-
1qft	a	Ligand Bindeprotein	Histamine Bindeprotein	-.-.-
1tni	a	Hydrolase	Trypsin	3.4.21.4
2f6t	a	Hydrolase	Tyrosin Proteinphosphatase	3.1.3.48
1d5r	a	Hydrolase	Phosphoinositid Phosphotase pten	3.1.3.48
1utn	a	Hydrolase	Trypsinogen	3.4.21.4
1fkb	a	Isomerase	Fk506 Bindeprotein	-.-.-
1c5x	a	Blutgerinnung	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1imx	a	Wachstumsfaktor	Insulinartiger Wachstumsfaktor 1a	-.-.-
1wc1	a	Lyase	Adenylatcyclase	4.6.1.1
2fqo	a	Lyase	S-Ribosylhomocysteine Lyase	4.4.1.21
1qy2	a	Transportprotein	Harnprotein	-.-.-
1y20	a	Ligand Bindeprotein	Glutamatrezeptor	-.-.-
1qb9	a	Hydrolase	Trypsin	3.4.21.4
2fqy	a	Transportprotein	Membranlipoprotein tmpc	-.-.-
1zky	a	Hormone rezeptor	Östrogenrezeptor	-.-.-
1nny	a	Hydrolase	Tyrosinphosphatase	3.1.3.48
1k1n	a	Hydrolase	Trypsin	3.4.21.4
1o0m	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1drk	a	Bindeprotein	D-ribose-Bindeprotein	-.-.-
1m1b	a	Isomerase	Phosphoenolpyruvat Phosphomutase	5.4.2.9
1gz9	a	Lectin	Erythrina Lectin	-.-.-

1z4o	a	Isomerase	Beta-Phosphoglucomutase	5.4.2.6
1rql	a	Hydrolase	Phosphonoacetaldehyde Hydrolase	-.-.-
220l	a	Hydrolase	T4 lysozyme	3.2.1.17
9abp	a	Bindeprotein	L-arabinose-Bindeprotein	-.-.-
1o3f	a	Hydrolase	Beta Trypsin	3.4.21.4
1vjj	a	Transferase	Glutamyltransferase E	2.3.2.13
1wvj	a	Membranprotein	Ionotropher Glutamatrezeptor 2	-.-.-
1bap	a	Bindeprotein	L-arabinose-Bindeprotein	-.-.-
1xt8	a	Transportprotein	Aminosäuretransporter	-.-.-
1tng	a	Hydrolase	Trypsin	3.4.21.4
1rpj	a	Zuckerrezeptor	Periplasmatischer Zuckerrezeptor	-.-.-
1r9l	a	Proteinbindung	Glycin Betainbindeprotein	-.-.-
1a4r	a	Hydrolase	G25k gtp-Bindeprotein	-.-.-
1f0r	a	Hydrolase	Faktor Xa	3.4.21.6
1sw1	a	Proteinbindung	Osmoprotektion Protein (prox)	-.-.-
1f0t	a	Hydrolase	Trypsin	3.4.21.4
1ocq	a	Hydrolase	Endoglucanase 5a	3.2.1.4
1j17	a	Hydrolase	Trypsin II	3.4.21.4
1z4n	a	Isomerase	Beta-Phosphoglucomutase	5.4.2.6
1v2n	a	Hydrolase	Trypsin	3.4.21.4
1w1d	a	Transferase	Protein Kinase-1	2.7.1.37
1rle	a	Transferase	Glutamyltransferase E	2.3.2.13
1af2	a	Hydrolase	Cytidin Deaminase	3.5.4.5
2ihq	a	Hormonrezeptor	Androgenrezeptor	-.-.-
1g36	a	Hydrolase	Trypsinogen	3.4.21.4
1g3e	a	Hydrolase	Beta Trypsin	3.4.21.4
2bza	a	Hydrolase	Trypsin	3.4.21.4
1o2n	a	Hydrolase	Beta Trypsin	3.4.21.4
1sgx	a	Transferase	Glutamyltransferase E	2.3.2.13
1ec9	a	Lyase	Glucarat Dehydratase	4.2.1.40
1igb	a	Aminopeptidase	Aminopeptidase	3.4.11.10
1ecv	a	Hydrolase	Protein-Tyrosine Phosphatase 1B	3.1.3.48
1y2g	a	Zellzyklus	Zellteilungsprotein Zipa	-.-.-
1pa9	a	Hydrolase	Protein-Tyrosin Phosphatase yoph	3.1.3.48
1z6s	a	Hydrolase	Pancreatische Ribonuclease	3.1.27.5
1abf	a	Bindeprotein	L-Arabinose-Bindeprotein	-.-.-
1rgk	a	Endoribonuclease	Ribonuclease t1	3.1.27.3
1y1z	a	Ligand Bindeprotein	Glutamatrezeptor	-.-.-
1fkf	a	Isomerase	Fk506 Bindeprotein	-.-.-
1w3l	a	Hydrolase	Endoglucanase 5a	3.2.1.4
4fiv	a	Aspartatprotease	Immunodefizienz Virus Protease	3.4.23.16
1d2e	a	Rna Bindeprotein	Elongationsfaktor tu	-.-.-
2h4k	a	Hydrolase	Tyrosin-Protein Phosphatase	3.1.3.48
1n46	a	Transkription	Thyroid Hormonrezeptor beta-1	-.-.-
2fx6	a	Hydrolase	Trypsin	3.4.21.4
2c4v	a	Lyase	3-Dehydroquinat Dehydratase	4.2.1.10
1dud	a	Hydrolase	Nucleotidhydrolase	3.6.1.23
1v2j	a	Hydrolase	Trypsin	3.4.21.4
2gv7	a	Hydrolase	Tumorsupressor 14	3.4.21.-
1pbk	a	Isomerase	Fkbp25	-.-.-
1tnj	a	Hydrolase	Trypsin	3.4.21.4
1ax0	a	Lectin	Lectin	-.-.-
2azr	a	Hydrolase	Tyrosin Proteinphosphatase	3.1.3.48

2i3h	a	Apoptose	Baculoviral Protein 7	-.-.-
1bzh	a	Hydrolase	Tyrosine Phosphatase 1b	3.1.3.48
1ppc	a	Hydrolase	Trypsin	3.4.21.4
1rzx	a	Zellzyklus	Cg5884-pa	-.-.-
1a1c	a	Transferase	C-src Tyrosinkinase	2.7.1.112
1fiv	a	Hydrolase	Fiv Protease	3.4.23.16
1ele	a	Hydrolase	Elastase	3.4.21.36
1elc	a	Hydrolase	Elastase	3.4.21.36
1j4r	a	Isomerase	Fk506-Bindeprotein	5.2.1.8
1elb	a	Hydrolase	Elastase	3.4.21.36
1fwv	a	Sugar Bindeprotein	Mannoserezeptordomäne	-.-.-
1ulg	a	Sugar Bindeprotein	Galectin-2	-.-.-
1pph	a	Hydrolase	Trypsin	3.4.21.4
1obx	a	Adhäsion	Syntenin 1	-.-.-
1bma	a	Hydrolase	Elastase	3.4.21.36
1a1e	a	Transferase	C-src Tyrosinkinase	2.7.1.112
1eld	a	Hydrolase	Elastase	3.4.21.36
1oai	a	Kerntransport	Nucleärer RNA Exportfaktor	-.-.-
1a1b	a	Transferase	C-src Tyrosinkinase	2.7.1.112
1jzs	b	Ligase	Isoleucyl-trna synthetase	6.1.1.5
2euk	b	Lipidtransport	Glycolipid Transferprotein	-.-.-
1pme	b	Transferase	Erk2	2.7.1.-
1v48	b	Transferase	Purine Nucleosidphosphorylase	2.4.2.1
1qiw	b	Calcium-Bindeprotein	Calmodulin	-.-.-
1ikt	b	Oxidoreductase	Estradiol beta-Dehydrogenase	1.1.1.62
1k1y	b	Transferase	4-alpha-Glucanotransferase	2.4.1.25
1hmt	b	Lipid-Bindeprotein	Muskel Fettsäure Bindeprotein	-.-.-
1fq5	b	Hydrolase	Saccharopepsin	3.4.23.25
1wvc	b	Transferase	Cytidylyltransferase	2.7.7.33
1mmr	b	Metalloprotease	Matrilysin	3.4.24.23
1q84	b	Hydrolase	Acetylcholinesterase	3.1.1.7
1yc4	b	Zellzyklus	Hitzeschockprotein 90-alpha	-.-.-
1tvo	b	Transferase	Mitogenactivierte Proteinkinase 1	2.7.1.37
1uho	b	Hydrolase	Phosphodiesterase	3.1.4.17
1ndz	b	Hydrolase	Adenosin Deaminase	3.5.4.4
1oyt	b	Hydrolase	Thrombin	3.4.21.5
1h22	b	Hydrolase	Acetylcholinesterase	3.1.1.7
1mrn	b	Transferase	Thymidylat Kinase	2.7.4.9
1g9r	b	Transferase	Glycosyl transferase	2.4.1.-
1sc8	b	Hydrolase	Plasminogen Activator, Urokinase	3.4.21.73
1w96	b	Ligase	Acetyl-Coenzyme A Carboxylase	6.4.1.2
1ppi	b	Hydrolase	Alpha-amylase	3.2.1.1
1bxo	b	Hydrolase	Penicillopepsin	3.4.23.20
1amw	b	Chaperon	Hitzeschockprotein 90	-.-.-
2g94	b	Hydrolase	Beta-Secretase 1	3.4.23.46
2b7d	b	Blutgerinnung	Faktor VII	3.4.21.21
2cf8	b	Hydrolase	Thrombin	3.4.21.5
2bak	b	Transferase	Mitogenaktivierte Proteinkinase 14	2.7.1.37
1njs	b	Transferase	Formyltransferase	2.1.2.2
1m0n	b	Lyase	2,2-Dialkylglycin Decarboxylase	4.1.1.64
1zvx	b	Hydrolase	Neutrophil Collagenase	3.4.24.34
1fhd	b	Hydrolase	Beta-1,4-Xylanase	3.2.1.91
1k4g	b	Transferase	Trna-Guanine Transglycosylase	2.4.2.29

2flb	b	Hydrolase	Faktor VII	-.-.-
1adl	b	Lipid-Bindeprotein	Adipocyt Lipidbindeprotein	-.-.-
1uvt	b	Hydrolase	Thrombin	3.4.21.5
1lee	b	Hydrolase	Plasmepsin 2	3.4.23.39
2evl	b	Lipidtransport	Glycolipid Transferprotein	-.-.-
1ejn	b	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1caq	b	Hydrolase	Stromelysin-1	3.4.24.17
1ndv	b	Hydrolase	Adenosin Deaminase	3.5.4.4
1mrs	b	Transferase	Thymidylat Kinase	2.7.4.9
1b8y	b	Hydrolase	Stromelysin 1	3.4.24.17
1lf2	b	Hydrolase	Plasmepsin 2	3.4.23.39
1h23	b	Hydrolase	Acetylcholinesterase	3.1.1.7
2am4	b	Transferase	Mannosyl-Glycoprotein	2.4.1.101
1vja	b	Hydrolase	Plasminogen Activator, Urokinase	3.4.21.73
5tmp	b	Transferase	Thymidylat Kinase	2.7.4.9
1sqo	b	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1rej	b	Transferase	Camp-Kinase	2.7.1.37
4tmk	b	Transferase	Thymidylate Kinase	2.7.4.9
2bal	b	Transferase	Mitogenactivierte Proteinkinase 14	2.7.1.37
1m13	b	Transkription	Kernrezeptor pxr	-.-.-
1ro7	b	Transferase	Alpha-2,3/8-Sialyltransferase	2.4.99.-
1bxq	b	Hydrolase	Penicillopepsin	3.4.23.20
1nvq	b	Transferase	Ser/Thr Proteinkinase chk1	2.7.1.-
1yc1	b	Zellzyklus	Hitzeschockprotein HSP 90-alpha	-.-.-
1xd1	b	Hydrolase	Alpha-Amylase	3.2.1.1
1xjd	b	Transferase	Proteinkinase C	2.7.1.-
1gjb	b	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1m2q	b	Transferase	Casein kinase II	2.7.1.37
1g2l	b	Hydrolase	Faktor Xa	3.4.21.6
1w7x	b	Hydrolase	Faktor VIIa	3.4.21.21
1vyg	b	Transportprotein	Fettsäure Bindeprotein	-.-.-
2bvd	b	Hydrolase	Endoglucanase H	3.2.1.4
1sb1	b	Hydrolase	Prothrombin	3.4.21.5
1ork	b	Transkriptionsregulation	Tetracyclin Repressor	-.-.-
1lnm	b	Bindeprotein	Diga16	-.-.-
1d9i	b	Hydrolase	Thrombin	3.4.21.5
2cf9	b	Hydrolase	Thrombin	3.4.21.5
1rek	b	Transferase	Camp Kinase	2.7.1.37
1b39	b	Transferase	Zellteilungs Proteinkinase 2	2.7.1.37
1ex8	b	Transferase	6-Hydroxymethyl-7,8-Dihydropterin	2.7.6.3
1ciz	b	Metalloproteinase	Stromelysin-1	3.4.24.17
1ndy	b	Hydrolase	Adenosine Deaminase	3.5.4.4
1hmr	b	Lipid-Bindeprotein	Muskel Fettsäure Bindeprotein	-.-.-
456c	b	Matrixmetalloprotease	Mmp-13	3.4.24.-
1b38	b	Transferase	Zellteilungs Proteinkinase 2	2.7.1.37
2bok	b	Hydrolase	Faktor Xa	3.4.21.6
1re8	b	Transferase	Camp-Kinase	2.7.1.37
1qy5	b	Chaperon	Endoplasmin	-.-.-
1t32	b	Hydrolase	Cathepsin G	3.4.21.20
1xd0	b	Hydrolase	Alpha-Amylase	3.2.1.1
1hms	b	Lipid-Bindeprotein	Muskel Fettsäure Bindeprotein	-.-.-
966c	b	Matrixmetalloprotease	Mmp-1	3.4.24.-
1sqa	b	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73

1udt	b	Hydrolase	Cgmp-Phosphodiesterase	3.1.4.17
1vzq	b	Hydrolase	Thrombin	3.4.21.5
1q91	b	Hydrolase	5(3)-Desoxyribonucleotidase	3.1.3.5
1q65	b	Transferase	Queuin tRNA-ribosyltransferase	2.4.2.29
2ayr	b	Transkription	Östrogenrezeptor	-.-.-
1fdq	b	Lipid Bindeprotein	Fettsäure Bindeprotein	-.-.-
1nvs	b	Transferase	Ser/Thr Proteinkinase chk1	2.7.1.-
1nt1	b	Hydrolase	Thrombin	3.4.21.5
1n1t	b	Hydrolase	Sialidase	3.2.1.18
1nvr	b	Transferase	Ser/Thr Proteinkinase chk1	2.7.1.-
2hu6	b	Hydrolase	Macrophage Metalloelastase	3.4.24.65
2d3u	b	Transferase	Polyprotein	2.7.7.48
1vyf	b	Transportprotein	Fettsäure Bindeprotein	-.-.-
2fdp	b	Hydrolase	Beta-Secretase 1	3.4.23.46
1g7v	b	Lyase	Desoxyphosphooctonat Aldolase	4.1.2.16
1b7h	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1b46	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1epo	b	Hydrolase	Endothiapepsin	3.4.23.22
1b3h	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1ent	b	Hydrolase	Endothiapepsin	3.4.23.22
1b6h	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1fkn	b	Hydrolase	Memapsin 2	3.4.23.-
2er6	b	Hydrolase	Endothiapepsin	3.4.23.6
1b05	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1a4w	b	Hydrolase	Alpha-Thrombin	3.4.21.5
1b5j	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1ppl	b	Hydrolase	Penicillopepsin	3.4.23.20
1gvw	b	Hydrolase	Endothiapepsin	3.4.23.22
1b40	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1m4h	b	Hydrolase	Beta-secretase	3.4.23.-
1ppk	b	Hydrolase	Penicillopepsin	3.4.23.20
5er2	b	Hydrolase	Endothiapepsin	3.4.23.6
1ppm	b	Hydrolase	Penicillopepsin	3.4.23.20
1jev	b	Peptidtransport	Oligo-Peptidbindeprotein	-.-.-
1b51	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1b3f	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1b0h	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1apw	b	Hydrolase	Penicillopepsin	3.4.23.20
4er1	b	Hydrolase	Endothiapepsin	3.4.23.6
1gvx	b	Hydrolase	Endothiapepsin	3.4.23.22
1b4h	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1b3g	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1b4z	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
2er9	b	Hydrolase	Endothiapepsin	3.4.23.6
1b5i	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
2jxr	b	Hydrolase	Proteinase A	3.4.23.25
1b3l	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
4er2	b	Hydrolase	Endothiapepsin	3.4.23.6
1b58	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1qka	b	Peptidtransport	Periplasmat. Peptidbindeprotein	-.-.-
1b5h	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1b32	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1qkb	b	Peptidtransport	Periplasmat. Peptidbindeprotein	-.-.-

1b52	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1apv	b	Hydrolase	Penicillopepsin	3.4.23.20
1b2h	b	Peptidbindeprotein	Periplasmat. Peptidbindeprotein	-.-.-
1jet	b	Peptidtransport	Oligo-Peptidbindeprotein	-.-.-
2olb	b	Bindeprotein	Oligo-Peptidbindeprotein	-.-.-
5er1	b	Hydrolase	Endothiapepsin	3.4.23.6
1b9j	b	Peptidbindeprotein	Oligo-Peptidbindeprotein	-.-.-
1b8o	c	Transferase	Purin Nucleosid Phosphorylase	2.4.2.1
1w6y	c	Isomerase	Steroid delta-Isomerase	5.3.3.1
1v2l	c	Hydrolase	Trypsin	3.4.21.4
1mq6	c	Hydrolase	Faktor Xa	3.4.21.6
2std	c	Lyase	Scytalone Dehydratase	4.2.1.94
1tsl	c	Methyltransferase	Thymidylate Synthase	2.1.1.45
1f0u	c	Hydrolase	Trypsin	3.4.21.4
8cpa	c	Peptidase	Carboxypeptidase a	3.4.17.1
1utm	c	Hydrolase	Trypsin I	3.4.21.4
1tnl	c	Hydrolase	Trypsin	3.4.21.4
1lpk	c	Hydrolase	Faktor Xa	3.4.21.6
1gj8	c	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1bra	c	Proteinase	Trypsin	3.4.21.4
1ony	c	Hydrolase	Tyrosinphosphatase	3.1.3.48
1o3k	c	Hydrolase	Beta Trypsin	3.4.21.4
1ksn	c	Hydrolase	Faktor Xa	3.4.21.6
6cpa	c	Hydrolase	Carboxypeptidase a	3.4.17.1
1ghz	c	Hydrolase	Beta Trypsin	3.4.21.4
1pyn	c	Hydrolase	Tyrosinphosphatase	3.1.3.48
1o3j	c	Hydrolase	Beta Trypsin	3.4.21.4
1h6h	c	Px Domäne	Neutrophiler Cytosol Faktor 4	-.-.-
1uj6	c	Isomerase	Ribose 5-Phosphat Isomerase	5.3.1.6
1onz	c	Hydrolase	Tyrosinphosphatase	3.1.3.48
1o38	c	Hydrolase	Beta Trypsin	3.4.21.4
1utp	c	Hydrolase	Trypsinogen	3.4.21.4
1k1i	c	Hydrolase	Trypsin	3.4.21.4
1o3d	c	Hydrolase	Beta Trypsin	3.4.21.4
1b11	c	Hydrolase	Immunodefizienzvirus Protease	3.4.23.16
1g3b	c	Hydrolase	Beta Trypsin	3.4.21.4
1v2t	c	Hydrolase	Trypsin	3.4.21.4
1o2w	c	Hydrolase	Beta Trypsin	3.4.21.4
1c5q	c	Hydrolase	Trypsin	3.4.21.4
2afx	c	Transferase	Glutaminyl-Peptide Cyclotransferase	2.3.2.5
1gi4	c	Hydrolase	Beta Trypsin	3.4.21.4
1c5p	c	Hydrolase	Trypsin	3.4.21.4
1qb1	c	Hydrolase	Trypsin	3.4.21.4
1t7d	c	Hydrolase	Signal peptidase I	3.4.21.89
1nhu	c	Transferase	Hepatitis C virus ns5b RNA	2.7.7.48
1o2j	c	Hydrolase	Beta Trypsin	3.4.21.4
1qb6	c	Hydrolase	Trypsin	3.4.21.4
1v2s	c	Hydrolase	Trypsin	3.4.21.4
1nz7	c	Hydrolase	Tyrosinphosphatase	3.1.3.48
1agm	c	Hydrolase	Glucoamylase-471	3.2.1.3
2gv6	c	Hydrolase	Tumorsupressor 14	3.4.21.-
1o2q	c	Hydrolase	Beta Trypsin	3.4.21.4
1v2u	c	Hydrolase	Trypsin	3.4.21.4

1o2o	c	Hydrolase	Beta Trypsin	3.4.21.4
2br6	c	Hydrolase	A IIa ähnliches Protein	-.-.-
1vwn	c	Biotin-Bindeprotein	Streptavidin	-.-.-
1kjr	c	Zuckerbindeprotein	Galectin-3	-.-.-
1icj	d	Hydrolase	Peptide Deformylase	3.5.1.31
1g1d	d	Lyase	Carboanhydrase II	4.2.1.1
1fcz	d	Genregulation	Retinsäurerezeptor Gamma-1	-.-.-
1lgt	d	Oxidoreductase	Biphenyl-2,3-diol 1,2-Dioxygenase	1.13.11.39
1x9d	d	Hydrolase	Mannosyl-Oligosaccharide Hydrolase	3.2.1.113
1zs0	d	Hydrolase	Neutrophil Collagenase	3.4.24.34
1zdp	d	Hydrolase	Thermolysin	3.4.24.27
5tln	d	Hydrolase	Thermolysin	3.4.24.27
1gar	d	Transferase	Ribonucleotidtransformylase	2.1.2.2
1c5z	d	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1ydd	d	Hydrolyase	Carboanhydrase II	4.2.1.1
1cny	d	Lyase (oxo-acid)	Carboanhydrase II	4.2.1.1
1i9q	d	Lyase	Carboanhydrase II	4.2.1.1
4tln	d	Hydrolase	Thermolysin	3.4.24.27
1loq	d	Lyase	Orotidin Decarboxylase	4.1.1.23
1qf1	d	Hydrolase	Thermolysin	3.4.24.27
1g4o	d	Lyase	Carboanhydrase II	4.2.1.1
1yda	d	Hydrolyase	Carboanhydrase II	4.2.1.1
1qf2	d	Hydrolase	Thermolysin	3.4.24.27
1xap	d	Transkription	Retinsäurerezeptor beta	-.-.-
1i9o	d	Lyase	Carboanhydrase II	4.2.1.1
1os5	d	Transferase	Hepatitis C virus RNA Polymerase	-.-.-
1okl	d	Lyase	Carboanhydrase II	4.2.1.1
1f2o	d	Hydrolase	Aminopeptidase	3.4.11.-
1ydb	d	Hydrolyase	Carboanhydrase II	4.2.1.1
1fcy	d	Gene regulation	Retinsäurerezeptor gamma-1	-.-.-
1tlp	d	Hydrolase	Thermolysin	3.4.24.27
2tmn	d	Hydrolase	Thermolysin	3.4.24.27
5tmn	d	Hydrolase	Thermolysin	3.4.24.27
1mmp	d	Metalloprotease	Gelatinase a	3.4.24.23
1tmn	d	Hydrolase	Thermolysin	3.4.24.27
1cim	e	Lyase(oxo-acid)	Carboanhydrase II	4.2.1.1
1o2h	e	Hydrolase	Beta Trypsin	3.4.21.4
1rbp	e	Retinol transport	Plasma Retinol-Bindeprotein	-.-.-
1qhc	e	Hydrolase	Ribonuclease A	3.1.27.5
1bq4	e	Isomerase	Phosphoglyceratmutase 1	5.4.2.1
1xhy	e	Membranprotein	Glutamatrezeptor 2	-.-.-
1moq	e	Glutamintransferase	Glucosamin 6-phosphat Synthase	2.6.1.16
2b07	e	Hydrolase	Tyrosin Proteinphosphatase	3.1.3.48
1vjd	e	Transferase	Phosphoglycerat Kinase	2.7.2.3
2bzz	e	Hydrolase	Nichtsekretorische Ribonuclease	3.1.27.5
1v0k	e	Hydrolase	Endo-1,4-beta-Xylanase a	3.2.1.8
1gah	e	Hydrolase	Glucoamylase-471	3.2.1.3
1i9m	e	Lyase	Carboanhydrase II	4.2.1.1
2usn	e	Hydrolase	Stromelysin-1	3.4.24.17
1hp0	e	Hydrolase	Inosin-Adenosin-Guanosine Hydrol.	3.2.2.1
1g48	e	Lyase	Carboanhydrase II	4.2.1.1
1bnt	e	Lyase	Carboanhydrase	4.2.1.1
2c3j	e	Transferase	Ser/Thr Proteinkinase chk1	2.7.1.37

2qwd	e	Hydrolase	Neuraminidase	3.2.1.18
1sld	e	Biotinbindeprotein	Streptavidin	-.-.-
1c1r	f	Hydrolase	Trypsin	3.4.21.4
1t31	f	Hydrolase	Chymase	3.4.21.39
1ezq	f	Hydrolase	Faktor Xa	3.4.21.6
1c5y	f	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1nfu	f	Hydrolase	Faktor Xa	3.4.21.6
1cps	f	Peptidase	Carboxypeptidase a	3.4.17.1
1mq5	f	Hydrolase	Faktor Xa	3.4.21.6
1gai	f	Hydrolase	Glucoamylase-471	3.2.1.3
1h0a	f	Endocytosis	Epsin	-.-.-
1fjs	f	Hydrolase	Faktor Xa	3.4.21.6
1m7y	f	Lyase	Carboxylat Synthase	4.4.1.14
2qwc	f	Hydrolase	Neuraminidase	3.2.1.18
1xpz	f	Lyase	Carboanhydrase II	4.2.1.1
1hi5	f	Hydrolase	Neurotoxin Hydrolase	3.1.27.5
1f0s	f	Hydrolase	Faktor Xa	3.4.21.6
1if7	f	Lyase	Carboanhydrase II	4.2.1.1
1nfy	f	Hydrolase	Faktor Xa	3.4.21.6
1f5l	f	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1gj7	f	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1ela	g	Hydrolase	Elastase	3.4.21.36
1i9l	g	Lyase	Carboanhydrase II	4.2.1.1
2adm	g	Methyltransferase	Methyltransferase taqi	2.1.1.72
1lbf	g	Lyase	Indol-3-glycerol phosphat synthase	4.1.1.48
1bn3	g	Lyase	Carboanhydrase	4.2.1.1
1q66	g	Transferase	Queuine tRNA-Ribosyltransferase	2.4.2.29
7cpa	g	Peptidase	Carboxypeptidase a	3.4.17.1
1ttm	g	Lyase	Carboanhydrase II	4.2.1.1
2bq7	g	Hydrolase	Faktor Xa	3.4.21.6
1cnw	g	Lyase	Carboanhydrase II	4.2.1.1
2hb1	g	Hydrolase	Tyrosin-Protein Phosphatase	Ec 3.1.3.48
1i9p	g	Lyase	Carboanhydrase II	4.2.1.1
1g45	g	Lyase	Carboanhydrase II	4.2.1.1
2i0g	g	Transkription	Östrogenrezeptor beta	-.-.-
1x8t	g	Transferase	Carboxyvinyltransferase	2.5.1.19
1usn	g	Hydrolase	Stromelysin-1	3.4.24.17
1c5t	g	Hydrolase	Trypsin	3.4.21.4
1zc9	g	Lyase	2,2-Dialkylglycine Decarboxylase	4.1.1.64
1qf0	h	Hydrolase	Thermolysin	3.4.24.27
1qq9	h	Hydrolase	Aminopeptidase	3.4.11.-
1nfw	h	Hydrolase	Faktor Xa	3.4.21.6
1rdt	h	Wachstumsfaktorrez.	Retinsäurerezeptor rxr-alpha	-.-.-
1bnu	h	Lyase	Carboanhydrase	4.2.1.1
1g54	h	Lyase	Carboanhydrase II	4.2.1.1
1bnn	h	Lyase	Carboanhydrase	4.2.1.1
2j2u	h	Hydrolase	Faktor Xa	3.4.21.6
2fgi	h	Transferase	Fibroblast Wachstumsfaktor Rezeptor	2.7.1.112
2gzl	h	Lyase	Cyclodiphosphat Lyase	4.6.1.12
1cin	h	Lyase	Carboanhydrase II	4.2.1.1
2j34	h	Hydrolase	Faktor Xa	3.4.21.6
1r0p	h	Transferase	Hepatocyt Wachstumsfaktor Rezeptor	2.7.1.112
1bnv	h	Lyase	Carboanhydrase	4.2.1.1

1y3g	h	Hydrolase	Thermolysin	3.4.24.27
1os0	h	Hydrolase	Thermolysin	3.4.24.27
4tmn	h	Hydrolase	Thermolysin	3.4.24.27
1lor	i	Lyase	Orotidine Decarboxylase	4.1.1.23
1kzn	i	Isomerase	Dna Gyrase	5.99.1.3
1g3d	i	Hydrolase	Beta Trypsin	3.4.21.4
1hi3	i	Hydrolase	Eosinophil Neurotoxin	3.1.27.5
1nl9	i	Hydrolase	Tyrosinphosphatase	3.1.3.48
1bn1	i	Lyase	Carboanhydrase	4.2.1.1
1x38	i	Hydrolase	Exohydrolase Isoenzyme exoi	3.2.1.58
1o2x	i	Hydrolase	Beta Trypsin	3.4.21.4
2qwe	i	Hydrolase	Neuraminidase	3.2.1.18
1x39	i	Hydrolase	Exohydrolase Isoenzyme exoi	3.2.1.58
1no6	i	Hydrolase	Tyrosinphosphatase	3.1.3.48
1b8n	i	Transferase	Purin Nucleoside Phosphorylase	2.4.2.1
2g00	i	Hydrolase	Faktor Xa	3.4.21.6
1o36	i	Hydrolase	Beta Trypsin	3.4.21.4
1v2k	i	Hydrolase	Trypsin	3.4.21.4
1aj6	j	Topoisomerase	Gyrase	5.99.1.3
2fzz	j	Hydrolase	Faktor Xa	3.4.21.6
2sim	j	Hydrolase	Sialidase	3.2.1.18
1m0q	j	Lyase	2,2-Dialkylglycine Decarboxylase	4.1.1.64
1vj9	j	Hydrolase	Plasminogen Aktivator, Urokinase	3.4.21.73
1nq7	j	Transkription	Nuclear Rezeptor ror-beta	-.-.-
1bxr	j	Amidotransferase	Carbamoyl-Phosphat Synthase	6.3.5.5
1g46	j	Lyase	Carboanhydrase II	4.2.1.1
1cnx	j	Lyase	Carboanhydrase II	4.2.1.1
1db1	j	Gene regulation	Vitamin D Nucleärer Rezeptor	-.-.-
1xka	j	Hydrolase	Faktor Xa	3.4.21.6
1add	j	Hydrolase	Adenosine deaminase	3.5.4.4
1rd4	j	Immunsystem	Integrin alpha-l	-.-.-
1jao	j	Metalloprotease	Matrix metallo proteinase-8	3.4.24.34
1z6e	k	Hydrolase	Faktor Xa	3.4.21.6
2d3z	k	Transferase	Polyprotein	2.7.7.48
1w2g	k	Transferase	Thymidylate Kinase tmk	2.7.4.9
2gvv	k	Hydrolase	Phosphotriesterase	3.1.8.2
2ada	k	Hydrolase	Adenosine Deaminase	3.5.4.4
1hi4	k	Hydrolase	Eosinophil-derived neurotoxin	3.1.27.5
2fxu	k	Strukturprotein	Actin, Alpha Skelettmuskel	-.-.-
1z9g	k	Hydrolase	Thermolysin	3.4.24.27
2br1	k	Transferase	Ser/Thr Proteinkinase chk1	2.7.1.37
1h1p	k	Transferase	Zellteilung Proteinkinase 2	2.7.1.-
2fqx	l	Transportprotein	Membranlipoprotein tmpc	-.-.-
2brm	l	Transferase	Ser/Thr Proteinkinase chk1	2.7.1.37
1erb	l	Retinol transport	Retinol Bindeprotein	-.-.-
1q54	l	Isomerase	Isopentenyl Diphosphat Isomerase	5.3.3.2
1fd0	l	Gene regulation	Retinsäurerezeptor gamma-1	-.-.-
1qji	l	Metalloproteinase	Astacin	3.4.24.21
1mmq	l	Metalloprotease	Matrilysin	3.4.24.23
2qwf	l	Hydrolase	Neuraminidase	3.2.1.18
2c02	l	Hydrolase	Nonsekretorische Ribonuclease	3.1.27.5
1qan	m	Transferase	Ermc Methyltransferase	2.1.1.48
1hlk	m	Hydrolase	Beta-Lactamase, type II	3.5.2.6

2bz6	m	Hydrolase	Faktor VIIa	3.4.21.21
2boh	m	Hydrolase	Faktor Xa	3.4.21.6
1h1s	m	Transferase	Zellteilung Proteinkinase 2	2.7.1.-
2j47	m	Inhibitor	Glucosaminidase	3.2.1.52
1m0o	m	Lyase	2,2-Dialkylglycine Decarboxylase	4.1.1.64
1n4h	m	Kenrezeptor	Nuclearrezeptor ror-beta	-.-.-
1lke	n	Ligandenbindeprotein	Diga16	-.-.-
1o3p	n	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
2d1n	n	Hydrolase	Collagenase 3	3.4.24.-
2fzk	n	Transferase	Aspartate Carbamoyltransferase	2.1.3.2
1hfs	n	Hydrolase	Stromelysin-1	3.4.24.17
1dqn	n	Transferase	Guanine Phosphoribosyltransferase	2.4.2.8
2h4n	o	Lyase	Carboanhydrase II	4.2.1.1
1bn4	o	Lyase	Carboanhydrase	4.2.1.1
1i9n	o	Lyase	Carboanhydrase II	4.2.1.1
1g53	o	Lyase	Carboanhydrase II	4.2.1.1
1gi7	o	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1xq0	p	Lyase	Carboanhydrase II	4.2.1.1
1fcx	p	Gene regulation	Retinsäurerezeptor Gamma-1	-.-.-
1xbb	p	Transferase	Tyrosine-Proteinkinase syk	2.7.1.112
1u33	p	Hydrolase	Pankreatische Alpha-Amylase	3.2.1.1
1q63	p	Transferase	Queuine trna-Ribosyltransferase	2.4.2.29
1n2v	q	Transferase	Queuine trna-Ribosyltransferase	2.4.2.29
1if8	q	Lyase	Carboanhydrase II	4.2.1.1
1f2p	q	Hydrolase	Aminopeptidase	3.4.11.-
2f7p	q	Hydrolase	Alpha-Mannosidase II	3.2.1.114
1gjc	r	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
1a42	r	Lyase	Carboanhydrase II	4.2.1.1
1bnw	r	Lyase	Carboanhydrase	4.2.1.1
1lpz	s	Hydrolase	Faktor Xa	3.4.21.6
1ta6	s	Hydrolase	Thrombin	3.4.21.5
1b1h	s	Signalprotein	Oligo-Peptidbindeprotein	-.-.-
1g52	t	Lyase	Carboanhydrase II	4.2.1.1
2hxm	t	Hydrolase	Uracil-DANN Glycosylase	3.2.2.-
1g4j	t	Lyase	Carboanhydrase II	4.2.1.1
1gja	u	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
2bfq	u	Hydrolase	Hypothetisches Protein af1521	-.-.-
1gja	v	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
2bfq	v	Hydrolase	Hypothetisches Protein af1521	-.-.-
1gja	w	Hydrolase	Urokinase-Typ Plasminogen Aktivator	3.4.21.73
2bfq	w	Hydrolase	Hypothetisches Protein af1521	-.-.-

8 Eidesstattliche Erklärung

Die vorliegende Dissertation wurde selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Alle Stellen die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche gekennzeichnet.

Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung nicht vorgelegt worden.

Frankfurt am Main, den ……... ……
(Martin Weisel)

9 Lebenslauf

Zur Person

Name	Martin Weisel
Geburtsdatum	09.06. 1980
Geburtsort	Gießen
Staatsangehörigkeit	deutsch

Schulische Ausbildung

1986 – 1992	Wilhelm-Leuschner-Schule Heuchelheim
1992 – 1999	Herderschule Gießen Abschluss: Abitur

Zivildienst

1999 – 2000	St. Josefs Krankenhaus Gießen

Hochschulbildung

2000 – 2006	**Studium der Bioinformatik** an der Goethe-Universität in Frankfurt am Main Abschluss: Diplom-Bioinformatiker
10/05 – 04/06	**Diplomarbeit** in der Arbeitsgruppe von Prof. Dr. Gisbert Schneider, Goethe-Universität Frankfurt am Main „Entwicklung einer Rezeptor-basierten Pharmacophor-funktion in PyMOL“
09/2006 – 04/2009	**Promotion** in der Arbeitsgruppe von Prof. Dr. Gisbert Schneider, Goethe-Universität Frankfurt am Main *„Analyse von Form, Eigenschaften und Druggability von Proteinbindetaschen“*

seit 04/06	Wissenschaftlicher Mitarbeiter am Lehrstuhl für Chemie- und Bioinformatik, Goethe-Universität Frankfurt am Main

10 Publikationsliste

Im Rahmen dieser Arbeit sind folgende Publikationen entstanden:

6. Weisel, M., Kriegl, J.M., Schneider, G. PocketGraph: Analysis of Binding Site Topologies using Growing Neural Gas Networks, 2009, *manuscript in preparation*.
5. Nietert, M., Weisel, M., Proschak, E., Kestner, E., Gohlke, H., Schneider, G. Pocket Surface Dynamic Ligand Binding Pockets for Structure-Based Virtual Screening, 2009, *manuscript in preparation*.

4. Stauch, B., Hofmann, H., Weisel, M., Cichutek, K., Münk, C., Schneider, G. Model Structure of APOBEC3C Supports a Critical Role of Protein Dimerization for Cell-Intrinsic Antiviral Activity, *Proc. Natl. Acad. Sci. USA* 2009, *in revision*.

3. Weisel, M., Proschak, E., Kriegl J.M., Schneider, G. Form Follows Function: Shape Analysis of Protein Cavities for Receptor-based Drug Design, *Proteomics* 2009, **9**, 451-489.

2. Proschak, E., Zettl, H., Tanrikulu, Y., Weisel, M., Kriegl, J.M., Rau, O., Schuber-Zsilavecz, M., Schneider, G. From molecular shape to potent bioactive agents I: Bioisosteric replacement of molecular fragements, *ChemMedChem.* 2009, **9**, 41-44.

1. Weisel, M., Proschak, E., Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors, *Chem. Cent. J.* 2007, 1:7.

Ergebnisse aus dieser Arbeit wurden auf folgenden Postern präsentiert:

12. Nietert, M., Brecht, M., Weisel, M., Zacharias, M., Schneider, G., Göringer, H.U. RNA editing as a drug target – decomposition of RNA/inhibitor interaction, *CRC 579 Symposium*, March 23-25, 2009, Frankfurt am Main – Germany.

11. Klenner, A., Reisen, F., Hartenfeller, M., Weisel, M., Tanrikulu, Y., Rupp, M., Schneider, G. Structure based virtual screening for RNA ligands, *CRC 579 Symposium*, January 23-25, 2009, Frankfurt am Main – Germany.

10. Weisel, M., Kriegl, J.M., Schneider, G. PocketGraph: Graph Representation of Binding Site Volumes, *4th German Conference on Chemoinformatics*, November 9-11, 2008, Goslar - Germany.

9. Löwer, M., Tanrikulu, Y., Weisel, M., Weydig, C., Wessler, S., Schneider, G. Fuzzy Virtual Ligands for Virtual Screening. *Chemical Computing Group European User Group Meeting 2008*, October 21-22, Hinxton Hall – United Kingdom.

8. Klenner, A., Becker, S., Nietert, M., Hartenfeller, M., Schüller, A., Weisel, M., Proschak, E., Tanrikulu, Y., Schneider, G. RNA Ligand Design, *CRC 579 Symposium*, March 4-5, 2008, Frankfurt am Main – Germany.

7. Becker, S., Klenner, A., Nietert, M., Hartenfeller, M., Nietert, M., Proschak, E., Schüller, A., Tanrikulu, T., Weisel, M. Molecular Recognition in RNA-Ligand Complexes and Molecular Design, *CRC 579 Symposium*, March 4-5, 2008, Frankfurt am Main - Germany.

6. Proschak, E., Tanrikulu, Y., Rau, O., Zettl, H., Weisel, M., Schubert-Zsilavecz, M., Schneider, G. The SQUIRREL Hops Over the Scaffolds: Novel PPARα antagonists, *Frontiers of Medicinal Chemistry*, March 2-5, 2008, Regensburg - Germany.

5. Weisel, M., Kriegl, J., Schneider, G. The Druggable Pocketome: Predicting the Suitability of Protein Cavities for Drug-Design, *3rd German Conference on Chemoinformatics*, November 11-13, 2007, Goslar - Germany.

4. Proschak, E., Tanrikulu, Y., Hofmann, B., Rau, O., Zettl, H., Weisel, M., Kriegl, J., Steinhilber, D., Schubert-Zsilavecz, M., Schneider, G. The SQUIRREL Hops Over the Scaffolds: Application of a Novel Virtual Screening Tool, *3rd German Conference on Chemoinformatics*, November 11-13, 2007, Goslar - Germany.

3. Nietert, M., Weisel, M., Proschak, E., Kestner, E., Gohlke, H., Schneider, G. Automated Prediction of Putative Binding Sites in Flexible RNA-Targets, *International Symposium "RNA-Ligand-Interactions"*, September 27-29, 2007, Frankfurt am Main - Germany.

2. Weisel, M., Proschak, E., Kriegl, J., Schneider, G. PocketPicker: Introduction of a Pocket Prediction Method for Ligand Binding Site Analysis with Shape Descriptors, *4th Joint Conference on Chemoinformatics*, June 18-20, 2007, Sheffield - United Kingdom.

1. Weisel, M., Proschak, E., Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape-Descriptors, *2nd German Conference on Chemoinformatics*, November 12-14, 2006, Goslar - Germany.

Die Publikationen Weisel *et al.*, 2007, sowie Weisel *et al*., 2009 sind auf den folgenden Seiten abgedruckt.

Chemistry Central Journal

ChemistryCentral

Research article

Open Access

PocketPicker: analysis of ligand binding-sites with shape descriptors

Martin Weisel, Ewgenij Proschak and Gisbert Schneider*

Address: Johann Wolfgang Goethe-Universität, Beilstein Endowed Chair for Cheminformatics, Institut für Organische Chemie und Chemische Biologie, Siesmayerstr. 70, D-60323 Frankfurt am Main, Germany

Email: Martin Weisel - Martin.Weisel@bioinformatik.uni-frankfurt.de; Ewgenij Proschak - Proschak@bioinformatik.uni-frankfurt.de; Gisbert Schneider* - g.schneider@chemie.uni-frankfurt.de

* Corresponding author

Published: 13 March 2007

Chemistry Central Journal 2007, **1**:7 doi:10.1186/1752-153X-1-7

Received: 15 December 2006
Accepted: 13 March 2007

This article is available from: http://journal.chemistrycentral.com/content/1/1/7

Abstract

Background: Identification and evaluation of surface binding-pockets and occluded cavities are initial steps in protein structure-based drug design. Characterizing the active site's shape as well as the distribution of surrounding residues plays an important role for a variety of applications such as automated ligand docking or *in situ* modeling. Comparing the shape similarity of binding site geometries of related proteins provides further insights into the mechanisms of ligand binding.

Results: We present PocketPicker, an automated grid-based technique for the prediction of protein binding pockets that specifies the shape of a potential binding-site with regard to its buriedness. The method was applied to a representative set of protein-ligand complexes and their corresponding *apo*-protein structures to evaluate the quality of binding-site predictions. The performance of the pocket detection routine was compared to results achieved with the existing methods CAST, LIGSITE, LIGSITE^{cs}, PASS and SURFNET. Success rates PocketPicker were comparable to those of LIGSITE^{cs} and outperformed the other tools. We introduce a descriptor that translates the arrangement of grid points delineating a detected binding-site into a correlation vector. We show that this shape descriptor is suited for comparative analyses of similar binding-site geometry by examining induced-fit phenomena in aldose reductase. This new method uses information derived from calculations of the buriedness of potential binding-sites.

Conclusion: The pocket prediction routine of PocketPicker is a useful tool for identification of potential protein binding-pockets. It produces a convenient representation of binding-site shapes including an intuitive description of their accessibility. The shape-descriptor for automated classification of binding-site geometries can be used as an additional tool complementing elaborate manual inspections.

Background

Accurate structural information of validated target proteins provides a basis for the design and development of novel therapeutic agents. The increased number of high resolution protein structures available from the RCSB Protein Databank (PDB) [1] has opened new opportunities for structure-based rational drug design [2,3]. Still, the identification of potential protein binding pockets and occluded cavities remains a central issue, as the capability to interact with other proteins or small ligands determines the biological function of a protein. The size and shape of ligand binding sites and the distribution of functional

groups in these pockets are of particular interest for the design of selective low-molecular weight ligands. This renders binding-site analysis pivotal for rational drug design, such as ligand docking or *de novo* molecular design. These methods require exact structural information of the binding-site as a starting-point.

A variety of computational methods already exists for the location of possible ligand binding-sites. Most of these pocket detection algorithms solely rely on geometric criteria to find clefts and surface depressions. Empirical studies show that the actual ligand binding-site usually coincides with the largest pocket of a protein's surface [4,5]. The program SURFNET [6] successfully predicted the ligand binding-site as the biggest pocket in 83% of the cases on a test set of 67 single-chain enzymes [7]. SURFNET identifies voids between two or more molecules as well as internal cavities and pockets by fitting virtual spheres into the solvent-accessible space between protein atoms. So-called "initial gap spheres" are placed midway between the vander-Waals surfaces of two atoms and scaled down when penetrated by neighboring atoms. All remaining gap spheres exceeding a minimal predefined radius (default is 1.0 Å) are denoted as "final spheres" and used to define surface pockets and cavities (Figure 1).

The program CAST [8,9] uses an approach based on alpha shapes [10,11] and triangulations of complex shapes. This method makes use of the concepts of Voronoi diagrams [12] and Delaunay [13] triangulations. The pocket prediction process of CAST specifies the calculation of the so-called "dual complex" (or alpha shape) and is summed up for a simplified two-dimensional depiction of binding site atoms (Figure 2). The procedure includes the calculation of the Voronoi diagram which consists of Voronoi cells (Figure 2A). Each Voronoi cell contains one protein atom and controls all spatial points that are closest to the respectively considered atom. The Voronoi diagram is mathematically equivalent to the Delaunay triangulation of the complex hull drawn around the protein atom centers (Figure 2B). The Delaunay triangulation can be obtained directly from the Voronoi diagram. Therefore a line is drawn across every Voronoi edge separating two Voronoi cells connecting the two corresponding atoms centers. For each Voronoi vertex where three Voronoi cells meet, a Delaunay triangle is placed connecting the three atom centers of the considered cells. To obtain the dual complex, Voronoi edges and vertices are disregarded in the triangulation, if they are situated completely or in part outside of the molecule (Figure 2B, grey lines). A triangle with one ore more omitted edges is denoted as "empty". Neighboring empty triangles are combined in the "disrete-flow" method to outline continuous voids in the protein surface. In the course of this process an obtuse empty triangle flows to its neighboring triangle, whereas acute empty triangles act as sinks to collect the flow of neighboring triangles (Figure 2C). CAST was tested on 51 of 67 monomeric complexes used for SURFNET [6] and achieved a success rate of 74%.

PASS [14] (Putative Active Sites with Spheres) uses an iterative placing of probe spheres to identify surface concavities. An initial layer of probe spheres coating the entire protein surface is created in the first step (Figure 3). For each probe sphere a "burial count" is calculated which gives the number of protein atoms within a preset radius of 8 Å. This measurement is used to identify probe spheres located in protein surface pockets and cavities. Probe spheres residing in convex parts of the surface are omitted from further calculations. Additional layers are then accreted to the remaining probe set to completely fill protein cavities with probe spheres. A "probe weight" is calculated for each probe sphere of the final set comprising the burial count and the number of neighboring probe spheres. Finally, a small number of "active site points" (ASPs) is selected to represent the centers of potential binding pockets. ASPs are identified by picking central probes from regions containing many spheres of high burial counts. Putative binding sites are defined by keeping a reduced set of ASPs separated by a minimum threshold of 8 Å. Pockets are ranked by the probe weights of their corresponding ASPs. PASS yielded correct predictions of 63% on a set of 30 complexed structures and 60% for a test set of 20 *apo*-protein structures.

Another pocket detection method is POCKET [15]. This algorithm operates on a rectangular grid, which is constructed around the protein and denotes grid points as either solvent-accessible or inaccessible to the solvent. The program searches for cavities by scanning along the *x*-, *y*- and *z*-axes to locate groups of solvent-accessible grid points that are enclosed by grid points not accessible to solvent on both sides (Figure 4). Such arrangements were denoted as PSP events (protein-solvent-protein). Results of POCKET may be unsatisfying as pockets with an orientation of 45° to the orthogonal axes will not be properly detected or even be totally ignored. To compensate for this deficiency LIGSITE [16] was developed as an extension to POCKET. In this approach the scanning process was extended to the four cubic diagonals so that a proper pocket prediction became possible, which is independent from the orientation of the protein in the grid. LIGSITE^{cs} and LIGSITE^{csc} were introduced as enhanced implementations of the original LIGSITE [16] algorithm and resulted in improved pocket prediction results [17]. LIGSITE^{cs} (cs = Conolly Surface) differs from the original LIGSITE [16] method by capturing surface-solvent-surface events using the protein's Conolly surface instead of detecting protein-solvent-protein events. LIGSITE^{csc} (csc = Conolly Surface and Conservation) performs a re-ranking of the top-three

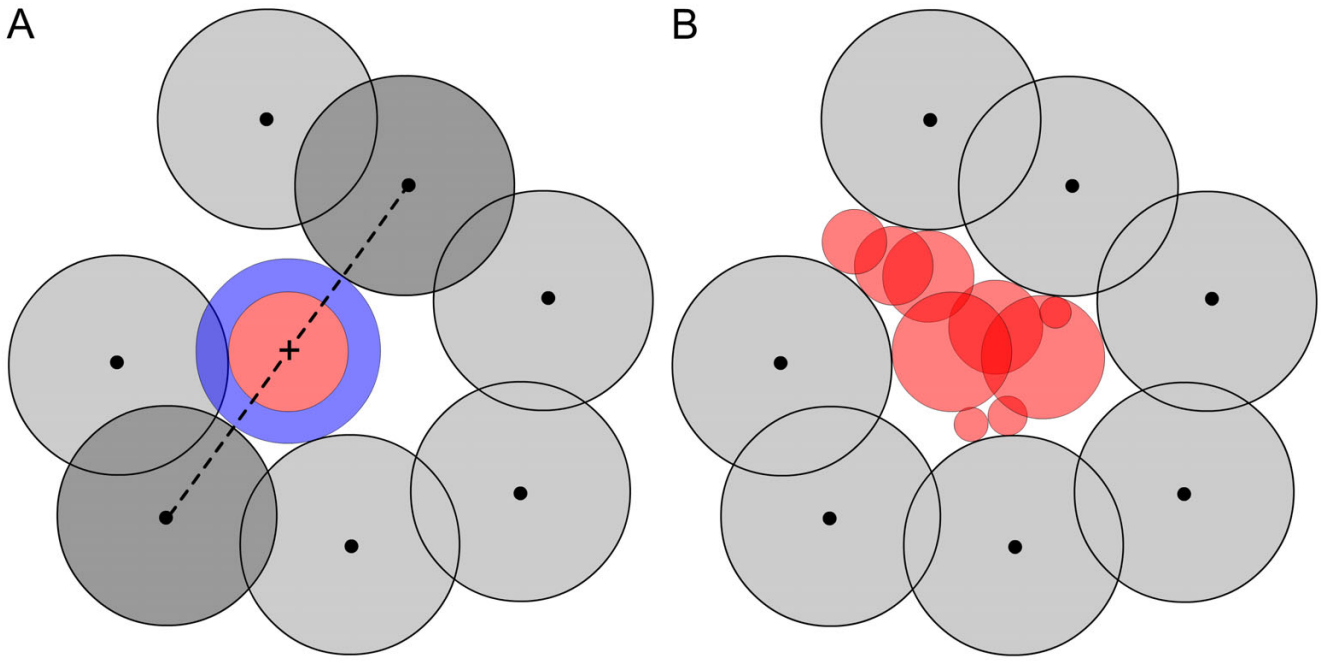


Figure 1
Two-dimensional depiction of the pocket detection process of SURFNET. **A:** An initial gap sphere (blue disc) is placed midway between the van der Waals surfaces of a pair of atoms. The radius of this gap sphere is then reduced until it is not penetrated by any of the neighboring atoms. The resulting final gap sphere is shown in red. **B:** The arrangement of final gap spheres is used to describe the shapes and sizes of protein cavities in SURFNET.

predicted pockets by the degree of conservation of the closest surface residues. The average conservation of the residues within 8 Å of the center of a predicted pocket is used as a conservation score applied for re-ranking. Note that LIGSITE^{csc} is not a purely geometric approach to pocket prediction as it considers conservation scores obtained from the ConSurf-HSSP [18] database as an additional source of information. A refinement of the pre-

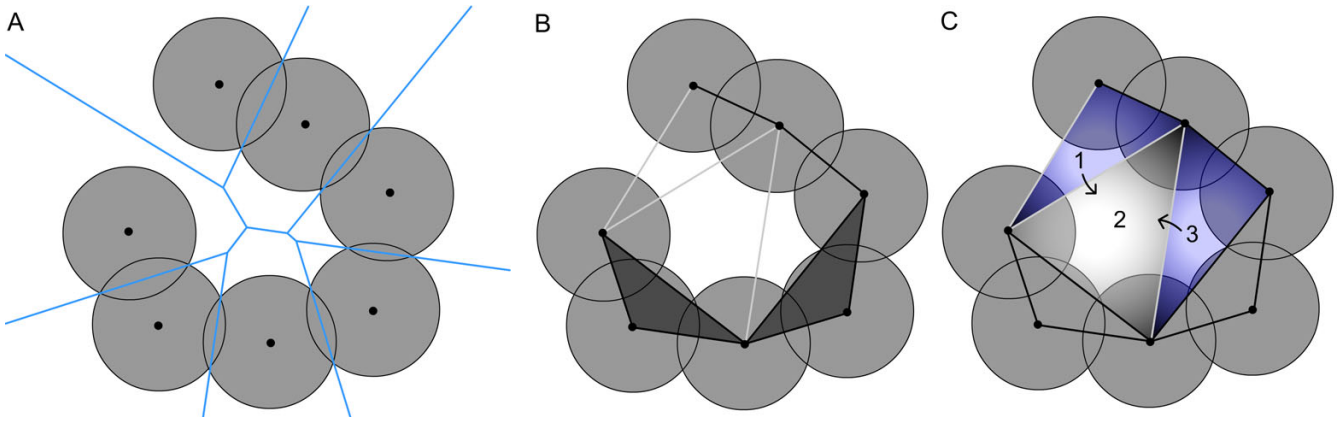


Figure 2
Illustration of the alpha shape theory and discrete-flow method used in CAST. **A:** Two-dimensional depiction of pocket atoms represented as disks of uniform radii. The blue lines show the Voronoi diagram for the pocket atoms. **B:** The seven bordering lines running through the atom centers represent the convex hull, which is triangulated into Delaunay triangles using information of the Voronoi diagram. The "alpha shape" or "dual complex" is defined by the shaded triangles and the black lines. Three "empty triangles" having at least one grey bordering line are shown. **C:** Two obtuse empty triangles (1, 3) are assigned to the obtuse triangle (2) by the discrete-flow method.

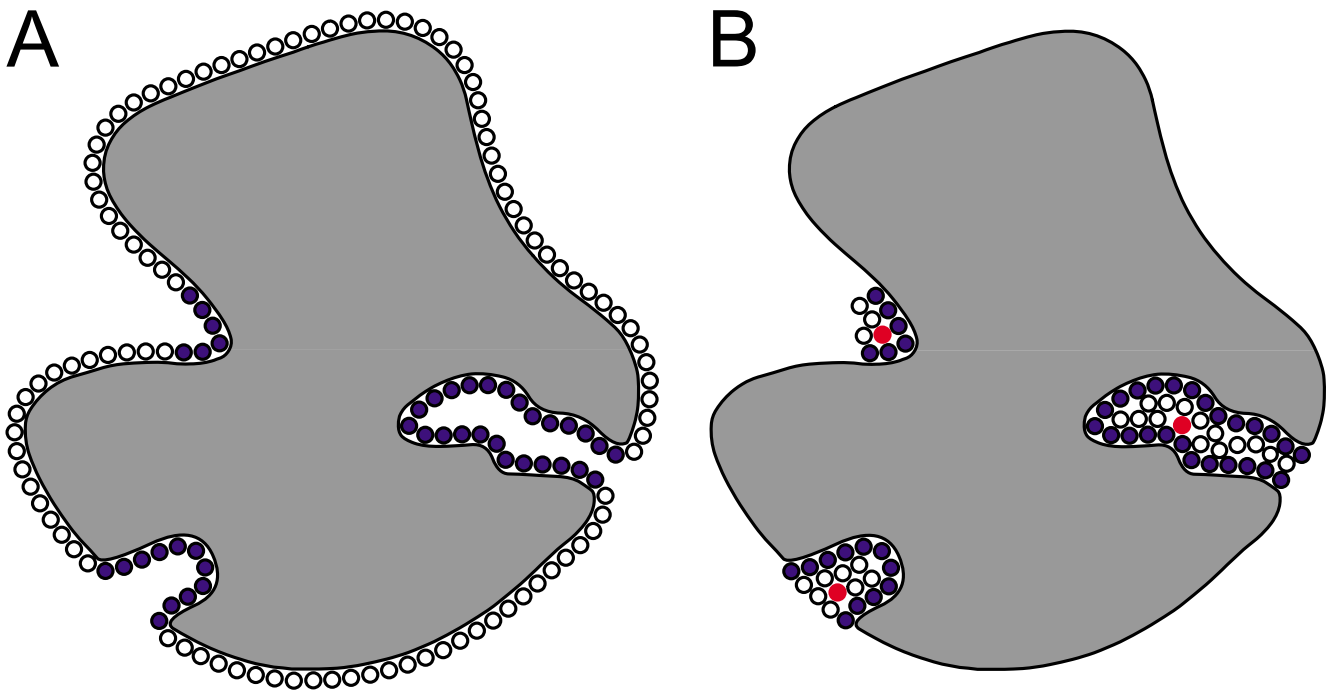


Figure 3
Placement of spheres for a two-dimensional molecule in PASS. **A:** The entire surface of the molecule is coated with virtual spheres and an initial layer of spheres residing in buried parts of the protein is specified (blue shaded circles). **B:** Additional layers are attached onto the initial layer in an iterative process and active site points (red disks) are exposed for potential binding pockets.

dictions made by SURFNET [6] using conservation scores for re-ranking is also available from a subsequent recent study [19].

Further algorithms exclusively operating on geometric criteria are Cavity Search [20], VOIDOO [21], APROPOS [22], and Travel Depth [23]. DrugSite [24] and PocketFinder [25] evaluate shape and physicochemical properties for identification of ligand binding envelopes. An energy-based method for protein pocket detection is Q-SiteFinder [26], which uses the interaction energy between the protein and a van der Waals probe to detect energetically favorable binding sites.

In this study, we present a new geometric pocket prediction method that translates the form and accessibility of identified binding-sites into correlation vectors for rapid pocket comparisons. A similar approach was pursued by Stahl *et al*. with the aim to classify matrix metalloproteinase active sites [27]. The pocket detection routine is based on a regular rectangular grid and employs a sophisticated scanning process to locate protein surface depressions. The scanning procedure comprises the calculation of "buriedness" of probe points installed in the grid to determine their atom environment. The buriedness of grid points is interpreted as a pocket accessibility index. The enhanced information content of both the buriedness and the shape of a predicted binding pocket is summarized in a shape descriptor. This descriptor has been designed to conduct automated comparisons between different binding-site conformations. The essential steps of our method can be summed up as follows:

(1) Calculation of buriedness values of grid probes installed in areas closely above the protein surface.

(2) Clustering of adjoining grid probes indicating buried regions of the structure to find potential binding-sites.

(3) Preparation of shape descriptors to enable comparisons of different pocket shapes.

Materials and computational methods

Protein data collection

To evaluate the accuracy of binding site predictions performed by PocketPicker we used a test set comprising 48 ligand-receptor complexes from the RCSB Protein Database (PDB [1]) as well as their corresponding 48 unbound *apo*-forms. This test set was presented in a previous study [17] to compare success rates of pocket predictions by the

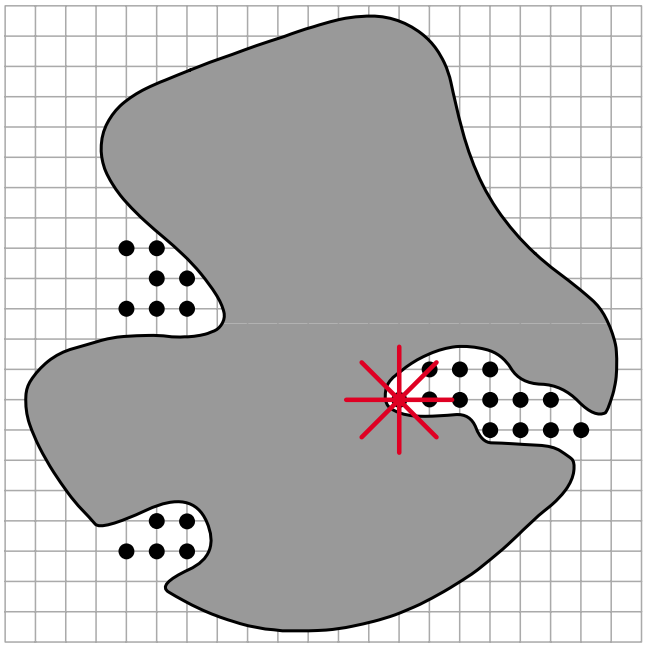


Figure 4
Pocket detection method used in POCKET, LIGSITE and its derivatives. Grid probes are installed at the edges of an artificial grid generated around the protein (shaded area). A scanning process is applied to detect protein-solvent-protein events (POCKET and LIGSITE) or surface-solvent-surface events (LIGSITE^{cs} and LIGSITE^{csc}).

programs CAST, PASS, SURFNET, LIGSITE, LIGSITE^{cs}, and LIGSITE^{csc}. We used this protein collection to validate the predictions made by PocketPicker compared to the findings of these algorithms. All protein structures were downloaded from the RCSB PDB database [1], and ligands denoted with the HET (heteroatom) identifier were removed from each PDB-file prior to computations. Binding site predictions were carried out for monomeric structures (results for protein multimers are provided as additional files [see Additional files 1, 2]). Unbound structures were aligned with the corresponding complex using the "align" command of PyMOL [28]. Structural alignments were performed to compare active site predictions for the unbound structures with the actual binding pocket given by the protein-ligand complex.

The capability of comparing induced-fit phenomena with the proposed shape descriptor was tested on a set of 13 aldose reductase crystal structures discussed by Sotriffer and coworkers [29]. This selection contained nine structures of human aldose reductase: 1ads, 1el3, 1iei, 1us0, 2acq, 2acr, 2acs, the Tyr48His mutant 2acu, and the Cys298Ala/Trp219Tyr double mutant 1az1. Additional four structures were from the porcine enzyme and carried one mutation each: 1ah0, 1ah3, 1ah4, and 1eko. The crystal structure of 1us0 with an ultrahigh resolution of 0.66 Å served as a reference. All selected structures shared a sequence identity of ≥ 85% with the reference and had resolution of at least 2.5 Å. Coordinates of 1ah0, 1ah3, 1ah4 and 1eko were rotated -45° around the *z*-axis to meet the orientation of the other aldose reductase structures. Pocket predictions were performed for structures in complex with the cofactor NADPH or $NADP^+$. All other ligands were removed prior to computation.

Strategy for identification of surface pockets and cavities

A rectangular grid with 1 Å mesh size is generated around the protein, adjusted to its spatial extent. The pocket detection routine is focused on grid points that are located closely above the protein surface: grid points that exceed a maximal distance of 4.5 Å to the closest protein atom or are situated under the protein surface are excluded from further calculations (Figure 5a). Note that these areas can be omitted from further investigation, since they are not relevant for pocket detection. Probes are attached to the remaining grid points to examine their accessibility on the protein surface.

The buriedness value indicates whether a grid point is situated next to a convex part of the surface or locates in a less accessible part of the surface. This information can be used for the identification of clefts and surface concavities: A straightforward clustering algorithm is applied to combine neighboring grid points with an appropriate buriedness-index into disjoint groups highlighting those parts of the grid located in less accessible parts of the protein surface (Figure 5b). Cavities and pockets identified in this manner are afterwards sorted by the number of the consisting grid points to specify the largest existing protein concavity.

Calculation of buriedness

The buriedness-index is calculated by investigating the molecular environment of a grid probe in an elaborate scanning process: Scans are being performed along 30 directions that are approximately equally distributed around a grid probe. The optimal distribution of vectors in three-dimensional space is not a trivial problem and resembles the task of equally distributing points on a sphere [30]. In fact, there are only three completely symmetric arrangements of points ($n > 2$) on the sphere: The vertices of the tetrahedron, the octahedron and the icosahedron are equally distributed [31] on a commemorated sphere (Figure 6a).

We use a series of triangulations to subdivide the eight faces of the octahedron in order to arrange three additional vectors on each face (Figure 6b). These newly added vectors are elongated toward the surface of a virtual sphere to adopt the length of the primary vectors of the octahe-

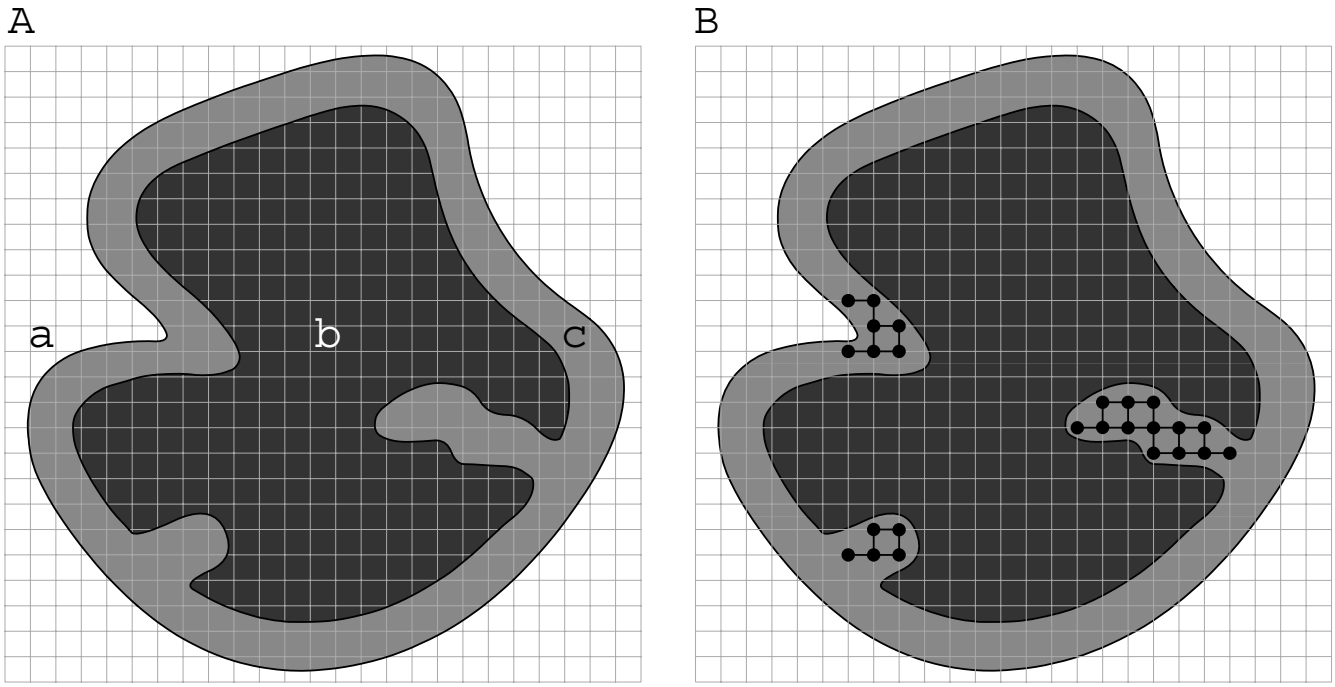


Figure 5
Schematic view of the pocket detection process of PocketPicker. **A:** Grid points located far off the protein (a) or hidden under the surface (b) are excluded from calculations. Buriedness values are calculated solely for grid points close to the protein surface (c). **B:** Grid probes indicating surface depressions are collected in clusters.

dron running along the Cartesian axes (Figure 6c). The 30 vectors created in this manner can be reflected in the *x, z*-plane which is required for a subsequent part of the computation.

The accessibility of a grid probe is calculated by scanning the molecular surrounding along 30 search rays of length 10 Å and width 0.9 Å. Whenever a protein atom is encountered within the dimensions of a search ray, the buriedness-index of the probe is increased by one and the next direction vector is regarded. As a result, the calculated indices range from 0 to 30 indicating a growing buriedness of the probe in a protein. The clustering of grid probes for pocket identification is restricted to those probes with buriedness-indices ranging from 16 to 26.

Direction vectors $\vec{u}$ are aligned along 30 straight lines *G* arranged by octahedron triangulation and scaled to the length of one. Search rays scanning the molecular environment of a grid probe *P* (represented by vector $\vec{p}$) are arranged along the direction vectors and scaled to the proposed dimensions. A neighboring protein atom *Q* ($\vec{q}$) is detected during scanning when the length of its orthogonal projection *d* onto the actual direction vector does not exceed the preset width of the search ray.

$$d = \left|(\vec{q} - \vec{p}) \times \vec{u}\right|, \text{ if } \left|\vec{u}\right| = 1$$

The projected point *X* has to reside within the length of the search ray. The distance between *X* and the actual grid point *P* can be determined as the length of the direction vector $\vec{u}$ scaled by *t*.

Factor *t* was calculated as follows:

$$t = (\vec{q} - \vec{p}) \cdot \vec{u}, \text{ if } \left|\vec{u}\right| = 1$$

The scanning process is summarized in Figure 7a. Position vectors of all atoms and grid points use the Cartesian origin *O* as their reference point.

In order to avoid distance calculations to all protein atoms, the search grid is subdivided into smaller cuboidal compartments of same size, and represented by centroids denoting the geometric center of a cuboid. In a first step, neighboring centroids are detected in an extended search radius along the actual direction vector. Distance calculations are then performed solely to protein atoms assigned to the cuboids of the regarded centroids (Figure 7b).

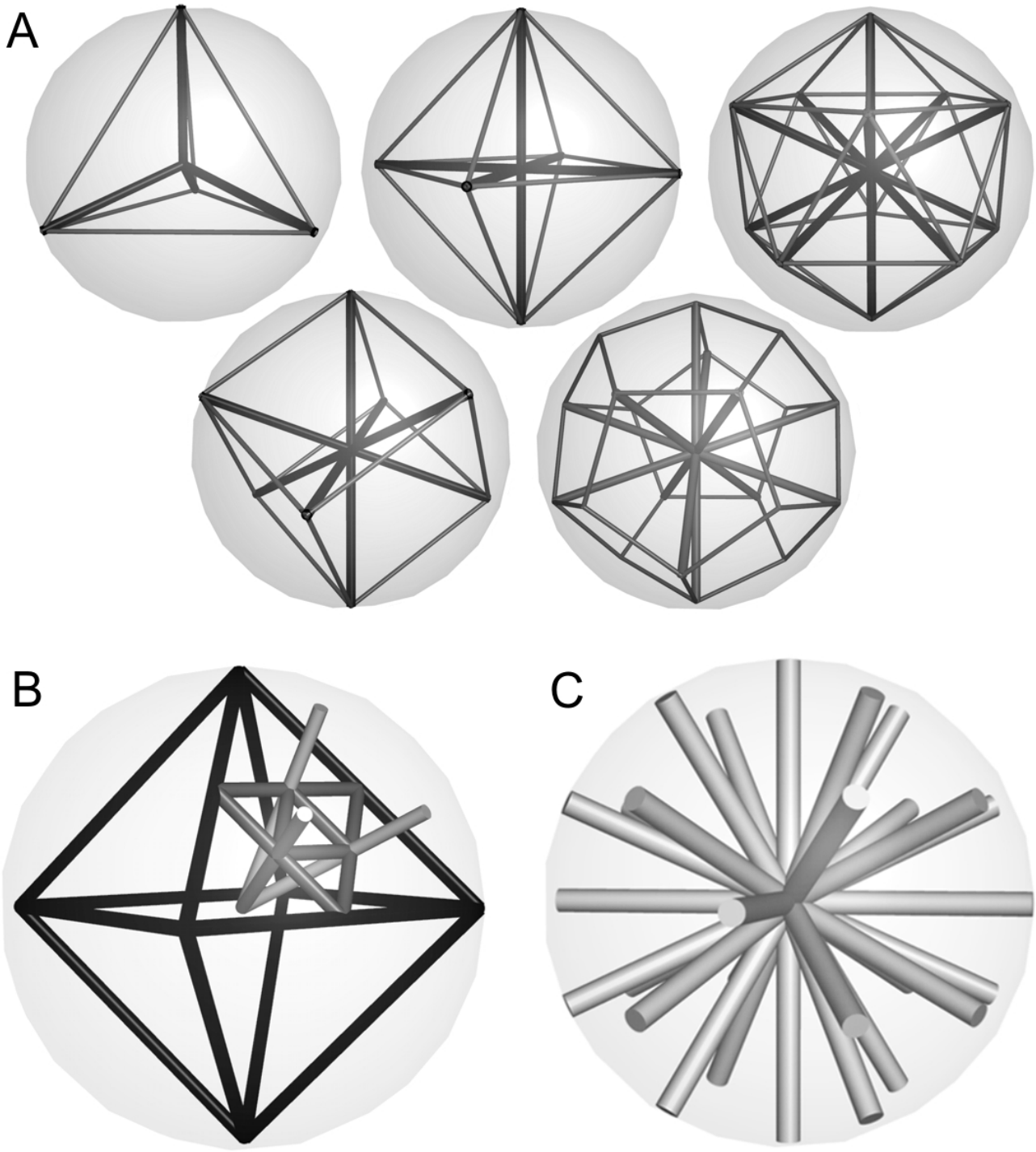


Figure 6
Triangulations of the sphere. **A:** The five Platonic bodies offer a symmetric decomposition of the sphere, but only the tetrahedron, the octahedron and the icosahedron (upper row) describe an exact spherical equidistribution of vectors. **B:** Triangulation of the octahedron was used to arrange additional vectors on the sphere. **C:** Distribution of 30 search rays obtained from octahedron triangulation.

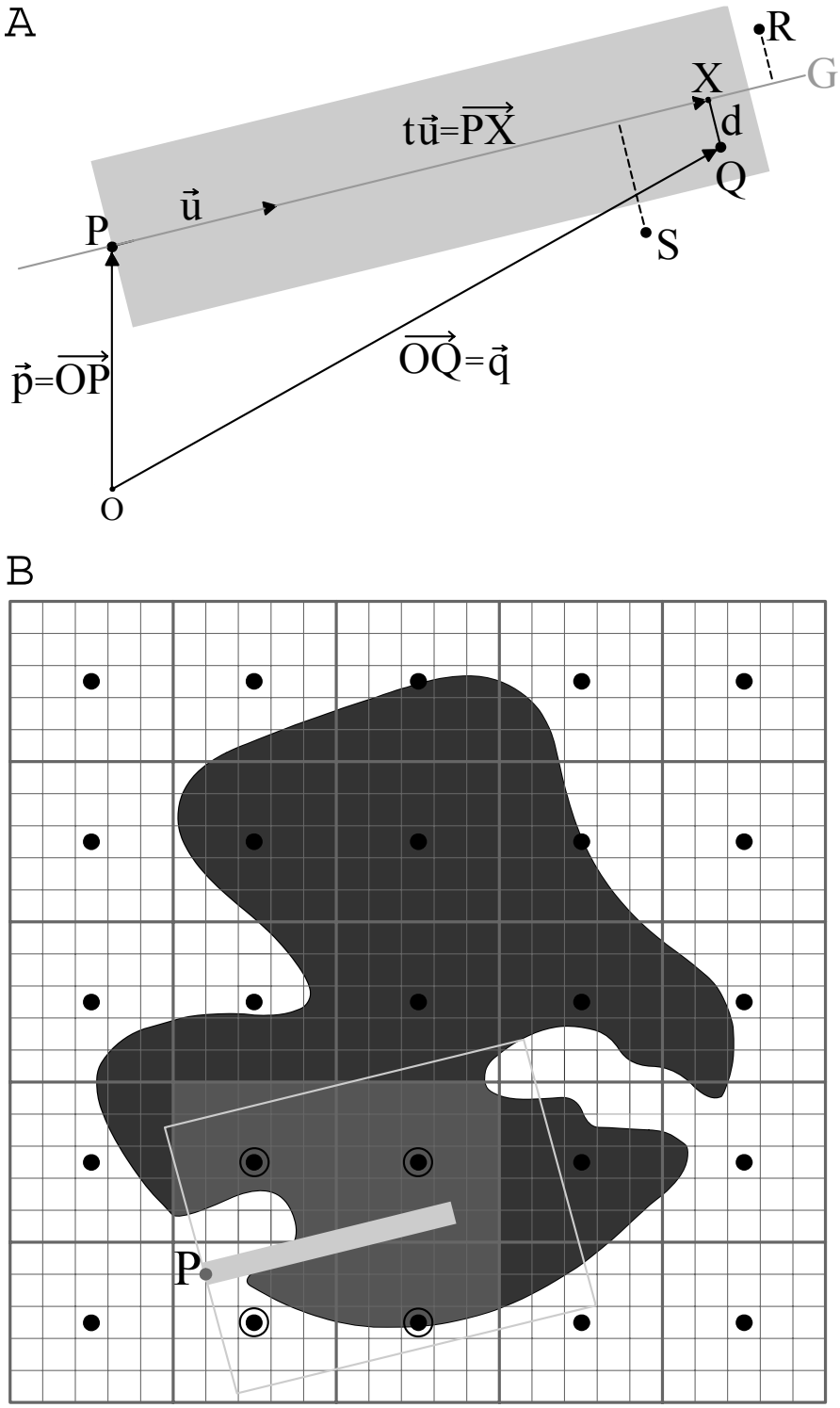


Figure 7
Calculation of the buriedness-index of a grid point *P*. **A:** A search ray (grey plane) scans the room for atoms. Atom *Q* is detected, since it is located within the dimensions of the search vector. Atoms *R* and *S* are not detected, since they are not covered by the search vector. **B:** Distance calculations are restricted to areas controlled by neighboring centroids (encircled). Neighboring centroids are identified by scanning an extended search space (grey border).

Table 1: Collection of 48 complexes and their corresponding *apo*-forms to evaluate pocket prediction results.

Complex	Unbound	Protein Description	Pocket Ligand[1]	Other Ligands[2]
1bid	3tms	Thymidylate synthase	UMP	CBX
1cdo	8adh	Alcohol dehydrogenase	NAD	zn
1dwd	1hxf	Alpha thrombin	MID	chains i, l
1fbp	2fbp	Fructose 1,6-bisphosphatase	AMP	F6P, mg
1gca	1gcg	Glucose/galactose-binding protein	GAL	ca
1hew	1hel	Hen egg white lysozyme	NAG	-
1hyt	1npc	Thermolysin	BZS	DMS, ca, zn
1inc	1esa	Elastase	ICL	ca, so4
1rbp	1brq	Retinol binding protein	RTL	-
1rob	8rat	Ribonuclease A	C2P	-
1stp	1swb	Streptavidin	BTN	-
1ulb	1ula	Purine nucleoside phosphorylase	GUN	so4
2ifb	1ifb	Fatty acid binding protein	PLM	-
3ptb	3ptn	Beta trypsin	BEN	ca
2ypi	1ypi	Triose phosphate isomerase	PGA	-
4dfr	5dfr	Dihydrofolate reductase	MTX	ca, cl
4phv	3phv	HIV 1 protease	VAC	-
5cna	2ctv	Concanavalin A	MMA	ca, cl, mn
7cpa	5cpa	Carboxypeptidase A	FVF	zn
1a6w	1a6u	B1-8 FV fragment	NIP	-
1acj	1qif	Acetylcholinesterase	THA	-
1apu	3app	Penicillopepsin	[IVA-VAL-VAL-STA-OET]	MAN
1blh	1djb	Beta-lactamase	FOS	-
1byb	1bya	Beta amylase	GLC	so4
1hfc	1cge	Fibroblast collagenase	HAP	ca, zn
1ida	1hsi	HIV 2 protease	[QND-VAL-HPB-PPL-PY2]	-
1igj	1a4j	Immunoglobulin	DGX	chain y
1imb	1ime	Inositol monophosphatase	LIP	gd
1ivd	1nna	Hydrolase	ST1	FUC, NAG, MAN, ca
1mrg	1ahc	Alpha momorcharin	ADN	-
1mtw	2tga	Trypsin	DX9	ca
1okm	4ca2	Carbonic anhydrase II	SAB	hg, zn
1pdz	1pdy	Enolase	PGA	ace, mn
1phd	1phc	Camphor 5-monoxygenase	PIM	HEM
1pso	1psn	Pepsin 3a	[IVA-VAL-VAL-STA-ALA-STA]	-
1qpe	3lck	Lck kinase	PP2	PTR, so4
1rne	1bbs	Renin	C60	NAG
1snc	1stn	Staphylococcal nuclease	PTP	ca
1srf	1pts	Streptavidin	MTB	-
2ctc	2ctb	Carboxypeptidase A	LOF	zn
2h4n	2cba	Carbonic anhydrase II	AZM	zn
2pk4	1krn	Plasminogen kringle	ACA	-
2sim	2sil	Sialidase	DAN	-
2tmn	1l3f	Thermolysin	[PHO-LEU-NH2]	ca, zn
3gch	1chg	Gamma chymotrypsin	CIN	-
3mth	6ins	Methylparaben insulin	MPB	cl, zn
5p2p	3p2p	Phosphilipase	DHG	ca
6rsa	7rat	Ribonuclease A	UVC	dod

[1]Considered ligand defining the active site. Brackets indicate ligands composed of multiple residues or fragments.
[2]Additional ligands and ions were removed prior to computations.

Table 2: Comparison of success rates for 48 complexed und 48 unbound protein structures.

	Top 1		Top 3	
	Unbound	**Bound**	**Unbound**	**Bound**
PocketPicker	69	72	85	85
LIGSITE^{cs}	60	69	77	87
LIGSITE	58	69	75	87
CAST	58	67	75	83
PASS	60	63	71	81
SURFNET	52	54	75	78
LIGSITE^{csc}[1]	71	79	-	-

[1]Results of LIGSITE^{csc} are given for the sake of completeness. Note that the comparability to the other methods findings is limited, due to the fact that LIGSITE^{csc} is not a purely geometric approach.

Comparison of pocket shapes

A descriptor was designed to describe the shape of a pocket with respect to the buriedness of the site. Grid probes were grouped into six categories A, B, C, D, E, F holding grid point coordinates with ascending buriedness values: A: 15–16, B: 17–18, C: 19–20, D: 21–22, E: 23–24, F: 25–26. The shape descriptor was developed to record the appearance of distances between pairs of these categories. Distances were staggered in 20 distance bins covering ranges up to 20 Å for 21 possible combinations of the six categories. Pocket shapes were compared with respect to their buriedness by calculating the Euclidean distance *d* between the resulting 420-dimensional shape descriptors of two molecules, *r* and *s*:

$$d = \sqrt{\sum_{i=1}^{420} (r_i - s_i)^2}$$

Results

Evaluation of pocket prediction

To assess the quality of PocketPicker's binding-site predictions we refer to an evaluation method already applied in previous studies [14,17]. Thus, we define a prediction to be a hit, if the geometric center of the presumed pocket lies within 4 Å to any atom of the ligand. Predictions that do not meet this criterion were excluded for calculation of prediction success rates.

The search routine of PocketPicker was evaluated on a test set of 48 protein-ligand complexes and the respective *apo*-structures. Evaluation of pocket predictions for uncomplexed structures is of special interest for geometric search algorithms as the absence of a pocket-inducing ligand might complicate pocket identifications.

Success rates of pocket predictions were compared to the findings of other prediction methods presented in a study published by Huang and Schröder [17]. The test set was compiled as described therein to allow for a comparison of results. Note that slight discrepancies to the original test set cannot be ruled out due to differences in data preparation.

Pocket prediction results were divided into different categories for quality assessment: Correct predictions were termed "TOP1-hits" whereas "TOP3-hits" are predictions where the respective ligand is found within the three largest predicted pockets. Success rates of pocket predictions are summarized in Table 2. Prediction results are given for the proposed methods and their performance on the dataset of 48 bound/unbound structures indicating TOP1- and TOP3-hits.

PocketPicker outperformed CAST, PASS and SURFNET, and showed advantages over LIGSITE and LIGSITE^{cs}. These two programs only showed slightly better success rates for the TOP3-hits on bound protein structures. Results of pocket predictions on the two test data sets are provided for PocketPicker in Table 3, indicating the rank of the proposed binding site and the distance between the pocket center and the nearest ligand atom. The summary of results obtained with LIGSITE^{csc}, LIGSITE, PASS, SURFNET and CAST is available in the work of Huang and Schröder [17].

Analysis of induced fit phenomena

The capability of the proposed shape descriptor to detect conformational similarities in pocket shapes of aldose reductase structures was assessed with respect to the structural analyses presented by Sotriffer and coworkers [29]. Four distinct binding-sites conformations were distinguished by visual inspection, named after the respective ligand characterizing a separate class of pocket shapes: the "IDD594"-conformation, the "holo"-conformation (the cofactor-bound, but ligand-free conformation), the "tolrestat"-conformation, and the "zenarestat"-conformation. In our study, we used these terms to address different classes of structural conformations caused by induced fit phenomena upon ligand binding.

Table 3: Prediction success of PocketPicker on 48 bound and unbound structures.

Complex	Hits[1]	D_{near}/Å)[2]	Unbound	Hits[1]	D_{near}/Å)[2]
1bid	1	2.7	3tms	1	2.3
1cdo	1	2.3	8adh	1	1.9
1dwd	1	0.5	1hxf	1	0.3
1fbp	1	1.3	2fbp	4	1.2
1gca	1	2.4	1gcg	1	1.4
1hew	1	1.4	1hel	1	1.2
1hyt	1	1.3	1npc	1	1.7
1inc	1	0.5	1esa	(1)	4.1
1rbp	1	0.8	1brq	1	1.0
1rob	1	1.7	8rat	1	1.9
1stp	1	2.4	1swb	1	1.0
1ulb	1	1.0	1ula	(1)	4.4
2ifb	1	1.7	1ifb	1	2.5
3ptb	1	0.4	3ptn	3	1.0
2ypi	5	1.0	1ypi	(1)	4.8
4dfr	(1)	7.8	5dfr	1	1.8
4phv	2	2.7	3phv	2	3.5
5cna	-	-	2ctv	-	-
7cpa	1	1.0	5cpa	1	1.1
1a6w	2	1.4	1a6u	3	1.2
1acj	1	0.8	1qif	2	1.2
1apu	1	0.6	3app	1	0.5
1blh	1	1.0	1djb	1	0.8
1byb	1	3.3	1bya	1	3.6
1hfc	1	1.2	1cge	1	1.0
1ida	1	1.5	1hsi	1	3.2
1igj	4	1.6	1a4j	3	1.4
1imb	(1)	5.5	1ime	1	3.4
1ivd	2	1.7	1nna	1	1.5
1mrg	(1)	5.8	1ahc	(1)	5.2
1mtw	2	0.8	2tga	4	0.6
1okm	2	1.2	4ca2	1	1.6
1pdz	1	2.2	1pdy	1	2.7
1phd	1	1.1	1phc	1	0.9
1pso	1	0.4	1psn	1	1.1
1qpe	1	0.9	3lck	1	1.1
1rne	1	1.7	1bbs	1	0.7
1snc	1	2.1	1stn	1	0.3
1srf	1	0.5	1pts	1	0.6
2ctc	1	1.2	2ctb	1	1.5
2h4n	1	0.8	2cba	1	2.1
2pk4	2	0.7	1krn	1	0.7
2sim	2	0.6	2sil	2	0.4
2tmn	1	1.3	113f	1	1.1
3gch	1	0.8	1chg	2	1.5
3mth	2	0.8	6ins	2	1.3
5p2p	1	1.0	3p2p	1	0.8
6rsa	1	3.0	7rat	1	0.9

[1]Rank of pocket centers within 4 Å of the considered ligand (brackets indicate hits exceeding the 4 Å criterion). Only the best hit is shown. Dashes indicate that the actual binding site is not found within the five largest predicted pockets.
[2]Distance from the geometric pocket center to the nearest atom in the ligand.

The complex formed between aldose reductase and the potent inhibitor (IC_{50} = 30 nM) IDD594 (PDB-ID 1us0 [32]) represents its own conformational class in the selected set of structures (Figure 8a). A structural similarity to the zenarestat-conformation (1iei, Figure 8b) was revealed using the calculated shape descriptors. Of all the structures observed in this study, the shape descriptor of the 1iei binding-site showed the smallest Euclidean distance to the IDD594-conformation (Table 4). The binding modes of zenarestat and IDD594 are reported as fairly

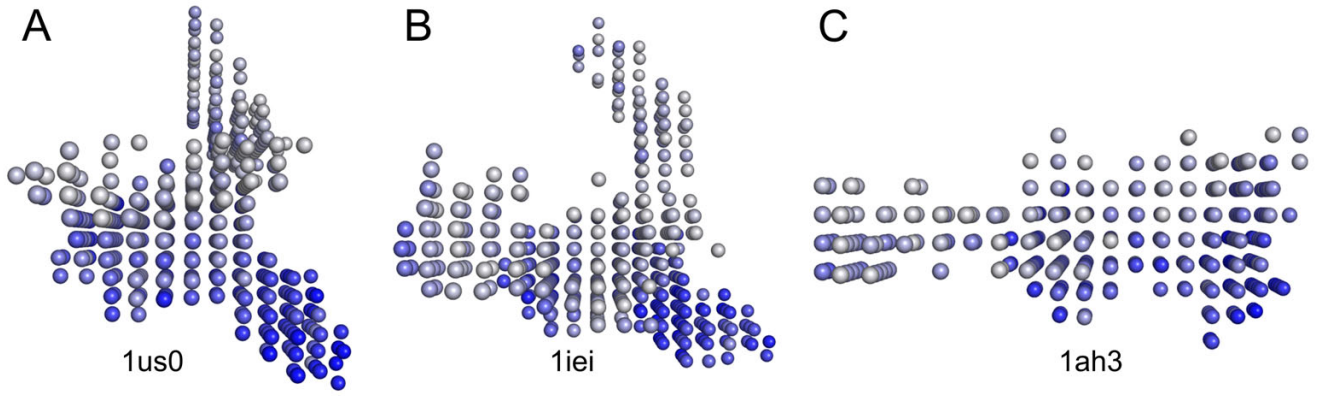


Figure 8
Shapes of pocket conformations induced by IDD594 (**A**), zenarestat (**B**) and tolrestat (**C**). Binding sites are given in PocketPicker representation with darker spheres indicating greater buriedness.

similar [29], which could explain the structural similarities of the binding-site conformations.

The tolrestat-complex (1ah3, Figure 8c) depicts a further binding-site conformation that is substantially different to the other pocket geometries discussed here [29]. This fact is again recognized by our shape descriptor showing pronounced Euclidean distances to the remaining structures (Table 4).

The majority of the binding-site conformations was assigned to the *holo*-conformation (1ads, 1ah0, 1ah4, 1az1, 1eko, 1el3, 2acq, 2acr, 2acs, 2acu) with three structures (1ah0, 1ah4, 1eko) forming a subset with only minor differences to the standard *holo*-conformation [29] (Figure 9).

The conformational similarity of the active sites of this subgroup is reflected in the calculated shape descriptors: Taking 1ah4 as a reference, the remaining members of this subset are correctly identified as the two entries with the lowest Euclidean distance (Table 4). Following this strategy, we were able to identify additional two subsets within the *holo*-conformation set: Considering 1el3 and 2acr as references presenting two strikingly similar entries (Euclidean distance < 2000), we detected the subsets {1ads, 1el3, 2acs} and {2acq, 2acr, 2acu}. Structural similarity in binding-site conformations can be comprehended by the visual information offered by PocketPicker (Figure 10).

PocketPicker was able to correctly predict the active sites of all aldose reductase structures tested, with the exception of the binding-site geometry of 1az1, which shows major differences compared to the *holo*-conformation.

Discussion

The pocket identification algorithm follows the concept of grid-based detection methods. The usage of an increased number of 30 scanning directions provides a finer resolution of the identified binding pockets compared to other implementations. This additional information was used to create a new descriptor combining knowledge of shapes

Table 4: Euclidean distances between pocket shape descriptors. Distances of very similar pocket shapes (*d* < 2000) are highlighted.

	1ah0	1ah3	1ah4	1eko	1el3	1iei	1us0	2acq	2acr	2acs	2acu
1ads	3735	6480	3762	3453	**1527**	3935	4773	2890	4263	2099	4966
1ah0		3749	**1758**	2528	3093	2943	2787	2516	2471	2152	3157
1ah3			3608	4116	5620	3873	3168	4739	3480	4939	3361
1ah4				**1727**	2967	3125	3130	2461	2254	2706	2512
1eko					2452	2972	3046	2693	2910	2884	2752
1el3						3414	4084	2499	3664	**1892**	4033
1iei							2415	3314	3401	2833	3863
1us0								3147	2894	3549	2940
2acq									**1946**	2267	2617
2acr										3058	**1862**
2acs											4019

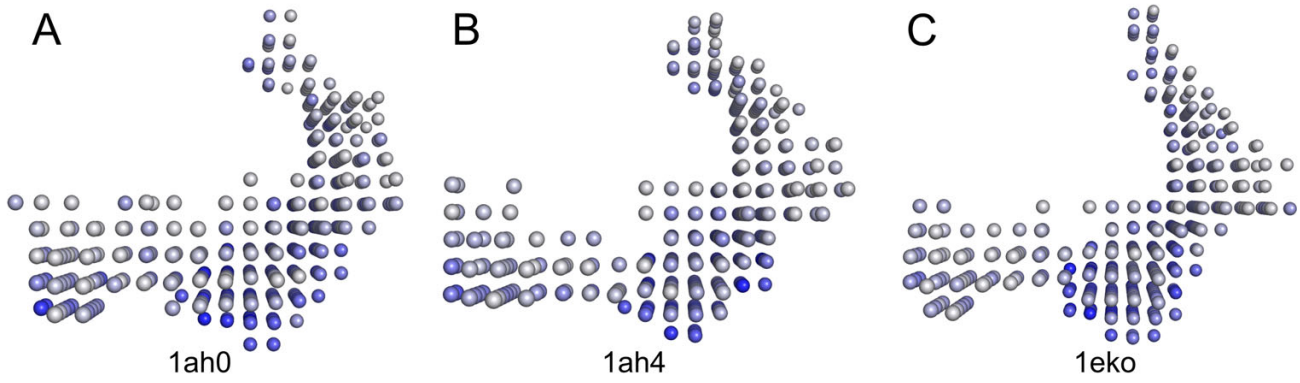


Figure 9
Pocket shapes of the *holo*-conformation subset 1ah0/1ah4/1eko.

with the buriedness of binding-sites. By this means we were able to classify active sites of homologue aldose reductase structures, thereby avoiding the application of sophisticated visual inspections. Results turned out to be promising for shape analyses of closely related enzymes, although 1az1 as the only exception was not properly assigned to the *holo*-conformation class of aldose reductase. This might be due to the fact that this crystal structure carries two mutations within its active site, appreciably changing the shape of the active site.

Pocket analyses revealed a considerable conformational similarity of the active sites of 1eko and 1el3. Although these two proteins originate from different species and share a sequence identity of only 87%, a pronounced adaptation to their common ligand IDD384 could be reg-

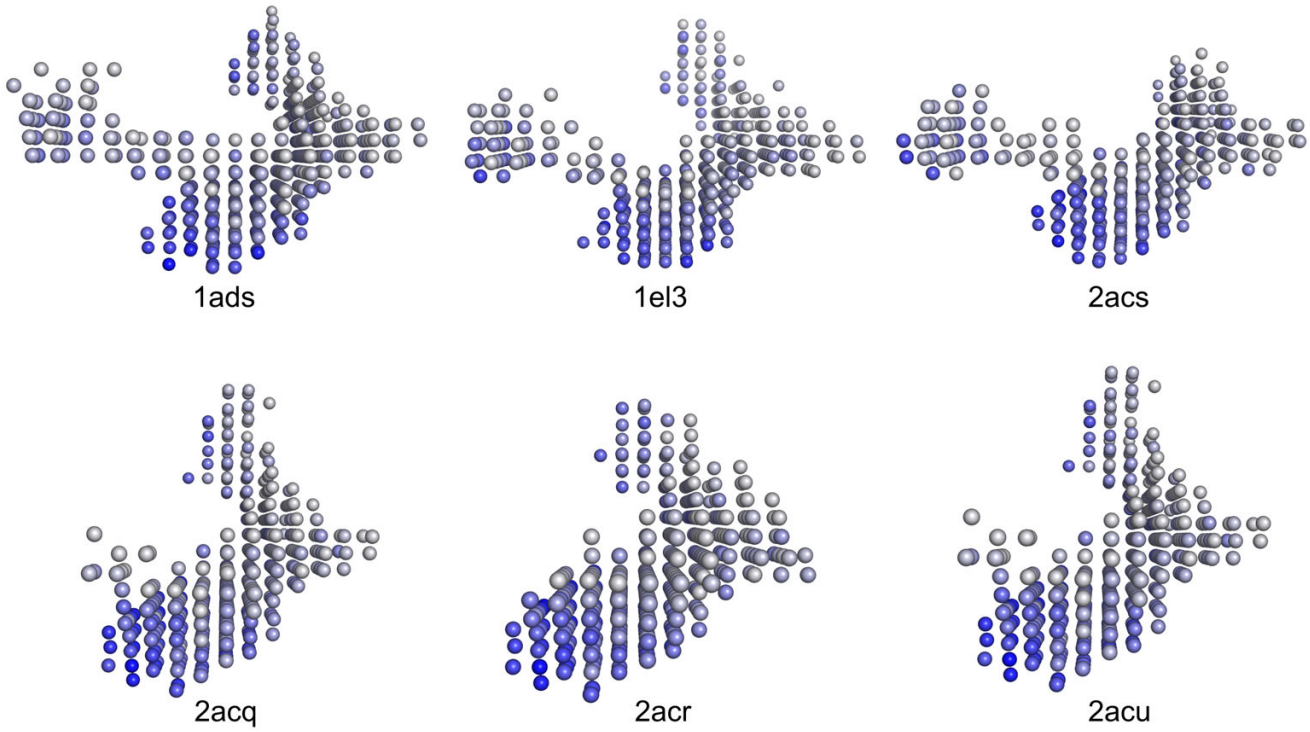


Figure 10
Pocket shapes of the *holo*-conformation with 1ads/1el3/2acs and 2acq/2acr/2acu forming similar subsets.

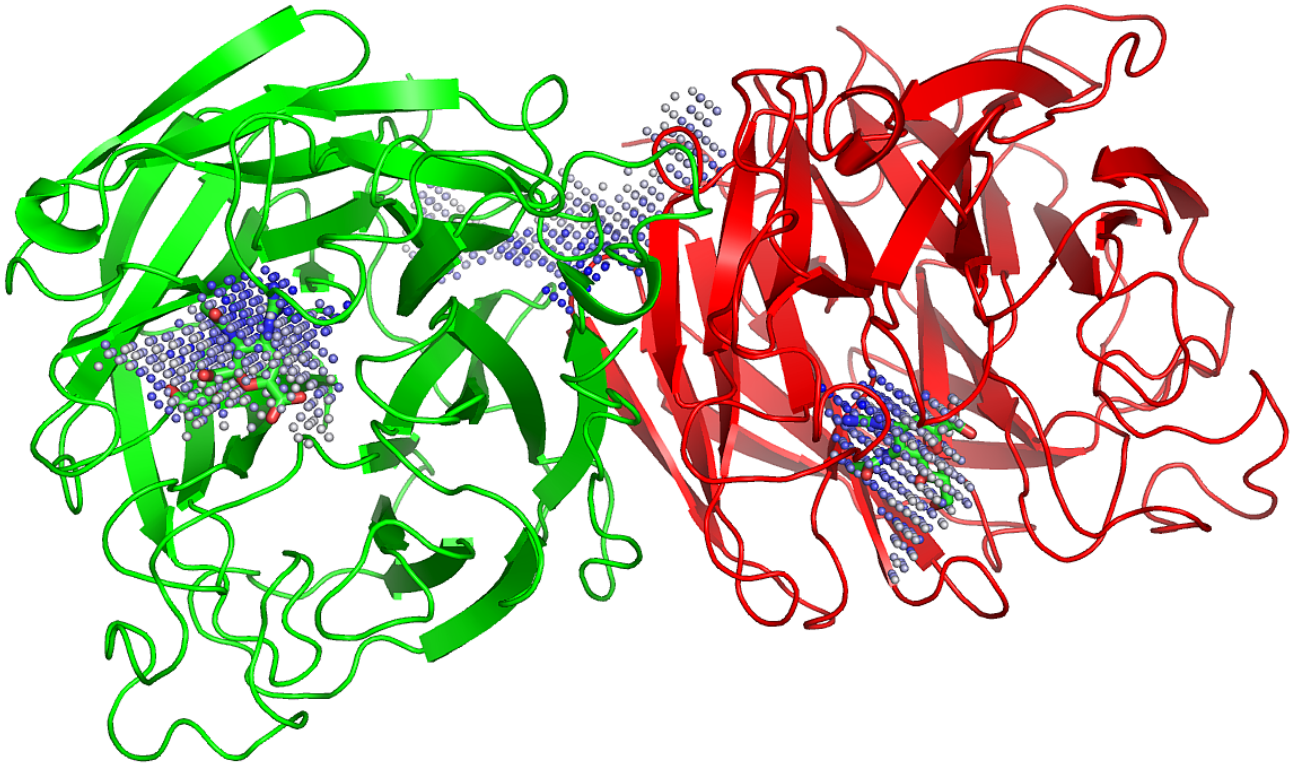


Figure 11
Pocket prediction for influenza virus neuramidase (PDB: 1a4g). A cleft formed between chains A and B is found to be the largest pocket and mistakenly predicted as the actual binding site. The binding sites for the ligands zanamivir (PDB: zmr) are identified as second and third largest pockets.

istered. This circumstance again emphasizes the ability of aldose reductase to react with induced-fit behavior upon ligand binding.

Best results in terms of prediction success rates were observed when applying PocketPicker to comparably small monomeric proteins (< 5000 atoms). Multimeric proteins composed of identical subunits often form clefts between contact faces that can be mistaken as binding sites (Figure 11).

It is therefore recommended to perform binding site predictions on monomeric structures. Predictions for large proteins (> 8000 atoms) turned out to be difficult as disjunct pockets were sometimes joined by narrow "tunnels" underneath the protein surface. The criterion used to assess the quality of pocket predictions raises further problems that can affect the actual prediction success. Thus, Top1-hits may not be considered as correct predictions for small ligands that reside in the distant end of an elongated pocket and, therefore, exceed the maximum distance of 4 Å towards the geometric pocket center (Figure 12).

For the sake of completeness we tested PocketPicker on a test set of 210 complexes compiled by Huang and Schröder [17]. This test set also contained multimeric structures. Success rates of PocketPicker predictions for Top1- and Top3-hits on this test set were considerably lower than on the set of 48 bound/unbound structures (Table 5). The reduced performance of PocketPicker might be due to the fact that this test set includes a considerable number of large proteins. It has been observed that the active site volume scales with the protein size whereas there is little correlation between protein volume and ligand volume [26]. This circumstance complicates predictions made by PocketPicker as the method is designed to identify ligand binding sites of limited size for shape comparisons.

It has been observed by us and others that predicted pockets are often larger than the volume occupied by a ligand [33]. This fact complicates automated shape comparison, because two pockets can possess a similar ligand binding site but have different volumes overall. Future work will be devoted to narrowing the definition of a "pocket" to the actually preferred ligand binding volume. This also

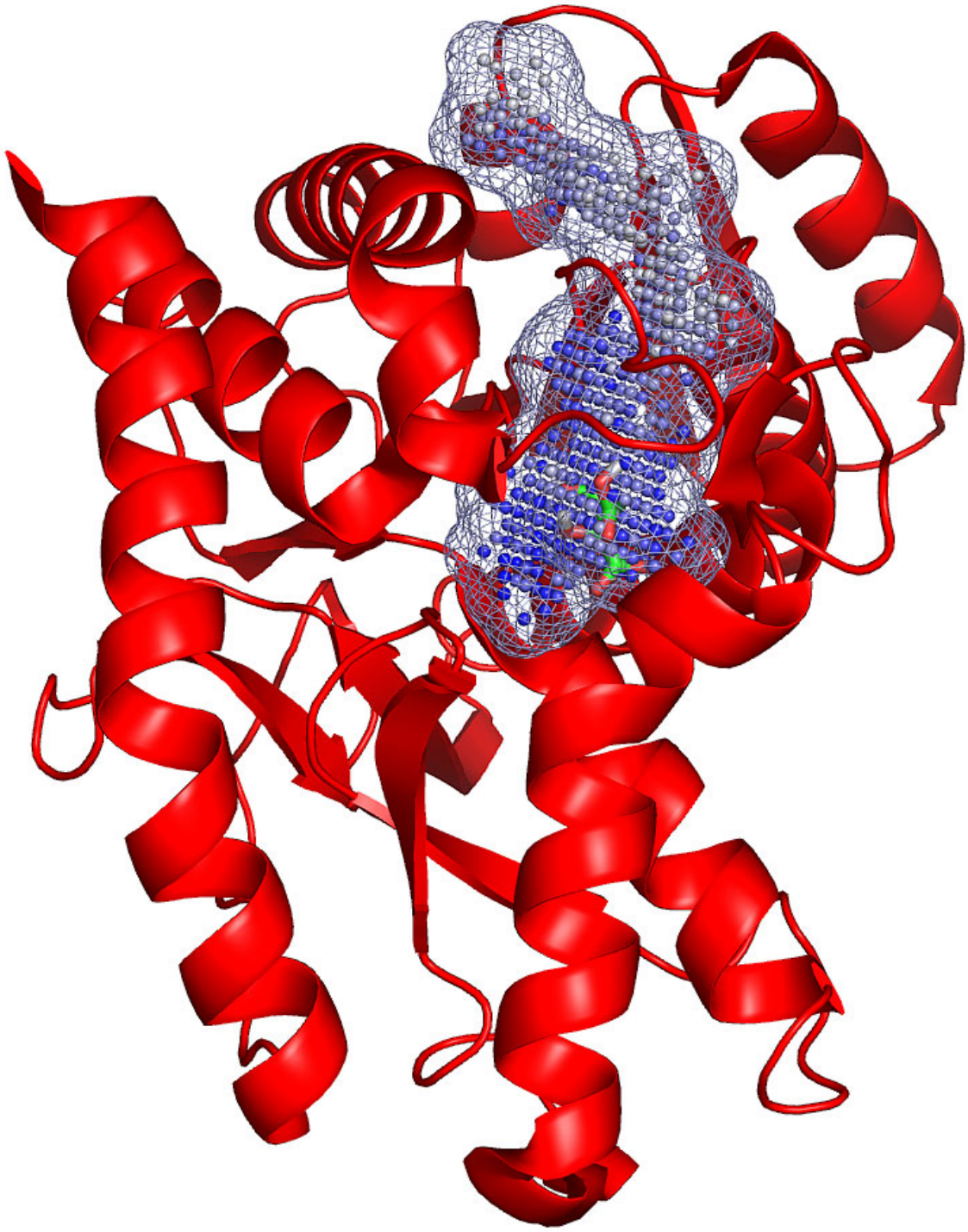


Figure 12
Binding site prediction for malate dehydrogenase (PDB: 2cmd). The ligand citrate (PDB: cit) is situated in the distant end of the elongated pocket (mesh representation) that is suggested as the largest pocket by PocketPicker (blue spheres). Due to the particular shape of the pocket this example is not rated as a correct prediction as the closest ligand atom exceeds the maximal preset distance of 4 Å towards the pocket center.

Table 5: Success rates [%] of PocketPicker on a test set of 210 complexes compared to the results published by Huang and Schröder [17].

	Top 1	Top 3
LIGSITE^{csc}[1]	75	-
LIGSITE^{cs}	67	87
LIGSITE	65	85
PocketPicker	59	71
PASS	54	79
SURFNET	42	56

[1]Results of LIGSITE^{csc} are given for the sake of completeness. Note that the comparability to the other methods findings is limited, due to the fact that LIGSITE^{csc} is not a purely geometric approach.

includes the introduction of an energy-based approach to complement the geometric algorithm used in PocketPicker.

Conclusion

We successfully developed and applied the automated pocket detection method PocketPicker to the task of identifying ligand binding sites in proteins, and the task of clustering structurally related binding sites by shape and a buriedness index. It was demonstrated that the search routine of PocketPicker is capable of identifying the active site within a protein structure with a high success rate on a representative test set.

Availability and requirements

PocketPicker was designed as a plugin for PyMOL [28] (version 0.98). The software is made available *via* our website http://www.modlab.de together with full documentation.

Project name: PocketPicker;

Project home page: http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/index.html;

Operating system: Platform independent;

Programming language: Python, PyMOL;

License: modified BSD; a valid license of PyMOL [28] is required.

Authors' contributions

This work was prepared in the research group of Professor Dr. Gisbert Schneider (Beilstein Endowed Chair For Cheminformatics). The PocketPicker algorithm was developed by Martin Weisel and Ewgenij Proschak. Shape analyses of aldose reductase active sites were carried out by Martin Weisel. This project was based on the idea and realized under the guidance and consultation of Gisbert Schneider.

Additional material

Additional file 1

Four complexed and three unbound structures (all multimers) were converted into their respective monomeric equivalents in the original test set by deleting chains. Additional Table 1 indicates which chains were deleted to form the monomeric proteins. List of structures converted into their monomeric versions by deleting chains from the original PDB file.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-153X-1-7-S1.xls]

Additional file 2

Four complexed and three unbound structures (multimers) were converted into their respective monomeric equivalents in the original test set by deleting chains (see additional Table 1). Results for calculations on the multimeric proteins are presented in additional file 2). Results of PocketPicker calculations on seven multimeric structures contained in the test data set. [1]Rank of pocket-centers within 4 Å of the considered ligand (brackets indicate hits exceeding the 4 Å criterion). Only the best hit is shown. Dashes indicate that the actual binding site is not found within the five largest predicted pockets. [2]Distance from the geometric pocket center to the nearest atom in the ligand. [3]Pocket detection for 1a4j failed, due to PyMOL's inability to align the dimer structure with its corresponding monomeric complex 1igj.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-153X-1-7-S2.xls]

Acknowledgements

The authors thank Michael Schmuker (Johann Wolfgang Goethe Universität, Frankfurt am Main, Germany) for suggestions on triangulations of the sphere. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, and the Deutsche Forschungsgemeinschaft (SFB 579, project A11).

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucl Acids Res* 2000, **28:**235-242.
2. Gane PJ, Dean PM: **Recent advances in structure-based rational drug design.** *Curr Opin Struct Biol* 2000, **10:**401-404.
3. Klebe G: **Recent developments in structure-based rational design.** *J Mol Med* 2000, **78:**269-281.

4. Sotriffer CA, Klebe G: **Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design.** *Farmaco* 2002, **57:**243-251.
5. Campbell SJ, Gold ND, Jackson RM, Westhead DR: **Ligand binding: functional site location, similarity and docking.** *Curr Opin Struct Biol* 2003, **13:**389-395.
6. Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.** *J Mol Graph* 1995, **13:**323-330.
7. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function.** *Protein Sci* 1996, **5:**2438-2452.
8. Liang J, Edelsbrunner H, Woodward C: **Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design.** *Protein Sci* 1998, **7:**1884-1897.
9. Binkowski T, Naghibzadeh S, Liang J: **CASTp: Computed Atlas of Surface Topography of proteins.** *Nucl Acids Res* 2003, **31:**3352-3355.
10. Edelsbrunner H, Mücke E: **Three-dimensional alpha shapes.** *ACM Trans Graph* 1994, **13:**43-72.
11. Edelsbrunner H, Facello M, Fu P, Liang J: **Measuring proteins and voids in proteins.** *Proceedings of the Twenty-Eighth Hawaii International Conference on System Sciences: 3–6 January 1995; Wailea* 1995, **5:**256-264.
12. Aurenhammer F: **Voronoi diagrams – a survey of a fundamental geometric data structure.** *ACM Comput Surv (CSUR)* 1991, **23:**345-405.
13. Lee DT, Schachter BJ: **Two Algorithms for constructing a Delaunay triangulation.** *Int J Comput Inf Sci* 1980, **9:**219-242.
14. Brady GP, Stouten PFW: **Fast prediction and visualization of protein binding pockets with PASS.** *J Comput-Aided Mol Des* 2000, **14:**383-401.
15. Levitt DG, Banaszak LJ: **POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids.** *J Mol Graph* 1992, **10:**229-234.
16. Hendlich M, Rippmann F, Barnickel G: **LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins.** *J Mol Graph Model* 1997, **15:**359-363.
17. Huang B, Schröder M: **LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6:**19-29.
18. Glaser F, Rosenberg Y, Kessel A, Tal P, Ben-Tal N: **The ConSurf-HSSP database: The mapping of evolutionary conservation among homologs onto PDB structures.** *Proteins* 2005, **58:**610-617.
19. Glaser F, Morris R, Najmanovich R, Laskowski R, Thornton J: **A method for localizing ligand binding pockets in protein structures.** *Proteins* 2006, **62:**479-488.
20. Ho CMW, Marshall GR: **Cavity Search: an algorithm for the isolation and display of cavity-like binding regions.** *J Comput-Aided Mol Des* 1990, **4:**337-354.
21. Kleywegt GJ, Jones TA: **Detection, delineation, measurement and display of cavities in macromolecular structures.** *Acta Crystallogr D Biol Crystallogr* 1994, **50:**178-185.
22. Peters KP, Fauck J, Frömmel C: **The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure Using only Geometric Criteria.** *J Mol Biol* 1996, **256:**201-213.
23. Coleman RG, Sharp KA: **Travel depth, a new shape descriptor for macromolecules: application to ligand binding.** *J Mol Biol* 2006, **362:**441-458.
24. An J, Totrov M, Abagyan R: **Comprehensive Identification of "Druggable" Protein Ligand Binding Sites.** *Genome Inform* 2004, **15:**31-41.
25. An J, Totrov M, Abagyan R: **Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes.** *Mol. Cell. Proteomics* 2005, **4:**752-761.
26. Laurie AT, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.** *Bioinformatics* 2005, **21:**1908-1916.
27. Stahl M, Taroni C, Schneider G: **Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network.** *Protein Eng* 2000, **13:**83-88.
28. DeLano WL: **The PyMOL Molecular Graphics System.** *DeLano Scientific* 2002 [http://pymol.sourceforge.net].
29. Sotriffer CA, Krämer O, Klebe G: **Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations.** *Proteins* 2004, **56:**52-66.
30. Saff EB, Kuijlaars ABJ: **Distributing Many Points on a Sphere.** *Math Intelligencer* 1997, **19:**5-14.
31. Thurston WP: **Shapes of polyhedra and triangulations of the sphere.** *Geom Topol Monographs* 1998, **1:**511-549.
32. Podjarny A, Cachau RE, Schneider T, Van Zandt M, Joachimiak A: **Subatomic and atomic crystallographic studies of aldose reductase: implications for inhibitor binding.** *Cell Mol Life Sci* 2004, **61:**763-773.
33. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM: **Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons.** *Bioinformatics* 2005, **21:**2347-2355.

RESEARCH ARTICLE

Form follows function: Shape analysis of protein cavities for receptor-based drug design

Martin Weisel[1], Ewgenij Proschak[1], Jan M. Kriegl[2] and Gisbert Schneider[1]

[1] Johann Wolfgang Goethe-Universität, Institut für Organische Chemie und Chemische Biologie, Frankfurt am Main, Germany
[2] Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riß, Germany

Received: January 30, 2008
Revised: June 23, 2008
Accepted: August 14, 2008

Identification of potential ligand-binding pockets is an initial step in receptor-based drug design. While many geometric or energy-based binding-site prediction methods characterize the size and shape of protein cavities, few of them offer an estimate of the pocket's ability to bind small drug-like molecules. Here, we present a shape-based technique to examine binding-site druggability from the crystal structure of a given protein target. The method includes the PocketPicker algorithm to determine putative binding-site volumes for ligand-interaction. Pocket shape descriptors were calculated for both known ligand binding sites and empty pockets and were then subjected to self-organizing map clustering. Descriptors were calculated for structures derived from a database of representative drug-protein complexes with experimentally determined binding affinities to characterize the "druggable pocketome". The new method provides a means for selecting drug targets and potential ligand-binding pockets based on structural considerations and addresses orphan binding sites.

Keywords:
Binding-site / Drug discovery / Orphan target / Pocketome / Protein structure / Virtual screening

1 Introduction

The ability to predict certain proteins suited for interaction with specific drug-like molecules is of particular interest in structure-based drug design and development. Identification of "druggable" protein-binding pockets and allosteric sites is considered a useful starting point when working on a potential target. In the context of this study, we use the term "druggable" to denote protein-surface pockets that can accommodate drug-like ligands. Therapeutically relevant targets have thus to be both disease modifying and druggable. It has been estimated that the effective number of such targets that can be exploited by the pharmaceutical industry comprises only several hundreds of potential targets [1]. It is therefore important to carefully distinguish non-druggable protein cavities from pockets suited for protein-ligand interaction at an early stage of the drug development process.

Different approaches have been used to examine the druggability of unknown potential binding sites. Common methods consider pocket size, surface roughness or polar/apolar surface area as key descriptors for druggability analyses. Previous findings describe the fact that the endogenous binding site is usually the largest, most hydrophobic and geometrically most complex pocket of a protein [2, 3]. While none of these parameters alone is sufficient to predict binding site druggability, regression analyses are commonly used

Correspondence: Professor Gisbert Schneider, Johann Wolfgang Goethe-Universität, Beilstein Endowed Chair for Cheminformatics, Institut für Organische Chemie und Chemische Biologie, Siesmayerstr. 70, D-60323 Frankfurt am Main, Germany
E-mail: G.Schneider@chemie.uni-frankfurt.de
Fax: +49-69-798-24880

Abbreviation: SOM, self-organizing map

to determine the importance of specific descriptors for druggability predictions.

In this study, we present the applicability of an automated pocket-prediction method providing autocorrelation vector-based shape descriptors of potential binding sites for druggability analyses. Prevalent pocket prediction methods solely rely on geometric aspects for identification of protein surface depressions, assuming the largest cleft to be the actual ligand-binding site. This assumption is supported by empirical studies emphasizing pocket volume as a prominent characteristic for binding-site identification [4, 5]. Geometric pocket detection algorithms cover a variety of techniques for the identification of protein surface cavities: SURFNET [6, 7] uses fitting of virtual spheres into the solvent-accessible space between protein atoms, which are scaled down when penetrating van-der-Waals radii of neighboring atoms. The program CAST [8, 9] makes use of the *alpha*-shapes approach [10, 11] comprising the concepts of Voronoi tesselation [12] and Delaunay triangulation [13] to define protein-binding sites. PASS [14] (putative active sites with spheres) coats the protein surface with a layer of virtual spheres and exposes spheres residing in buried parts of the protein. Additional layers are added and a central sphere is specified for each identified pocket to represent the respective potential site.

Geometric pocket prediction methods typically use artificial grids constructed around a regarded protein structure with probes installed in the grid to examine their respective molecular environment along predefined search vectors. POCKET [15] searches for potential pockets by scanning along the *x*-, *y*-, and *z*-axes to locate buried parts of the protein surface. LIGSITE [16] was introduced as an advancement to pocket providing a more precise scanning procedure by using an increased number of 14 search vectors. LIGSITE^{cs} [17] (cs = Conolly Surface) and LIGSITE^{csc} [17] (CSC = Conolly Surface and Conservation) further enhanced results by considering the Conolly surface and re-ranking of favored predicted pockets by the degree of conservation of the closest surface residues. LIGSITE^{csc} uses conservation scores obtained from the ConSurf-HSSP database [18] and is therefore not considered as a purely geometric approach. An improved method of SURFNET has been introduced, using conservation scores for re-ranking to further increase prediction accuracy [19]. Additional geometric pocket prediction methods are Cavity Search [20], VOIDOO [21], APROPOS [22] and Travel Depth [23]. DrugSite [24] and PocketFinder [25] evaluate shape and physicochemical properties for pocket prediction. Q-SiteFinder uses energetic calculations for the identification of ligand-binding sites [26].

In this work, our own geometric prediction method PocketPicker [27] was used for pocket prediction. PocketPicker provides a method to translate the shape and buriedness of potential binding sites into autocorrelation vectors, which allow for alignment-free pocket comparisons. Descriptors based on autocorrelation vectors are widely used in cheminformatics as this vector representation enables efficient quantification of molecular similarities [28, 29]. Here, we employ this approach to examine protein druggability, by applying PocketPicker shape descriptors of pockets known to bind small molecules with high-affinity.

2 Materials and methods

2.1 Protein data collection

Druggability analyses were performed on two sets of receptor-ligand complexes with experimentally determined binding affinities designed to characterize a representative profile of the pocket universe while preserving sequential diversity of the underlying protein structures. Both protein sets were derived from the *refined set* of the PDBbind database [30, 31], a collection of 1300 complexes meeting a variety of predefined quality criteria, such as an overall crystal structure resolution $\leq$2.5 Å as well as demands on ligand structures and the fitting accuracy of their noncovalent interactions.

Set A was derived from the refined set in a two-step procedure: (i) Monomeric structures and structures that could be converted into the respective monomers were selected yielding a subset of 909 complexes. Subsequent pocket analyses were focused on protein monomers to facilitate binding-site identification with PocketPicker [27]. This is because multimeric proteins tend to form clefts between contact faces that can be mistaken as binding sites by geometric pocket prediction methods. (ii) The set was then reduced to complexes with predicted ligand binding site volumes or pockets of additional ligands not exceeding the ligand volume for more than 50% resulting in Set A consisting of 623 structures.

Set B (98 complexes) is composed of set A members that are annotated in the PDBbind core set [30, 31], which is a subset deduced from the refined set holding 210 unrelated entries. Set B was introduced to ensure that clustering of druggable pocket descriptors is not driven by trivial clustering of homologous proteins.

2.2 Ligand collection

Reference ligands from PDBbind as well as additional ligands contained in the complexes of Sets A and B were added to the ligand collection and termed "main ligands" and "additional ligands", respectively. Ligands that did not fulfill the binding-site volume constraint (see Section 2.1) were removed from the set. A simple filter was applied to avoid misinterpretations of pockets binding artifacts (such as ions and solvents) as druggable. We therefore referred to the Collection of Bioactive Reference Analogues (COBRA) database [32] to specify the minimum size of annotated drug-like molecules. Analyses on 8311 pharmacologically active substances showed that all ligands within 2 SD from the mean were composed of at least nine heavy atoms. Ligands with less than nine heavy atoms were excluded from the ligand

collection as well as sugars due to their ubiquitous binding behavior. The final ligand collection consisted of 623 main ligands and 19 additional ligands for Set A, and 98 main ligands and 7 additional ligands for Set B.

2.3 Automated pocket prediction and shape descriptors

PocketPicker was used for the identification of protein cavities. The method places probes into an artificial grid covering an underlying protein structure and uses a sophisticated geometric scanning procedure for the identification of buried regions on the protein surface (see reference [27] for details). Neighboring grid probes residing in protein cavities are afterwards grouped into clusters depicting potential binding pockets (Fig. 1). PocketPicker comprises the calculation of a shape descriptor that describes the shape of a pocket with respect the distribution of buriedness over the site. A buriedness index is computed for every surface-near grid probe to determine its accessibility on the protein surface (Figs. 2a and b). Buriedness indices are natural numbers ranging from 16 to 26 indicating growing buriedness of a probe in the cavity.

A 210-dimensional shape descriptor was used in this work: Grid probes were grouped into six categories A–F holding grid point coordinates with ascending buriedness values (A: 15-16, B: 17-18, C: 19-20, D: 21-22, E: 23-24, F: 25-26). The shape descriptor records the appearances of distances between pairs of these categories staggered in ten distance bins covering ranges up to 10 Å. Considering the 21 possible combinations of the six categories A–F, the shape descriptor is composed of 210 integers (Fig. 2c). In a previous study [27], the concept of correlation-based shape descriptors was successfully applied to examine similarity of pocket shapes using Euclidean distance measurement, and the prediction performance of PocketPicker was evaluated on a representative set of protein-ligand complexes derived from the PDBBind [30, 31]. PocketPicker succeeded in locating the actual binding site as one of the top-three predicted sites in 85% of the proteins tested, both for the complexed and for the unbound set of protein structures. Prediction success rates outperformed the results of other established tools that had been applied to the same dataset [27].

Figure 1. Visualization of PocketPicker cavity detection routine (Dihydroxybiphenyl-Dioxygenase (DHBD), PDB entry: 1lgt). Pockets are depicted as clusters of grid probes defining the site. Pocket buriedness is expressed using color shading with darker shading for more buried parts of the pocket.

2.4 Clustering

Self-organizing maps (SOM, [33]) were used for classification of pocket shape descriptors for druggability analyses. The software molmap® was used to project data onto a SOM containing 10×15 neurons with toroidal topology as described previously [34, 35] (Gaussian neighborhood function with initial radius = 7 and linear adaptation; 400 000 training epochs). The Euclidean distance measure was used for data clustering.

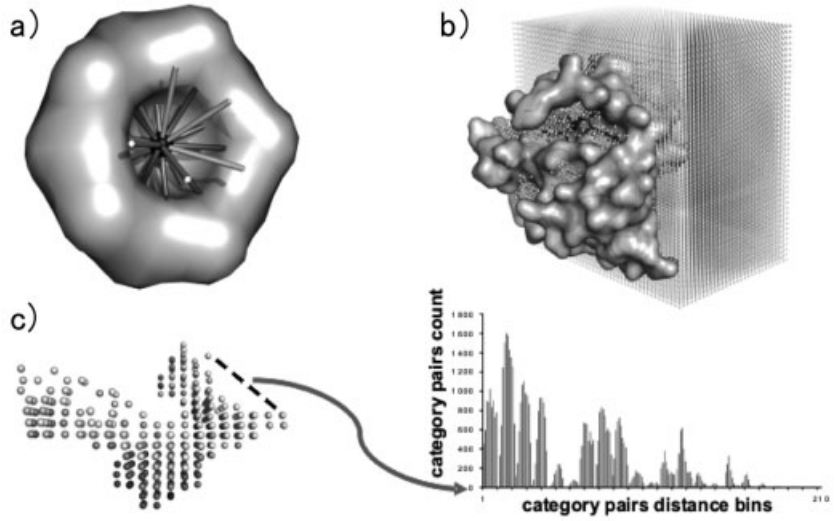


Figure 2. Schematic overview of buriedness calculation (a): A grid probe (center) examines its molecular environment along 30 direction vectors. A buriedness index of 25 is calculated here as 25 vectors interfere with the schematic receptor model. Buriedness indices are calculated for surface-near probes of the artificial grid (b). Grid probes are clustered into potential pockets and a topological correlation vector is derived for each pocket (c). The first ten bins hold occurrences of A-A category pairs with distances from 1 to 10 Å.

3 Results

PocketPicker extracted 13 859 potential binding pockets from 623 protein structures (Set A). Set B yielded 2257 cavities. In contrast to Set A, Set B contains only 98 complexes, which represent a diverse and unrelated selection of representative ligand-binding pockets. For Set A, PocketPicker yielded correct prediction for 477 (77%) of the 623 tested complexes. For 558 entries (90%) the actual binding site was identified as one of the top-three predicted cavities. These results were obtained using only a shape description of protein pockets.

Pocket size was expressed as the number of grid probes that define a surface depression. Regarding the mesh size of 1 Å used by PocketPicker in the present study, one grid probe within the cubic grid defines a volume of 1 $Å^3$. Cavities composed of less than ten such probes were excluded from pocket analyses. Pocket buriedness was described using PocketPicker buriedness indices, which account for grid probes that are surrounded by protein atoms from about 50% to deeply buried probes. Average pocket size and buriedness of the sets of extracted pockets are given in Table 1.

It is generally assumed that protein cavities suitable to bind small drug-like molecules show an increased pocket size and buriedness when compared to empty sites [27]. This holds true for both Sets A and B (Fig. 3): Pockets occupied by main or additional ligands show notably larger values for pocket size and buriedness. However, the two largest pockets from Sets A and B turned out to be cavities that are unoccupied, with volumes of 2121 $Å^3$ (carbamoyl phosphate synthetase, PDB entry: 1bxr [36]) and 2010 $Å^3$ (hepatitis C virus RNA polymerase, PDB entry: 1nhu [37]). Maximal pocket buriedness of 25.53 was observed for an unliganded site contained in both Sets A and B (beta secretase, PDB entry: 1fkn [38]).

Identification of protein cavities suitable to accommodate small drug-like molecules can be seen as a first step in receptor-based drug design. Aside from appropriate pocket volume, a certain degree of buriedness is considered a necessary precondition to enable the deployment of noncovalent interaction between the ligand and the receptor. Therefore, researchers working in the field of lead structure identification generally look for potential binding sites with reasonable buriedness *vs.* size performance. A detailed analysis of pocket size and pocket buriedness for druggable pockets is given in Fig. 4. It is evident that known liganded pockets are not uniformly distributed but rather possess preferred size and buriedness values.

Then, we used molmap® [34] to train SOM on PocketPicker shape descriptors of liganded and unliganded pockets to further examine the role of pocket size and buriedness in druggability (Fig. 5). The SOM achieved clearly recognizable clustering of liganded pockets for both sets of complexes. This demonstrates the usefulness of our shape descriptor for the distinction between liganded and unliganded cavities. Notably, a comparably small group of pocket descriptors (encoded as SOM "neurons") represents liganded pockets that are surrounded by a larger set of descriptors representing unliganded pockets. To obtain a quality estimation of the classification accuracy, Matthews' correlation coefficient, *cc*, Eq. (1), was calculated based on the SOM clustering (Figs. 5b and d).

$$cc = \frac{P * N - O * U}{\sqrt{(P + O)(P + U)(N + O)(N + U)}} \quad (1)$$

Liganded pockets were considered positive correct (*P*) when mapped onto black neurons, and underpredicted (*U*) when mapped onto white neurons. Unliganded pockets were interpreted as negative correct (*N*) when projected onto white parts of the binary SOM, and overpredicted (*O*) when assigned to black neurons. The SOM derived from Set A and Set B yielded Matthews correlation coefficients of 0.72 and 0.76, respectively (see Supporting Information Table 1), thereby confirming the suitability of the proposed method for a distinction of liganded and unliganded pockets.

4 Discussion

The concept of correlating information about pocket size and the distribution of pocket buriedness in a topological shape descriptor turned out to be a valuable tool for rapid alignment-free pocket shape comparison. In a previous study, PocketPicker shape descriptors have already proven to be suitable for detecting conformational similarities among pocket shapes of aldose reductase [27]. In this work, we

Table 1. Mean pocket size and pocket buriedness for pockets predicted for Sets A and B. Pocket sizes are expressed as the number of probes defining the site, PocketPicker buriedness indices are given as results from buriedness analyses

	Set A		Set B	
	Mean pocket size	Mean pocket buriedness	Mean pocket size	Mean pocket buriedness
Main ligand pockets	322.0	19.4	349.8	19.2
Additional ligand pockets	229.4	18.9	223.6	18.8
Nonliganded pockets	59.5	17.4	62.6	17.3
Entire dataset	71.5	17.5	75.6	17.4

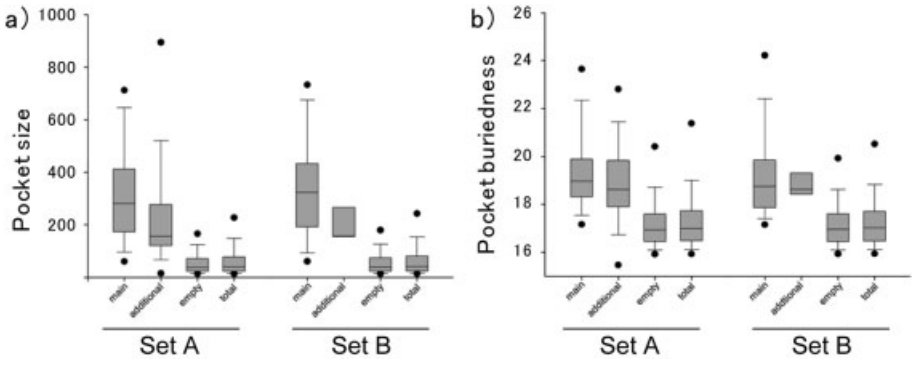


Figure 3. Box plot representations of pocket size (a) and pocket buriedness (b) of pockets extracted from the protein complexes contained in Set A and Set B. Results are shown for pockets with reference ligands (*main*), pockets occupied by ligands other than reference (*additional*), unliganded pockets (*empty*) as well as for the entire set (*total*). Black circles indicate 5th and 95th percentile.

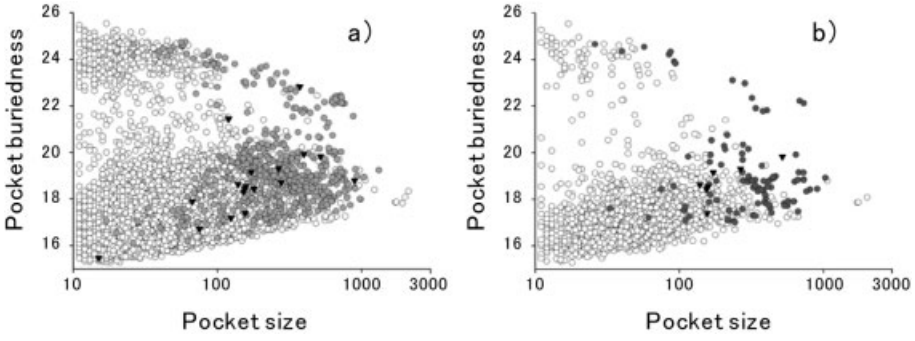


Figure 4. Size *vs.* buriedness of pockets extracted from Set A (a) and Set B (b). Distributions for unliganded pockets (white circles), additional ligands (black triangles), and main ligands (grey circles) are shown.

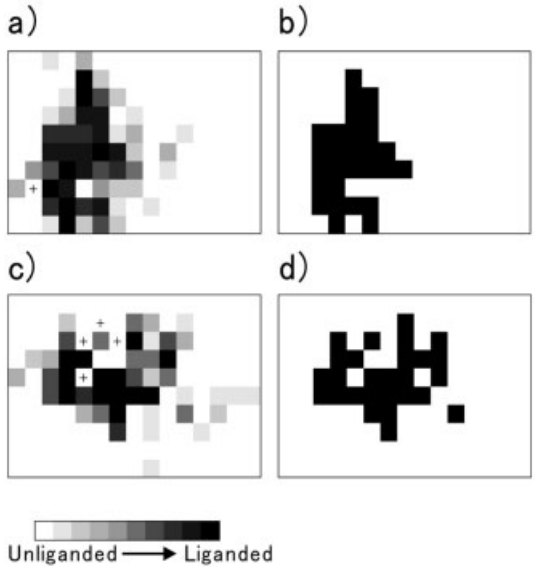


Figure 5. Results for shape descriptors of test pockets extracted from the protein complexes contained in Set A (a, b) and Set B (c, d) on 10 × 15 toroidal SOM projections. Relative frequencies are given in (a) and (c) showing black squares for neurons solely populated by liganded pockets while white squares indicate neurons holding unliganded pockets only. Neurons holding mixtures of liganded and unliganded pockets are shown as grey shaded squares. Empty neurons are labeled by '+'. Binary classification is given in (b) and (d) showing black or white squares for neurons mainly populated by liganded or unliganded pockets, respectively.

examined the applicability of PocketPicker shape descriptors for predicting general druggability of potential ligand-binding sites. We used a set of high-quality complexes taken from the PDBbind database [30, 31] to compile sets of druggable pockets with experimentally determined binding affinities. Confirming previous research, we demonstrate that the largest protein cavity is most likely the actual ligand binding-site: Cavities occupied by main or additional ligands are larger than empty sites (Fig. 3, Table 1). This holds true for both Sets A and B. The largest pockets calculated for Set A and Set B were not liganded, which underlines the fact that considerably large pocket size does not imply druggability. Adjoining shallow pockets can be mistaken as a large single cavity by pocket clustering algorithms, which is a drawback prevalent throughout geometric pocket prediction methods. Still, PocketPicker prediction success rates for Set A slightly exceed those of previous findings [27].

We wish to point out that our concept of target "druggability" primarily addresses the ability of a protein pocket to accommodate a drug-like ligand; we do not explicitly consider drug-like ligand properties as such. To get an idea of the drug-likeness of the ligands in Sets A and B, additional effort has been made to ensure that the selected ligands represent "drug-like" molecules. Lipinski descriptors were calculated using the MOE package (Chemical Computing Group, Montreal, Canada, http://www.chemcomp.com). Results showed that 80% of the ligands in Set A (79% for Set B) were considered "drug-like" in terms of Lipinski's Rule-of-Five (Ro5) [39]. As the Ro5 guidelines address potential oral bio-

availability of a substance, we additionally applied a second drug-likeness estimation, which is based on an artificial neural network following the concept of Sadowski and Kubinyi [40] and Ajay *et al.* [41]. Our own implementation of this drug-likeness concept was used [42]. This neural network prediction system computes a score between zero (non-drug-like) and one (drug-like) for a given molecular structure. Predictions show an approximately balanced distribution of the ligands (Fig. 6), with at least 50% of the data yielding a score >0.5. This result allowed us to distinguish between pronounced "drug-like" (score >0.9) and "non-drug-like" (score <0.1).

Then, we analyzed the distribution of pocket shape descriptors on the trained SOM with respect to the calculated drug-likeness of the corresponding ligands. We therefore derived subsets of Set B containing only pockets with ligand drug-likeness score > 0.9 (n = 28) according to neural network analysis, or drug-likeness score <0.1 (n = 25), respectively. These subsets were considered as representative drug/non-drug data and were re-projected on the original SOM trained before with descriptors of Set A and B (Fig. 7). Shape descriptors of pockets containing ligands with a high drug-likeness score clearly populate the center of the "druggability islands" (Fig. 5), while pockets with non-drug-like ligands reside in the outer margin of these clusters. This observation can be used as an additional criterion for druggability prediction: Apparently, pockets with shape descriptors related to neurons in cluster centers have an increased potential to interact with drug-like molecules (note that the descriptor selections are subsets of Set A and B and therefore only able to populate the same neurons as highlighted in the original SOM, *cf.* Fig. 5). It is remarkable that a distinction of pockets with pronounced different drug-likeness scores is noticeable within the given SOM islands. The overlay of island centers and their respective boundaries sum up to trace the outline of the original SOM islands obtained with the full Set A and B (Fig. 5). It should be mentioned here, that a ligand with a predicted low drug-likeness score does not necessarily have to be a "non-drug" as all ligands in the test data are annotated in the PDBBind database with an experimentally determined affinity to its respective receptor.

The SOM "druggability islands" derived from Sets A and B have a similar size, but exhibit differences in their shape

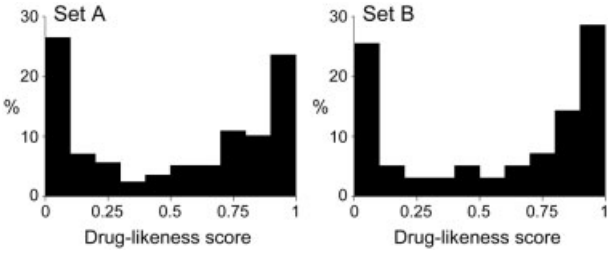


Figure 6. Calculated drug-likeness scores (artificial neural network) for the ligands from Set A and Set B.

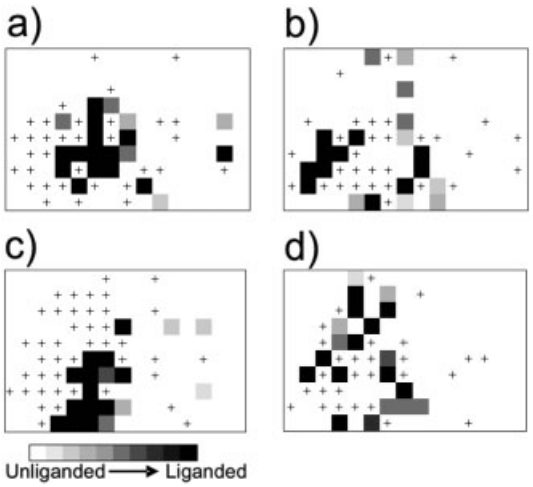


Figure 7. Projection of Set B drug/non-drug pocket descriptors onto the SOM from Fig. 5 (Set B, upper row), Set A, lower row)). Pockets with drug-like ligands (score >0.9) cluster in the center of the suggested druggability island (a), while pockets with non-drug-like ligands (score <0.1) populate the boundary (b). This also applies when projecting the descriptors on the SOM trained with Set A pockets (c, d).

(*e.g.* island shapes in Figs. 5a and c). This is because these SOM were trained on different datasets providing different information contents for pocket classification. Furthermore, competitive learning (which is the key principle of SOM) is a non-deterministic method, which easily leads to different SOM plots. Keeping these caveats in mind, our results clearly show that the SOM is actually able to visualize islands, which are densely populated druggable pockets. In fact, the SOM projections show an enrichment of a few hundred druggable pockets (dark-colored clusters) situated in an ocean of thousands of non-liganded pockets (white-colored clusters). This finding is again supported by assessing the discriminative quality of the SOM [Eq. (1)]. Results show correlation coefficients of cc = 0.88 and cc = 0.89 for mapping of drug-like pockets onto SOM trained with Set A and B respectively (Supporting Information Table 2). Mappings of non-drug-like pockets yield correlation coefficients of 0.88 and 0.79 for Set A and Set B trained SOM (Supporting Information Table 2; Matthews coefficients are calculated from binary SOM of Fig. 7 given in Supporting Information Fig. 1). These results affirm the aptitude of PocketPicker shape descriptors for pocket druggability predictions.

In a previous study, PocketPicker predicted binding-sites of *apo*-structures with high accuracy [27]. The datasets used in the present work for SOM-based druggability predictions consist of protein-ligand complexes, where liganded pockets were used to characterize druggable binding-sites. Generally, it would be useful to predict pocket druggability of *apo*-structures. To address this issue, shape descriptors were calculated for seven PDB entries representing *apo*-structures for seven complexes from Set A. *Apo*-structures were aligned to

Table 2. Coordinates of selected complexed structures (*holo*) and their corresponding *apo*-structures mapped on the SOM trained with Set A and Set B (*cf.* Fig. 5). Coordinates refer to the trained SOM, where (0,0) indicates the upper left neuron (first row, first column)

PDB-ID *holo*/*apo*	(x,y) Position on Set A SOM (*holo*)/(*apo*)	(x,y) Position on Set B SOM (*holo*)/(*apo*)
1rbp/1brq	(3,5)/(3,5)	(3,2)/(3,5)
7cpa/5cpa	(4,4)/(5,4)	(4,4)/(5,4)
2ctc/2ctb	(4,3)/(5,4)	(3,4)/(5,4)
2h4n/2cba	(5,4)/(5,3)	(6,6)/(5,3)
2sim/2sil	(5,5)/(5,8)	(5,4)/(5,8)
2tmn/ll3f	(7,6)/(7,6)	(8,4)/(7,6)

their respective complexed versions using the *align* function implemented in PyMOL (version 1.0r2, [43]) to locate the pocket corresponding to the actual ligand-binding site in the complex. Shape descriptors of the selected *apo*-pockets were then projected on the SOM shown in Fig. 5. Results demonstrate that most of the *apo*-pocket descriptors populate either the same or the neighboring neurons as their corresponding descriptors from the ligand-receptor complexes (Table 2). This holds true for the trained SOM from both Set A and Set B and suggests that the proposed method might also be applicable to unbound (*apo*-) protein structures.

As an additional descriptor to assist druggability predictions, we considered information about pocket buriedness. Results show that druggable pockets are clearly shifted towards larger and more buried sites (Figs. 3 and 4). This fact is also reflected for additional ligands contained in the complexes investigated. Analyses of the pocket size *vs.* buriedness performance revealed two distinct populations concerning pocket buriedness. Sets A and B are separated into two clusters: deeply buried (buriedness >22.5) and more shallow (buriedness <20) pockets. The area in between is only sparsely populated, while most pockets residing in this area are denoted as druggable. This observation could result from the compactness of protein structures [44] which renders the occurrence of tunnels ($20<$ buriedness <22.5) below the surface rather unlikely. Based on our analysis, however, existing tunnels most likely provide function. This coincides with the observation that large protein cavities are likely to be ligand-binding sites. Future research will concentrate on pockets situated in this area to propose them as candidates for receptor-based drug design.

A peculiarity in Fig. 4 is the distribution of ligand-binding sites with an average buriedness value greater than 22.5. Among these 55 pockets, four functions are prevalent: (total numbers of pockets in brackets): sugar binding/transport (15), amino acid binding (12), hormone/growth factor receptors (8), and hydrolases/lysozymes (8). It has been reported that sugars usually bind to shallow grooves formed by protein loop regions with low affinity, whereas sugar transporters bind saccharides in buried clefts with high affinity [45].

Next steps might include incorporation of machine learning methods like artificial neural networks and support vector machines for automated feature extraction and as an alternative for SOM clustering. Generally, an advantage of autocorrelation descriptors that were used for SOM clustering in this work is their usefulness for alignment-free comparisons of two or more pocket shapes. Autocorrelation vectors avoid the task of explicit spatial alignment of binding pockets. Therefore, this approach can offer advantages in terms of computational speed. Despite their appeal for rapid similarity calculation, however, autocorrelation vectors have drawbacks. For example, information about stereochemistry or mirror images of binding pockets is lost, which can result in erroneous similarity assignments. Future work will have to address this issue. Inclusion of additional alignment-free geometric descriptors might be useful for this purpose [46, 47].

In this work, SOM were used with the intent to classify druggable and non-druggable pockets. The resulting maps show that druggable pockets form a distinct cluster in descriptor space. An additional reduced set of sequentially independent protein complexes was used to ensure that clustering of druggable pockets was not driven by homologous similarities. These maps are valuable tools for projection of unknown pockets to examine their druggability and provide a systematic approach for the "de-orphanization" of protein cavities [48].

This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, and the Deutsche Forschungsgemeinschaft DFG (SFB 579, project A11.2). The authors thank Boehringer Ingelheim Pharma GmbH & Co. KGaA for funding this project.

The authors have declared no conflict of interest.

5 References

[1] Hopkins, A. L., Groom, C. R., The druggable genome. *Nat. Rev. Drug Discov.* 2002, *1*, 727–730.

[2] Hajduk, P. J., Huth, J. R., Tse, C., Predicting protein druggability. *Drug Discov. Today.* 2005, *10*, 1675–1682.

[3] Hajduk, P. J., Huth, J. R., Fesik, S. W., Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* 2005, *48*, 2518–2525.

[4] Sotriffer, C. A., Klebe, G., Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* 2002, *57*, 243–251.

[5] Campbell, S. J., Gold, N. D., Jackson, R. M., Westhead, D. R., Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* 2003, *13*, 389–395.

[6] Laskowski, R. A., SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* 1995, *13*, 323–330.

[7] Laskowski, R. A., Luscombe, N. M., Swindells, M. B., Thornton, J. M., Protein clefts in molecular recognition and function. *Protein Sci.* 1996, *5*, 2438–2452.

[8] Liang, J., Edelsbrunner, H., Woodward, C., Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* 1998, *7*, 1884–1897.

[9] Binkowski, T., Naghibzadeh, S., Liang, J., CASTp: Computed atlas of surface topography of proteins. *Nucl. Acids Res.* 2003, *31*, 3352–3355.

[10] Edelsbrunner, H., Mücke, E., Three-dimensional alpha shapes. *ACM Trans. Graph.* 1994, *13*, 43–72.

[11] Edelsbrunner, H., Facello, M., Fu, P., Liang, J., Measuring proteins and voids in proteins. *Proc. 28th Ann. Hawaii Internat. Conf. System Sciences* 1995, *5*, 256–264.

[12] Aurenhammer, F., Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Comput. Surv. (CSUR)* 1991, *23*, 345–405.

[13] Lee, D. T., Schachter, B. J., Two algorithms for constructing a Delaunay triangulation. *Int. J. Comput. Inf. Sci.* 1980, *9*, 219–242.

[14] Brady, G. P., Stouten, P. F. W., Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.–Aided Mol. Des.* 2000, *14*, 383–401.

[15] Levitt, D. G., Banaszak, L. J., POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* 1992, *10*, 229–234.

[16] Hendlich, M., Rippmann, F., Barnickel, G., LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* 1997, *15*, 359–363.

[17] Huang, B., Schröder, M., LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* 2006, *6*, 19–29.

[18] Glaser, F., Rosenberg, Y., Kessel, A., Tal, P., Ben-Tal, N., The ConSurf-HSSP database: The mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* 2005, *58*, 610–617.

[19] Glaser, F., Morris, R., Najmanovich, R., Laskowski, R., Thornton, J., A method for localizing ligand binding pockets in protein structures. *Proteins* 2006, *62*, 479–488.

[20] Ho, C. M. W., Marshall, G. R., Cavity Search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput.-Aided Mol. Des.* 1990, *4*, 337–354.

[21] Kleywegt, G. J., Jones, T. A., Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D. Biol. Crystallogr.* 1994, *50*, 178–185.

[22] Peters, K. P., Fauck, J., Frömmel, C., The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* 1996, *256*, 201–213.

[23] Coleman, R. G., Sharp, K. A.: Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J. Mol. Biol.* 2006, *362*, 441–458.

[24] An, J., Totrov, M., Abagyan, R., Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform.* 2004, *15*, 31–41.

[25] An, J., Totrov, M., Abagyan, R., Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* 2005, *4*, 752–761.

[26] Laurie, A. T, Jackson, R. M., Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 2005, *21*, 1908–1916.

[27] Weisel, M., Proschak, E., Schneider, G., PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* 2007, *1*:7.

[28] Moreau, G., Broto, P., Autocorrelation of molecular structures: application to SAR studies, *Nouv. J. Chim.* 1980, *4*, 757–764.

[29] Broto, P., Moreau, G., Vandyke, C., Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.* 1984, *19*, 66–70.

[30] Wang, R., Fang, X., Lu, Y., Yang, C.-Y., Wang, S., The PDBbind database: methodologies and updates. *J. Med. Chem.* 2005, *48*, 235–242.

[31] Wang, R., Fang, X., Lu, Y., Wang, S., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* 2004, *47*, 2977–2980.

[32] Schneider, P., Schneider, G., Collection of bioactive reference compounds for focused library design. *QSAR Comb. Sci.* 2003, *22*, 713–718.

[33] Kohonen, T., *Self-Organization and Associate Memory*, Springer-Verlag, Heidelberg 1984.

[34] Schneider, G., Wrede, P., Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* 1998, *70*, 175–222.

[35] Stahl, M., Taroni-Osterroth, C., Schneider, G., Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.* 2000, *13*, 83–88.

[36] Thoden, J. B., Wesenberg, G., Raushel, F. M., Holden, H. M., Carbamoyl phosphate synthetase: closure of the B-domain as a result of nucleotide binding. *Biochemistry* 1999, *38*, 2347–2357.

[37] Wang, M., Ng, K. K. S., Cherney, M. M., Chan, L. *et al.*, Non-nucleoside analogue inhibitors bind to an allosteric site on

HCV NS5B polymerase: Crystal structures and mechanism of inhibition. *J. Biol. Chem.* 2003, *278*, 9489–9495.

[38] Hong, L., Koelsch, G., Lin, X., Wu, S. *et al.*, Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. *Science* 2000, *290*, 150–153.

[39] Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* 1997, *23*, 3–25.

[40] Sadowski, J., Kubinyi, H., A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* 1998, *41*, 3325–3329.

[41] Ajay, A., Walters, W. P., Murcko, M. A., Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* 1998, *41*, 3314–3324.

[42] Schneider, G., Schneider, P., in: Kubinyi, H., Müller, G. (Eds.), *Chemogenomics in Drug Discovery*, Wiley-VCH, Weinheim 2004, pp. 341–376.

[43] DeLano, W. L., *The PyMOL Molecular Graphics System*, DeLano Scientific, Palo Alto, CA, USA, 2002.

[44] Fischer, H., Polikarpov, I., Craievich, A. F., Average protein density is a molecular-weight-dependent function. *Protein Sci.* 2004, *13*, 2825–2828.

[45] Qasba, P. K., Involvement of sugars in protein-protein interactions. *Carbohydrate Polymers* 1999, *41*, 293–309.

[46] Pastor, M., Cruciani, G., McLay, I., Pickett, S., Clementi, S., GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* 2000, *43*, 3233–3243.

[47] Cruciani, G., Pastor, M., Guba, W., VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* 2000, *11*, S29–S39.

[48] Cavasotto, C. N., Orry, A. J., Abagyan, R. A., Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins* 2003, *51*, 423–433.

