

**Virtuelles Screening nach Inhibitoren der Protease HtrA
aus *Helicobacter pylori***

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

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

von
Martin Löwer
aus Frankfurt am Main

Frankfurt 2009
(D 30)

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

Dekan: Prof. Dr. Volker Müller

Erster Gutachter: Prof. Dr. Gisbert Schneider
Zweite Gutachterin: PD Dr. Silja Wessler

Datum der Disputation:

Inhaltsverzeichnis

1. Abkürzungsverzeichnis	7
2. Zusammenfassung	8
3. Summary	10
4. Einleitung	12
4.1 Das Humanpathogen <i>Helicobacter pylori</i>	12
4.2 Virulenzfaktoren von <i>Helicobacter pylori</i>	14
4.3 Virtuelles Screening und Wirkstoffentwicklung	19
4.4 Die Protease HtrA von <i>Helicobacter pylori</i> als Ziel für ein virtuelles Screening	22
4.5 Strukturbasiertes virtuelles Screening	22
4.6 Mögliche Sekretionswege der Protease HtrA	24
4.7 Zielsetzungen der Arbeit	26
5. Material und Methoden	28
5.1 Verbrauchsmaterialien	28
5.1.1 Puffer, Lösungen und Reagenzien	28
5.1.2 Reaktionskits	28
5.1.3 Enzyme	35
5.1.4 Vektoren	35
5.1.5 Bakterienstämme	35
5.1.6 Antikörper	35
5.1.7 Polynukleotide	36
5.1.8 Sonstige Materialien	36
5.3 Geräte	37
5.3 Klonierung des Gens <i>hp1018/19</i>	38
5.3.1 Polymerasekettenreaktion	38

5.3.2 Herstellung von chemokompetenten Zellen	39
5.3.3 Klonierung in pGEM-T Easy	39
5.3.4 Transformation und Blau-Weiß Selektion	40
5.3.5 Plasmidpräparation und –extraktion	41
5.3.6 Restriktionsanalysen und –verdaue	41
5.3.7 Glycerinstock	41
5.3.8 Subklonierung in p-GEX-6P-1	41
5.3.9 Sequenzierung	42
5.3.10 Vektorkarten	42
5.4 Expression und Reinigung der Protease HtrA	42
5.4.1 Expressionsbedingungen	42
5.4.2 Reinigung	43
5.4.3 Elution	43
5.4.4 Verdau mit Prescission-Protease	43
5.4.5 Überwachung der Expression	44
5.5 Inhibitionsassays der Protease HtrA	44
5.5.1 Reaktionsbedingungen	44
5.5.2 Westernblot	44
5.5.3 Markierung	44
5.5.4 Entwickeln	45
5.5.5 Stripping	45
5.6 Informatische Methoden	45
5.6.1 PocketPicker	45
5.6.2 LUDI	46
5.6.3 LIQUID	47
5.6.4 Implementierung der Berechnung des virtuellen Liganden	49
5.7 Virtuelles Screening	53
5.8 Homologiemodellierung	56
5.9 Moleküldatenbanken	57
5.10 Software	59
5.10.1 GOLD	59

5.10.2 gnuplot	59
5.10.3 ImageJ	59
5.10.4 Jalview	59
5.10.5 Java	60
5.10.6 MOE	60
5.10.7 PyMOL	60
5.10.8 Python	60
6. Ergebnisse und Diskussion	62
6.1 Übersicht	62
6.2 Klonierung des Gens <i>hp1018/19</i>	62
6.3 Expression der Protease HtrA	65
6.4 HtrA schneidet E-Cadherin	67
6.5 Homologiemodellierung der Protease HtrA	68
6.6 Das virtuelle Ligandenmodell	77
6.7 Retrospektive Validierung	80
6.7.1 Screeningdatenbanken	80
6.7.2 Ergebnisse des retrospektiven Screenings	82
6.7.3 Einfluss der Bindetaschenvorhersage auf die Screeningergebnisse	85
6.7.4 Parameterauswahl für das prospektive Screening	88
6.8 Prospektives Screening	91
6.8.1 Virtuelles Ligandenmodell der Protease HtrA	91
6.8.2 Inhibitionsexperimente mit HtrA	92
6.8.3 Docking und Struktur-Wirkungs-Beziehung	96
7. Ausblick	103
8. Literaturverzeichnis	105
9. Danksagung	117

10. Anhang	119
A) Sequenzierungsdaten	119
B) Zusätzliche Ergebnistabellen der retrospektiven Validation	120
C) Westernblots des prospektiven Screenings	129
D) Parametereinstellungen von GOLD	130
E) Ergebnislisten der retrospektiven virtuellen Screenings	131
F) Prediction of Extracellular Proteases of the Human Pathogen <i>Helicobacter pylori</i> Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA	153
G) Prediction of Type III Secretion Signals in Genomes of Gram- negative Bacteria	162
11. Eidesstattliche Erklärung	173
12. Lebenslauf	174

1. Abkürzungsverzeichnis

Å	Ångström
BEDROC	Boltzmann enhanced discrimination of receiver operating characteristic
CagA	Cytotoxin associated antigen A
CagL	Cytotoxin associated antigen L
COBRA	Collection of Bioactive Reference Analogues
DNA	Desoxyribonukleinsäure
GST	Glutathion-S-Transferase
<i>H. pylori</i>	<i>Helicobacter pylori</i>
HTS	High-throughput screening
IARC	International Agency for Research on Cancer
IUPAC	International Union of Pure and Applied Chemistry
MOE	Molecular Operating Environment
MUV	Maximum Unbiased Validation Datasets
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PPP	Potentieller Pharmakophorpunkt
ROC	Receiver operating characteristic
ROCAUC	Receiver operating characteristic area under curve
SDS-PAGE	Sodium dodecylsulfate polyacrylamide gel electrophoresis
T3SS	Type III secretion system
T4SS	Type IV secretion system
VacA	Vacuolating cytotoxin A

2. Zusammenfassung

Helicobacter pylori (*H. pylori*) ist ein Gram-negatives, mikroaerophiles Bakterium. Es kolonisiert die menschliche Magenschleimhaut, wobei mehr als 50% der Menschheit befallen sind. Als Pathogen begünstigt es die Entstehung von Magengeschwüren und –krebs.

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 *Cytotoxin Associated Antigen A* (CagA) in die Wirtszelle zu schleusen. Diese *Integrine* genannten Membranproteine werden bei polaren Epithelzellen bevorzugt basolateral exprimiert. Außerdem können extrazellulär geschnittene E-Cadherinfragmente im Medium mit *H. pylori* infizierter Zellkulturen nachgewiesen werden. Beide Beobachtungen legen den Schluss nahe, dass eine Protease von *H. pylori* sezerniert wird, die Zell-Zell-Kontakte zerstört, um *H. pylori* den Zugang zur basolateralen Seite der Wirtszellen zu ermöglichen.

Das vom Gen *hp1019* des Stammes *H. pylori* 26695 kodierte Protein HtrA konnte im Rahmen einer Kooperation mit dem Paul-Ehrlich-Institut in Langen im Überstand von *H. pylori* mit proteolytischer Aktivität nachgewiesen werden. Um den Einfluss dieser extrazellulären Protease auf die Infektion von Kulturzellen mit *H. pylori* zu untersuchen, sollte ein niedermolekularer Inhibitor für HtrA gefunden werden. Ein Homologiemodell als Grundlage für ein strukturbasiertes virtuelles Screening wurde erstellt, wobei die aktive Konformation der Protease DegP von *Escherichia coli* als Vorlage diente (PDB ID 3CS0).

Für eine neue, im Rahmen der vorliegenden Arbeit entwickelte Methode, wurde die Software *PocketPicker* eingesetzt, um Größe und Form von potentiellen Bindetaschen auf der Proteinoberfläche von HtrA zu bestimmen. Durch die komplementäre Projektion von Proteintypen auf dieses definierte Volumen kann für eine von *PocketPicker* vorgesagte Bindetasche ein potentielles Pharmakophormodell berechnet und für Datenbanksuchen eingesetzt werden.

In retrospektiven Studien konnte die Funktion dieser Berechnungen für eine Auswahl an pharmakologisch wichtigen Proteinen aus verschiedenen Strukturklassen validiert werden. Dabei stellte sich heraus, dass eine

Überschätzung der Bindetaschengröße durch *PocketPicker* die Anreicherung im virtuellen Screening deutlich verringern kann. Dies lässt den Schluss zu, dass eine präzise Definition der Dimensionen der Bindetasche für das Gelingen eines strukturbasierten virtuellen Screening unerlässlich ist.

Für die Protease HtrA von *H. pylori* konnten erfolgreich 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. 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. 77% bei einer mittleren inhibitorischen Konzentration bei halbmaximaler Inhibition (IC_{50}) von ca. 26 μ M. Dieses Molekül stellt nun einen Startpunkt für die weitere Optimierung zur Leitstruktur dar.

3. Summary

Helicobacter pylori (*H. pylori*) is a Gram negative, microaerophilic bacterium. It colonizes the human gastric mucosa and more than 50 % of the human population is infected.

Experimental results hint that *H. pylori* needs to interact with certain membrane proteins of the host cells in order to build a type IV secretion system and translocate the primary virulence factor Cytotoxin Associated Antigen A (*CagA*) into the cytoplasm of the host cell. These membrane proteins are called *integrins* and are expressed mainly at the basolateral surface of polarized epithelium cells. Also E-cadherin fragments can be found in the medium of culture cells infected by *H. pylori*. Both observations suggest that *H. pylori* actively secretes a protease in order to degrade the cell-cell contacts, allowing access to the basolateral surface of the host cells.

The gene *hp1019* of *H. pylori* strain 26695 codes for the secreted protein HtrA, which was shown to be an active protease and is located in the supernatant of *H. pylori* cultures. The aim of this work is to find a small molecule inhibitor for the protease HtrA in order to study the influence of this protease on the pathogenesis of *H. pylori* infections. A homology model of HtrA based on the template DegP of *Escherichia coli* (PDB identifier 3CS0) is used to perform a structure based virtual screening.

Within this task a novel method for structure based virtual screening was developed. The software *PocketPicker* is used to predict shape and size of potential binding sites on protein surfaces. By projecting complementary interaction features of the protein atoms inside this defined space a structure based pharmacophore model is calculated and used for database screenings.

In retrospective studies this approach was validated for a range of protein targets of pharmaceutical interest. The performance of the screening was dependent on the quality of the *PocketPicker* predictions. An overestimation of the size of the binding site may lead to a decreased enrichment.

Three different structured based pharmacophore models were calculated for the protease HtrA, each including a different set of predicted binding sites. The compound databases of the commercial providers Asinex and Specs were

screened using these models and a total of 26 compounds were picked from the result lists. Six compounds inhibit the proteolytic activity on a recombinant substrate in an *in-vitro* assay. The best performing compound has a maximal inhibition of about 77%. The half maximum inhibition (IC_{50}) is achieved at a concentration of about 26 μ M. This molecule can serve as a starting point for a further optimization towards a lead compound.

4. Einleitung

4.1 Das Humanpathogen *Helicobacter pylori*

„*The experiment had succeeded – Helicobacter was a proven pathogen.*“

Barry J. Marshall über seinen Selbstversuch mit *Helicobacter pylori*,
Vortrag im Rahmen der Verleihung des Nobelpreises für Medizin 2005

Helicobacter pylori (*H. pylori*) ist ein Gram-negatives, mikroaerophiles, stäbchen-förmiges Bakterium (Abbildung 1; Brown, 2000) und kolonisiert die menschliche Magenschleimhaut (Marshall und Warren, 1984). Zwischen dieser Infektion und Krankheiten wie der gastroduodenalen Ulkusbildung, Adenokarzinomen und MALT-Lymphomen besteht eine Korrelation (Mbulaiteye *et al.*, 2009). *H. pylori* wird von der IARC (*International Agency for Research on Cancer*, Teil der Weltgesundheitsorganisation der Vereinten Nationen) als Gruppe 1 Karzinogen (*Group 1: Carcinogenic to humans*) eingestuft, wie zum Beispiel auch Asbest oder Tabakrauch (IARC 1994). Bis zu der Entdeckung von *H. pylori* im Jahre 1983 durch B. J. Marshall und J. R. Warren (Marshall und Warren, 1984) wurde angenommen, dass kein Mikroorganismus unter den aziden Bedingungen des Magens überleben kann. Auch wurden anstatt einer Infektion primär psychische Ursachen als Auslöser für Magengeschwüre vermutet. Marshall und Warren wurden für ihre Arbeit 2005 mit dem Nobelpreis für Physiologie oder Medizin ausgezeichnet.

Die übliche Behandlung einer festgestellten *H. pylori* Infektion ist eine Kombinationstherapie aus zwei Antibiotika und einem Protonenpumpenhemmer zur Behandlung der Symptome (z.B. Amoxicillin, Clarithromycin und Omeprazol; Graham 2008). Die Behandlung führt zu einer Eradikation der *H. pylori* Infektion. Allerdings sind bereits verschiedene Stämme von *H. pylori* bekannt, die gegen eines oder mehrere Antibiotika resistent sind (Graham und Shiotani, 2008). Die Genesungsraten durch die Dreifachtherapie sind teilweise auf unter 80% gesunken (Graham und Shiotani, 2008).

Eine gemeinsame Evolution von Mensch und *H. pylori* über einen langen Zeitraum wird angenommen (Linz *et al.*, 2007). Die phylogeografische Struktur von *H. pylori* ähnelt der des Menschen und es wird angenommen, dass, zusammen mit der Besiedelung der Erde durch den Menschen ausgehend von Ostafrika, auch *H. pylori* migriert ist. Dies lässt sich durch Vergleiche der angenommenen Wanderbewegungen des Menschen und der Unterschiede der genetischen Struktur verschiedener *H. pylori* Isolate postulieren (Linz *et al.*, 2007). Durch genetische und fossile Beobachtungen am Menschen wird angenommen, dass das so genannte ‚*out of Africa*‘ Ereignis, bei dem die menschliche Spezies Afrika erfolgreich verlassen hatte, 45.000 bis 75.000 Jahre zurückliegt (Liu *et al.*, 2006). Simulationen des genetischen Stammbaums von *H. pylori* datieren dieses Ereignis auf 58.000 Jahre zurück. Eine zu *H. pylori* verwandte Spezies ist *Helicobacter acionychis*, welche spezifisch große Katzen (*Felidae*) infiziert. Vergleiche von genetischen Sequenzdaten geben Hinweise darauf, dass ein Sprung vom menschlichen auf den katzenartigen Wirt in den letzten 200.000 Jahren stattgefunden hat (Eppinger *et al.*, 2006).

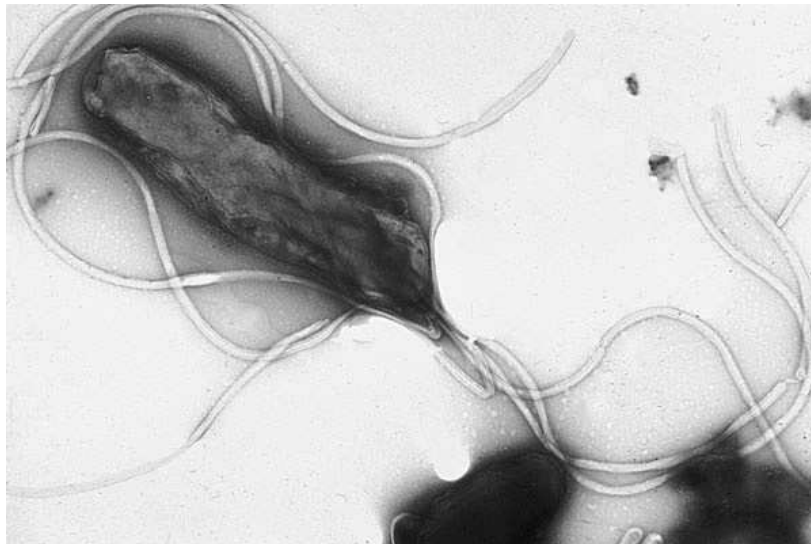


Abbildung 1. Elektronenmikroskopische Aufnahme einer *Helicobacter pylori* Zelle (Yutaka Tsutsumi, M.D., Fujita Health University).

Mehr als 50% der menschlichen Bevölkerung sind mit *H. pylori* infiziert (Ryan *et al.*, 2005). In „entwickelten“ Ländern ist dieser Anteil allerdings geringer, was an besseren hygienischen Verhältnissen liegt, da ein oral-fäkaler Infektionsweg für

H. pylori angenommen wird (Brown, 2000) und an der höheren Verfügbarkeit von Antibiotika, welche gegen andere Infektionen eingenommen werden und eine möglicherweise unentdeckte *H. pylori* Infektion kurieren.

Die gemeinsame Geschichte von Mensch und *H. pylori* ist länger als bei anderen humanpathogenen Bakterien. So ist zum Beispiel *Yersinia pestis*, der Erreger der Pest, erst in den letzten 1.500 bis 20.000 Jahren aus *Yersinia pseudotuberculosis*, einem Tierpathogen, entstanden (Achtman *et al.*, 2006). Dies könnte ein Hinweis auf eine Beziehung zwischen beiden Spezies sein, bei der auch der Mensch als Wirtsorganismus Vorteile hat. In der Tat existieren Hinweise, dass eine Infektion mit *H. pylori* das Risiko verringert, an Adenokarzinomen der Speiseröhre zu erkranken (Rokkas *et al.*, 2007). Auch auf eine Refluxösophagitis, also eine entzündliche Erkrankung der Speiseröhre, kann eine Infektion mit *H. pylori* eine positive Wirkung haben (Delaney und McColl, 2005). Im Allgemeinen wird angenommen, dass *H. pylori* den pH-Wert des Magens reguliert und so auf beide Krankheiten wirkt; die genauen Mechanismen sind jedoch unbekannt. Weiterhin zeigen Studien, dass *H. pylori* durch Stimulation des Immunsystems eine schützende Rolle gegen allergische Erkrankungen des Atemtrakts haben könnte (Cremonini und Gasbarrini, 2003). Aus diesen Gründen wird mittlerweile auch überlegt, ob eine Eradikation einer *H. pylori* Infektion immer von Vorteil ist oder ob es von Fall zu Fall besser wäre, die möglicherweise bedeutenden Vorteile gegen die potentiellen Nachteile abzuwägen (Cremonini und Gasbarrini, 2003). Die vorhandenen Daten lassen eine Beantwortung dieser Frage allerdings nicht zu, was weitere Forschung auf dem Gebiet der Wirts-Pathogen Beziehungen von *H. pylori* motiviert. So werden neue Behandlungskonzepte diskutiert (Nishimori *et al.*, 2008) und auch die Erforschung der molekularen Grundlagen der Infektion könnte neue Ziele für Wirkstoffe liefern.

4.2 Virulenzfaktoren von *Helicobacter pylori*

Verschiedene Virulenzfaktoren von *H. pylori* sind bekannt. *H. pylori* sezerniert das Enzym Urease, welches die Reaktion von Harnstoff zu Ammoniak

katalysiert und den pH-Wert in der unmittelbaren Umgebung der ansonsten säurenlabilen Bakterienzellen neutralisiert (Andersen, 2007).

Das Protein VacA ist ein multifunktionelles sezerniertes Cytotoxin. Zum einen unterstützt es die anfängliche Kolonisierung des Magens durch *H. pylori*, zum anderen werden Effekte vermutet, die das Wirtsimmunsystem beeinflussen und so eine persistente Infektion erleichtern (Cover und Blanke, 2005).

Die Beweglichkeit des Bakteriums ist ein essentieller Faktor für die Kolonisierung der Magenschleimhaut durch *H. pylori*. Dies wird durch mehrere Flagellen sichergestellt (Ryan *et al.*, 2005).

Zusätzlich besitzt *H. pylori* mindestens 32 Proteine in der äußeren Zellmembran, von denen viele eine Rolle für die Adhärenz der Bakterienzellen an die Wirtszellen einnehmen. Die Adhärenz verhindert ein Entfernen der Bakterien unter anderem durch peristaltische Bewegungen des Magens (Aspholm *et al.*, 2006). So vermitteln zum Beispiel die Proteine BabA und SabA eine Bindung an Blutgruppenantigene der Epithelzellen (Lu *et al.*, 2005).

Das Protein CagA wird durch ein Typ IV Sekretionssystem (T4SS; Backert und Selbach, 2008) in die Wirtszelle injiziert (Abbildung 2). Dort werden Tyrosinseitenketten von CagA spezifisch durch Kinasen der Familien Src und Abl phosphoryliert (Backert und Selbach, 2008). Von dieser Modifikation ist die Induktion des so genannten *scatter* Phänotyps abhängig (Abbildung 3; Backert und Selbach, 2008). Die infizierten Kulturzellen sind elongiert und depolarisiert (Moese *et al.*, 2004), was auf eine Veränderung des Cytoskeletts hindeutet. Zusätzlich wurde von Weydig *et al.* gezeigt (Weydig *et al.*, 2007), dass E-Cadherin abhängige Zell-Zell Verbindungen (wie z.B. *Adherens Junctions*) während der Infektion abgebaut werden. Dies geschieht unabhängig von CagA, da eine entsprechende Deletionsmutante von *H. pylori* dieselbe Wirkung hatte. Es konnte gezeigt werden, dass die Catenin vermittelte Bindung des E-Cadherin an das Cytoskelett (van Roy und Berx, 2008) aufgelöst und E-Cadherin extrazellulär proteolytisch geschnitten wird. Die Autoren vermuten einen mehrstufigen Infektionsablauf, in dessen früher Phase *H. pylori* eine Protease sezerniert, welche die Proteolyse katalysiert.

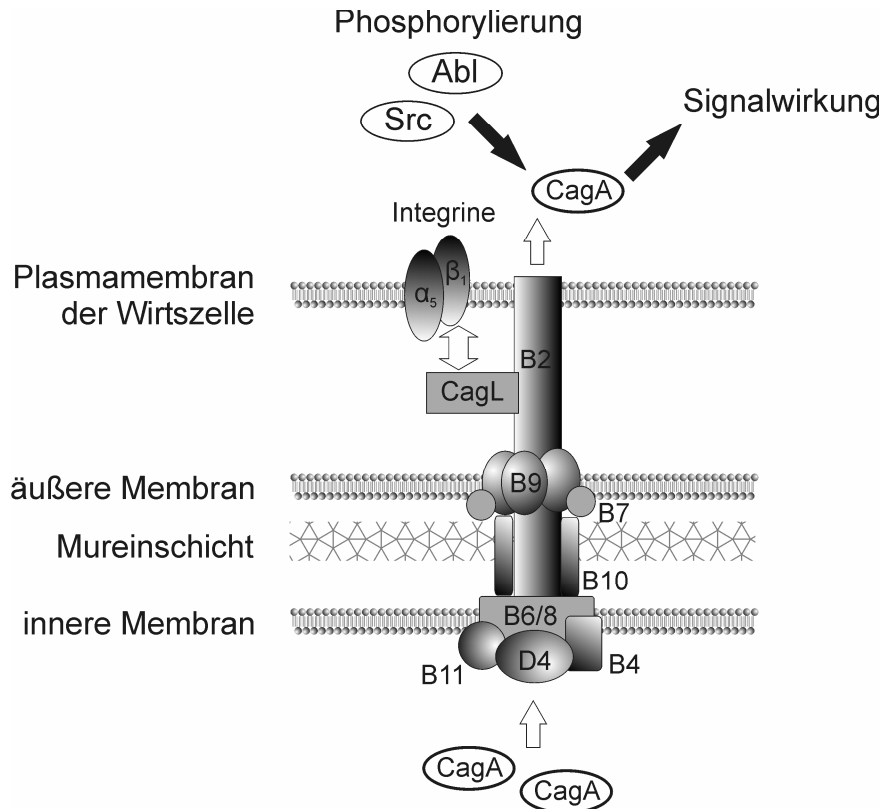


Abbildung 2. Schematische Darstellung des Typ IV Sekretionssystems (T4SS) von *H. pylori* und Injektion von CagA. Die Bezeichnung der Untereinheiten des T4SS erfolgt nach Kwok *et al.* (Kwok *et al.*, 2007).

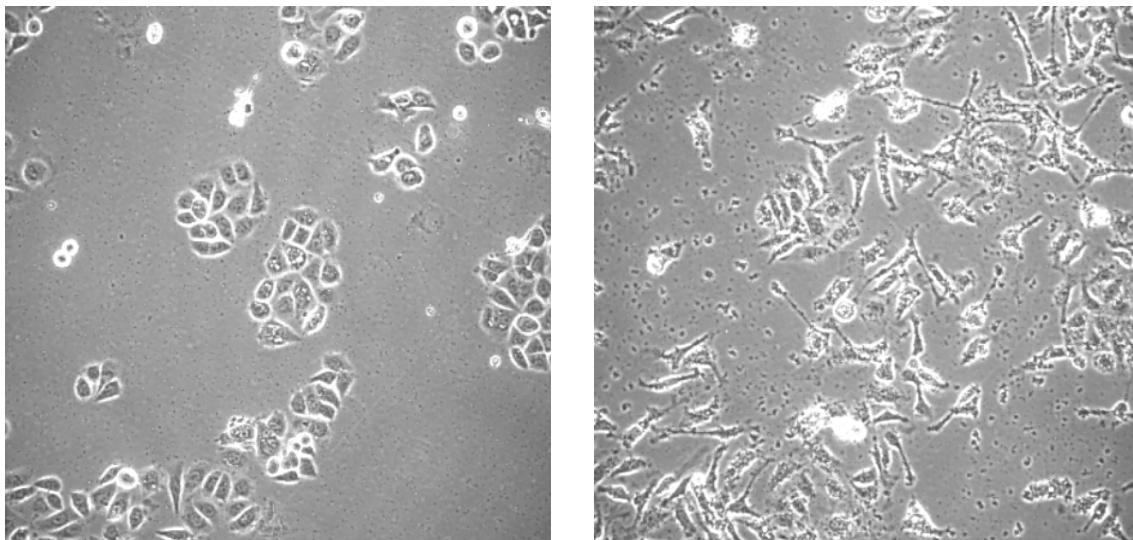


Abbildung 3. AGS Magenepithelzellen vor (links) und vier Stunden nach (rechts, *scatter* Phänotyp) einer Infektion mit *H. pylori* P12 (PD Dr. Silja Wessler, Paul Ehrlich Institut, Langen). Die typischen Änderungen des Phänotyps sind deutlich zu sehen.

Die Vermutung eines mehrstufigen Ablaufs der Infektion, wobei die frühen Phasen CagA unabhängig sind, wurde von Kwok *et al.* (Kwok *et al.*, 2007) bestätigt. Es wurde gezeigt, dass die Translokation von CagA in die Wirtszelle erst erfolgen kann, wenn die Untereinheit CagL des T4SS mit bestimmten Membranproteinen der Wirtszelle interagieren kann (Abbildung 2). Diese Membranproteine gehören zur Familie der Integrine (Hynes, 2002), welche in den polarisierten Epithelzellen bevorzugt auf der basolateralen Seite exprimiert werden. Diese Seite der Wirtszellen und damit die Injektion von CagA kann vermutlich nur erreicht werden, wenn *H. pylori* zuvor die Barriere der Zell-Zell Verbindungen überwindet.

Das Modell für den hypothetischen Infektionsweg, welches die Grundlage für die vorliegende Arbeit bildet, ist in Abbildung 4 dargestellt. Im frühen Stadium der Infektion befindet sich die Bakterienzelle im Magenlumen und sezerniert, neben Urease und möglicherweise VacA, eine oder mehrere Proteasen, welche die Proteine der Zell-Zell Verbindungen degradieren. Dies erlaubt es *H. pylori*, in das Epithelgewebe einzudringen, womit die räumliche Voraussetzung für die CagL- β 1 Integrin Interaktion erfüllt wird.

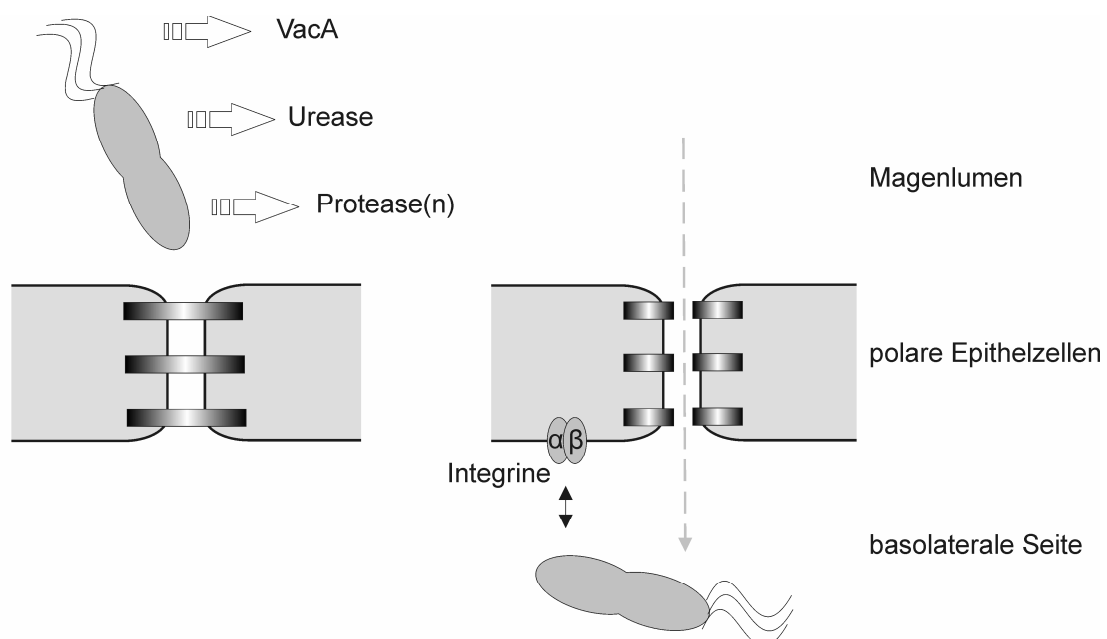


Abbildung 4. Schematische Darstellung der hypothetischen mehrstufigen Infektion von Epithelzellen der Magenschleimhaut durch *H. pylori*. Die linke Seite zeigt eine frühe Phase der Infektion, die rechte Seite einen angenommenen späteren Zustand.

Allerdings ist wenig bekannt, welche Protease sezerniert und durch welches Gen von *H. pylori* sie kodiert wird. Eine Metalloprotease mit einer Masse von ca. 200 kDa konnte im Überstand von *H. pylori* Flüssigkulturen nachgewiesen werden (Windle und Kelleher, 1997) und eine Kollagenase, kodiert durch das Gen *Hp0169*, wird von *H. pylori* sezerniert (Kavermann *et al.*, 2003). Das Genprodukt des Gens *hp1019*, welches als Protease HtrA annotiert ist (Tomb *et al.*, 1997), wurde erstmals 2002 im Kulturmedium von *H. pylori* 26695 nachgewiesen (Bumann *et al.*, 2002). Allerdings basierte die Beschreibung in der Genomdatenbank nur auf automatisierten Vorhersagen (Tomb *et al.*, 1997). Die proteolytische Wirkung dieses Enzyms wurde 2008 von unserer Gruppe beschrieben (Löwer *et al.*, 2008). In derselben Arbeit wurden weitere potentielle für Proteasen kodierende Gene von *H. pylori* durch einen bioinformatischen Ansatz gefunden.

Die Protease HtrA ist homolog zum Protein DegP von *Escherichia coli* (40,4% paarweise Sequenzidentität, E-Wert $4 \cdot 10^{-82}$, BLOSUM80 Substitutionsmatrix; Kim und Kim, 2005; Löwer *et al.*, 2008). DegP hat sowohl eine Chaperon- als auch ein Proteasefunktion. Die Aktivität dieser Funktionen wird über das Bilden von Homomultimeren reguliert. Die Hexamerform ist dabei inaktiv, kann aber ungefaltete Proteine binden und so in eine Dodeca- oder 24-merform übergehen, die die Faltung des Substrats begünstigt oder das Substrat proteolytisch spaltet (Krojer *et al.*, 2008). Dagegen konnte gezeigt werden, dass HtrA von *H. pylori* auch als Monomer proteolytisch aktiv ist (Löwer *et al.*, 2008). Mittlerweile sind verschiedene Strukturmodelle von DegP in der Protein Data Bank (PDB; Berman *et al.*, 2007) verfügbar (Krojer *et al.*, 2002; Krojer *et al.*, 2008; Jiang *et al.*, 2008), welche als Vorlage für ein Strukturmodell von HtrA dienen können.

4.3 Virtuelles Screening und Wirkstoffentwicklung

„*Space is big.*“

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

Der chemische Raum (engl. *chemical space*), aufgespannt durch die Menge aller chemischen Verbindungen, wird auf eine Größe von bis zu 10^{400} Molekülen geschätzt (Winkler und Burden, 2002). Diese riesige Zahl steht im Kontrast zu den Anforderungen der pharmazeutischen Industrie an die Entwicklung von Wirkstoffen. Während dieses Prozesses gilt es, eine einzige Substanz aus allen Möglichen auszuwählen, die optimal an das gegebene Wirkstoffziel bindet und eine hohe Bioverfügbarkeit aufweist (Reddy *et al.*, 2007). Im besten Fall geschieht dies mit minimalen Entwicklungskosten und in einem geringen Zeitrahmen (Oprea, 2005).

In der Vergangenheit haben oftmals zufällige Entdeckungen zu einem fertigen Produkt geführt (Reddy *et al.*, 2007). Eine systematische Methode dagegen ist zum einen das so genannte Hochdurchsatzscreening (engl. *high-throughput screening*, HTS), bei dem hunderttausende von Substanzen automatisiert auf ihre Wirkung in einem bestimmten Experiment getestet werden (Oprea, 2005). Solche Verfahren haben aber den Nachteil einer geringen Trefferrate (Assay Drug Dev Technol, 2008) und verursachen hohe Kosten (Seifert und Lang, 2008). In der CAS Registry Datenbank (Weisgerber, 1997) sind zurzeit ungefähr 43 Millionen synthetisierte Substanzen registriert, was die Obergrenze für die Anzahl an HTS Experimenten darstellt und um viele Größenordnungen kleiner als der tatsächliche chemische Raum ist.

Auf der anderen Seite kann versucht werden, eine bekannte Struktur einer aktiven Verbindung zu verändern, um neue Wirkstoffklassen zu finden (Jalaie und Shanmugasundaram, 2006). Möglich ist zum Beispiel eine Variation der Seitenketten einer Substanz oder auch das Suchen einer neuen Grundstruktur, das so genannte *scaffold-hopping* (Schneider *et al.*, 1999). Eine solche Struktur ist allerdings nicht in allen Fällen gegeben.

Durch Fortschritte auf dem Gebiet der Computertechnik und Proteinkristallographie sind mittlerweile rechnergestützte Ansätze ein wichtiger Bestandteil der akademischen und industriellen Forschung auf dem Bereich der Wirkstoffentwicklung geworden (Reddy *et al.*, 2007). Diese Ansätze schließen die experimentelle Forschung nicht aus, können aber theoretisch den gesamten chemischen Raum beschreiben und so helfen, eine geeignete Teilmenge an Substanzen für eine geringe Anzahl an Experimenten auszuwählen. Dies geschieht durch das so genannte *virtuelle Screening*, bei dem computergestützt schnell große Substanzdatenbanken bewertet werden. Ziel ist es, die Substanzen zu finden, die wahrscheinlich an ein gegebenes Wirkstoffziel binden (Fara *et al.*, 2006).

Virtuelles Screening wird hauptsächlich in den ersten Phasen eines Projekts zur Wirkstoffentwicklung genutzt (Abbildung 5), besonders in den Bereichen der Leitstruktursuche, also der Suche nach einer Substanz, die zumindest eine suboptimale Wirkung auf das gegebene Wirkstoffziel hat und der Optimierung dieser Leitstruktur im Hinblick auf ihre Aktivität (Reddy *et al.*, 2007). Bioinformatische Methoden können beim Verständnis der Biologie einer Krankheit hilfreich sein (Löwer *et al.*, 2008), andere computergestützte Ansätze berechnen pharmakokinetische und toxikologische Eigenschaften von Substanzen für die medizinische Chemie (Schneider und Baringhaus, 2007).

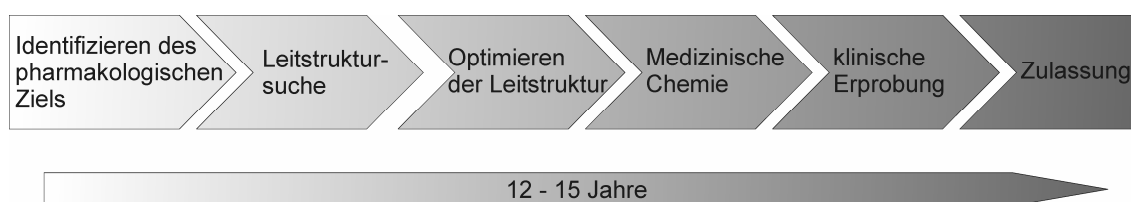


Abbildung 5. Typischer Ablauf eines Projekts zur Wirkstoffentwicklung (verändert übernommen von Roses, 2007 und Schneider und Baringhaus, 2008).

Die Wahl der Methode(n) für ein virtuelles Screening wird zum einen durch die vorhandenen Daten und zum anderen durch die vorhandene Rechenkapazität bedingt (Abbildung 6; Seifert und Lang, 2008). Die Datengrundlage kann entweder aus der Kristallstruktur eines Zielproteins bestehen oder aus einem

bekanntem Liganden. Im besten Fall ist beides bekannt oder jeweils mehrere Beispiele.

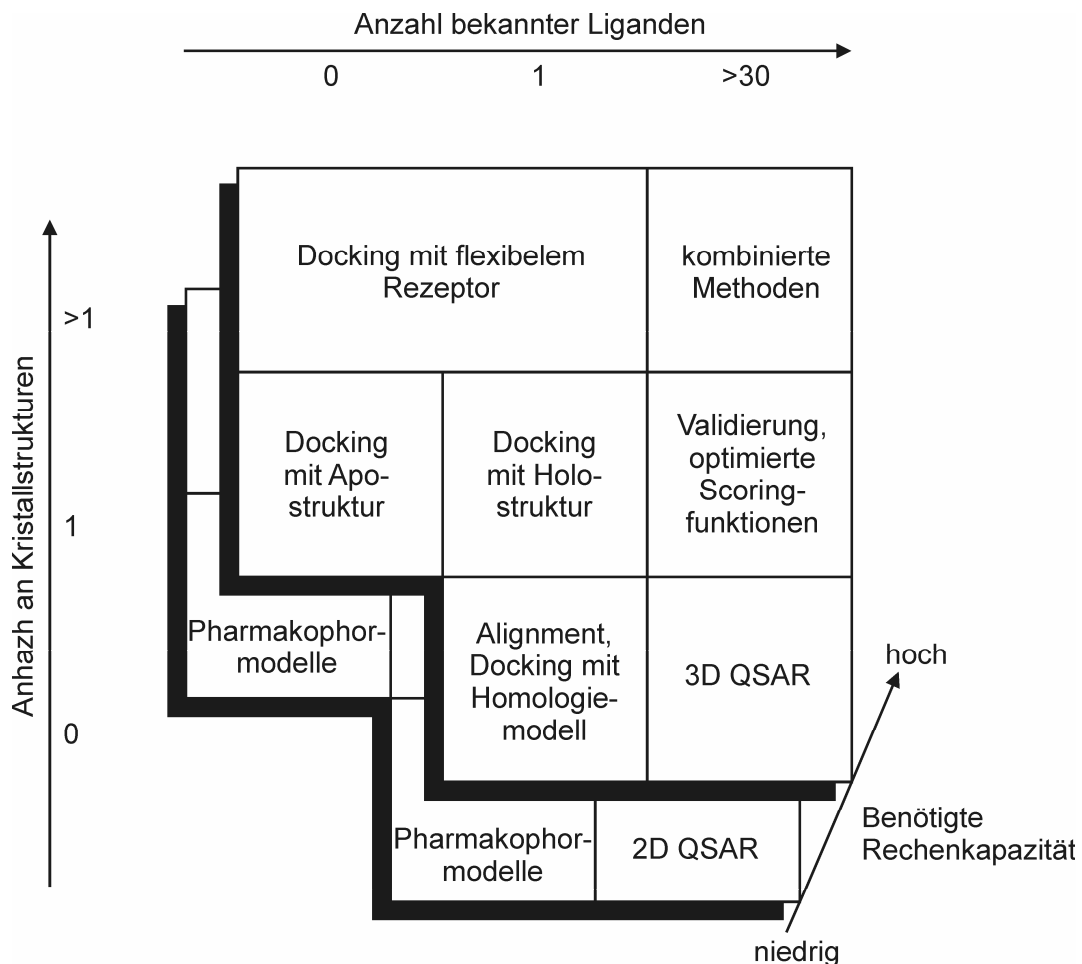


Abbildung 6. Schematische Darstellung möglicher Verfahren für das virtuelle Screening in Abhängigkeit von vorhandenen Daten und Ressourcen (übernommen aus Seifert und Lang, 2008).

Das Ziel eines strukturbasierten virtuellen Screenings (SBVS) ist es, Moleküle zu finden, die komplementär zur Vorlage (also einem Teil des Zielproteins) sind und in die entsprechende Bindetasche binden (Rarey, 2005). Das Molekül soll zum Protein passen wie ein Schlüssel in ein Schloss (Schlüssel-Schloss-Prinzip; Fischer, 1894), beziehungsweise wie eine Hand in einen Handschuh (*induced-fit*; Koshland, 1958), da sich in der Regel weder Protein noch Ligand während der Bindung starr verhalten.

Ein ligandenbasiertes virtuelles Screening (LBVS) soll möglichst zur Vorlage (also einem bekannten Liganden) ähnliche Moleküle liefern (Rarey, 2005). Der

Grundgedanke ist dabei das Ähnlichkeitsprinzip (oder auch *Prinzip der starken Kausalität*; Rechenberg, 1973), nach dem ähnliche Moleküle eine ähnliche biologische Wirkung haben (Johnson und Maggiora, 1990).

Einen Vergleich zwischen ligandenbasierten und strukturbasierten Methoden haben zum Beispiel Hawkins *et al.* vorgenommen (Hawkins *et al.*, 2006). Dabei wurde gezeigt, dass die untersuchte ligandenbasierte Methode gleich gute und oftmals bessere Ergebnisse erreicht.

4.4 Die Protease HtrA von *Helicobacter pylori* als Ziel für ein virtuelles Screening

Die Protease HtrA von *H. pylori* hat als sezerniertes Protein möglicherweise eine wichtige Rolle in den frühen Phasen der Pathogenese (Löwer *et al.*, 2008). Eine spezifische Inhibition von HtrA könnte helfen, diese Rolle in *in-vitro* Studien weiter zu untersuchen.

Bisher sind keine Inhibitoren für HtrA bekannt. Auch die Substratspezifität von HtrA ist unbekannt. Die Aminosäuren N- und C-terminal von der Schnittstelle der Protease wären ansonsten eine Vorlage für einen peptidanalogen Inhibitor gewesen (Böhm *et al.*, 1996).

Die Strukturmodelle der homologen Protease DegP von *Escherichia coli* (Krojer *et al.*, 2002; Krojer *et al.*, 2008; Jiang *et al.*, 2008) ermöglichen die Vorhersage eines Homologiemodells und damit ein strukturbasiertes virtuelles Screening nach Inhibitoren der Protease HtrA.

4.5 Strukturbasiertes virtuelles Screening

„Essentially, all models are wrong, but some are useful.“

George E. P. Box, Statistiker

Das am meisten genutzte Verfahren für ein strukturbasiertes virtuelles Screening ist das so genannte *Docking* (Seifert und Lang, 2008). Im

Allgemein wird zuerst eine Konformation des an das Zielprotein gebundenen Liganden vorhergesagt (engl. *pose prediction*) und dann die Interaktion des Liganden mit dem Protein bewertet (engl. *scoring*). Besonders wenn auch das Protein als flexibel behandelt wird, benötigt Docking große Rechnerressourcen (Jalaie und Shanmugasundaram, 2006). Ein weiterer Nachteil ist, dass die verschiedenen existierenden Bewertungsfunktionen zwar für das virtuelle Screening geeignet sind, aber daran scheitern, Bindungsaffinitäten verlässlich vorherzusagen (Warren *et al.*, 2006). Auch ist die Wahl des Dockingprogramms und der Bewertungsfunktion abhängig vom jeweiligen Zielprotein, wobei die optimale Kombination nicht im Voraus ermittelt werden kann (Barillari 2008).

Ein anderer Ansatz für das strukturbasierte virtuelle Screening sind strukturabgeleitete Pharmakophormodelle. Ursprünglich wurde der Begriff des Pharmakophors durch Paul Ehrlich geprägt (Ehrlich, 1909) und bezieht sich auf das Wirkstoffmolekül selbst. Nach der IUPAC Definition von 1998 (Wermuth *et al.*, 1998) ist ein Pharmakophor die Gesamtheit der sterischen und elektrostatischen Eigenschaften, die für eine optimale zwischenmolekulare Interaktion mit einer spezifischen biologischen Zielstruktur und für das Aktivieren oder Blockieren einer biologischen Antwort notwendig sind.

Die Grundannahme für strukturabgeleitete Pharmakophormodelle ist, dass diese Menge (oder zumindest eine Teilmenge) von Eigenschaften des Wirkstoffmoleküls allein von der Zielstruktur abgeleitet werden kann. Zwei verschiedene Methoden werden dabei häufig eingesetzt oder als Komponenten in umfangreichere Modellberechnungen miteinbezogen. Das Programm GRID (Goodford, 1985) bewegt virtuelle Sonden entlang der Punkte eines Gitters, welches über das dreidimensionale Strukturmodell des Zielproteins gelegt wurde und berechnet für jeden Gitterpunkt die Wechselwirkungsenergien zwischen der Sonde und dem Protein. Als Sonden können Moleküle, Molekülfragmente oder einzelne Atome eingesetzt werden. Für jede Sonde ist das Ergebnis eine dreidimensionale Karte von günstigen Wechselwirkungen der Sonde mit dem Protein. Das Programm LUDI (Böhm, 1992) ist ein Programm für das *de-novo* Design von Proteinliganden. Der erste Schritt ist dabei auch das Erstellen einer dreidimensionalen Karte von bevorzugten

Wechselwirkungspunkten. Allerdings wird dabei ein fester Satz von geometrischen Regeln benutzt um ausgehend von den Proteinatomen die Wechselwirkungspunkte in den Raum zu projizieren.

Ein Problem ist besonders bei den LUDI-Karten, dass oft mehr Interaktionen vorhergesagt werden als im biologischen Kontext sinnvoll wären (Barillari *et al.*, 2008). Es gilt also, aus allen möglichen Interaktionen die auszuwählen, die ein nützliches Pharmakophormodell ergeben. Verschiedene Lösungsvorschläge für dieses Problem wurden bisher publiziert. Das Programm HS-Pharm (Barillari *et al.*, 2008) benutzt einen Ansatz aus dem Bereich des maschinellen Lernens, um eine Auswahl an Proteinatomen zu treffen, die für die LUDI-Karte in Betracht gezogen werden. Schüller *et al.* (Schüller *et al.*, 2006) schlagen dagegen die Kodierung aller Eigenschaften in einen Deskriptorvektor vor, mit welchem ein aufwändiges dreidimensionales Alignment umgangen wird.

Weitere strukturbasierte Methoden für das virtuelle Screening basieren auf GRID, wie zum Beispiel das Programm *FLAP* (Baroni *et al.*, 2007), welches die Form von und potentielle Vier-Punkt-Pharmakophore in Proteinbindetaschen in einer Matrix beschreibt. Diese wird für die Ligand-Protein Vergleiche herangezogen.

Eine nicht alleine auf der Struktur des Zielproteins basierende Methode stellen zum Beispiel die Programme *LigandScout* (Wolber und Langer, 2005) oder *Pocket* (Chen und Lai, 2006) dar. Beide Programme starten mit einem Protein/Ligand-Komplex um ein Pharmakophormodell abzuleiten, welches auf Liganden- und Strukturinformationen beruht.

4.6 Mögliche Sekretionswege der Protease HtrA

Proteinsekretionssysteme von Gram-negativen Bakterien werden grob in sechs Klassen unterteilt (Bingle *et al.*, 2008). Dies geschieht aufgrund der jeweiligen Abstammung der einzelnen Systeme. So ist zum Beispiel das T4SS, wie es auch von *H. pylori* exprimiert wird, mit bakteriellen Konjugationssystemen verwandt (Backert und Selbach, 2008).

Trotz der nachgewiesenen extrazellulären Lokalisation von HtrA ist dessen genauer Sekretionsweg nicht bekannt. Bisher konnte gezeigt werden, dass HtrA wahrscheinlich ein N-terminales Signalpeptid für den *Sec*-abhängigen Transport besitzt (von Heijne, 1985; Löwer *et al.*, 2008). Das potentielle Signalpeptid wird durch das Gen *hp1018* kodiert, welches allerdings mit *hp1019* einen Leserahmen und ein Genprodukt bildet (Löwer *et al.*, 2008).

Dieser Transportweg erklärt allerdings nur die Translokation in das Periplasma. Der weitere Transport könnte über ein Typ II (Sandkvist, 2001) oder Typ V Sekretionssystem (Henderson *et al.*, 2004) erfolgen. Allerdings ist für Typ II Sekretionssysteme nur wenig über die Eigenschaften bekannt, die ein Effektorprotein dieses Sekretionswegs aufweisen muss (Sandkvist, 2001). Das Typ V Sekretionssystem (auch Autotransporter genannt) ist dagegen eine C-terminale Domäne des Effektorproteins, die den Transport über die äußere Membran ermöglicht (Henderson *et al.*, 2004). Diese Domäne ist in HtrA nicht vorhanden (Löwer *et al.*, 2008).

Im Gegensatz dazu sind die Sekretionssignale für Typ III Sekretionssysteme (T3SS; Sheng *et al.*, 2004) wesentlich besser erforscht (Galán und Wolf-Watz, 2006). T3SS und Flagellen haben eine gemeinsame Abstammung (Pallen und Matzke, 2006) und können wie T4SS Proteine in das Cytoplasma von eukaryotischen Zellen transportieren (Abbildung 7). Aufgrund dieser Eigenschaft sind T3SS oft an der Pathogenese verschiedener pathogener Mitglieder von Bakteriengattungen wie *Yersinia* oder *Salmonella* beteiligt (Shao, 2008; McGhie *et al.*, 2009). Die Sekretionssignale sind in der Primärstruktur kodiert und am N-terminalen Ende des jeweiligen Effektorproteins gelegen. Eine eingehende Untersuchung dieser Signale mit maschinellen Lernverfahren wurde begleitend zur vorliegenden Arbeit durchgeführt und publiziert (Löwer und Schneider, 2009). Das Manuskript der Publikation ist im Anhang F zu finden. Es konnte gezeigt werden, dass

1. eine hohe Korrelation zwischen dem Export eines Proteins und den ersten 30 Aminosäuren der Primärstruktur besteht und
2. die Signale in einer großen Menge von Spezies vorkommen, unabhängig ob ein T3SS im entsprechenden Organismus vorhanden ist.

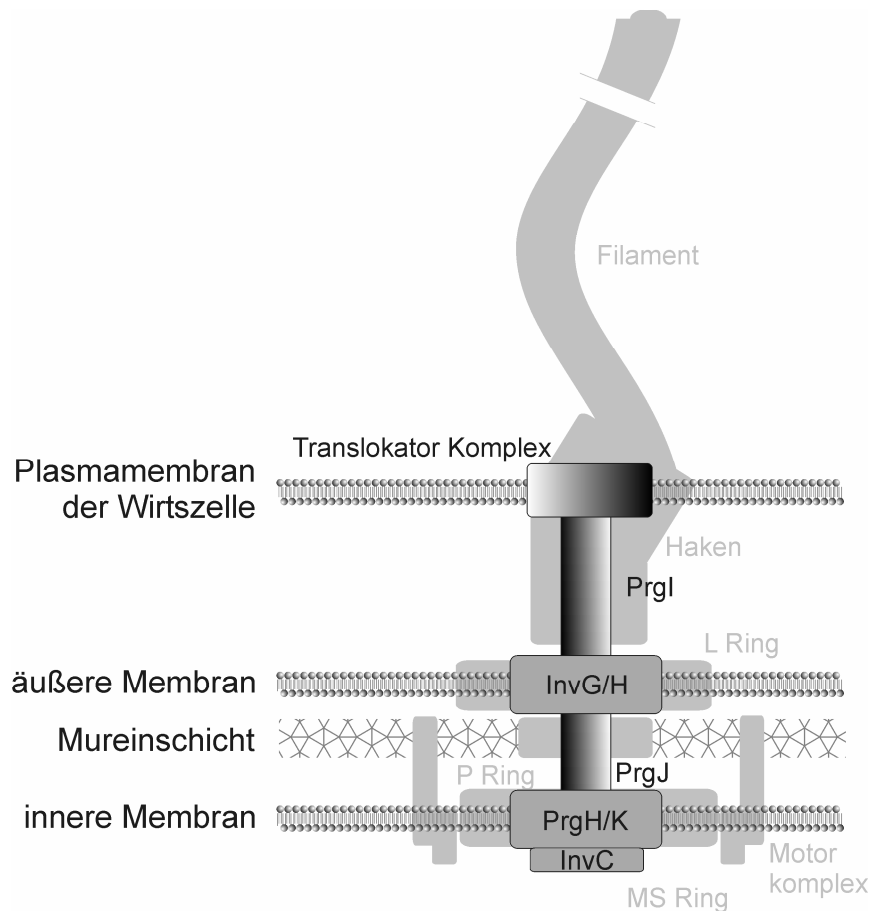


Abbildung 7. Aufbau bakterieller Typ 3 Sekretionssysteme (T3SS). Die Benennung der Komponenten erfolgt nach dem prototypischen T3SS der Gattung *Salmonella* (Sheng *et al.*, 2004). Der Aufbau einer bakteriellen Flagelle (Pallen und Matzke, 2006) ist grau schattiert im Hintergrund zu sehen und verdeutlicht den evolutionären Zusammenhang zwischen T3SS und Flagellen (Pallen und Matzke, 2006).

4.7 Zielsetzungen der Arbeit

Ziel der vorliegenden Arbeit war es, einen niedermolekularen spezifischen Inhibitor für die Protease HtrA von *H. pylori* zu finden. Diese bisher noch wenig beschriebene Protease könnte eine bedeutend für die Pathogenese von *H. pylori* sein. Kurzfristig kann ein solcher Inhibitor eine wichtige Rolle in der weiteren Erforschung der *H. pylori*-Wirt Interaktionen spielen.

Die Suche nach dem Inhibitor wird als strukturbasiertes virtuelles Screening ausgeführt. Als Datengrundlage steht nur ein Homologiemodell zur Verfügung. Docking mit Homologiemodellen wurde schon erfolgreich durchgeführt (Kairys *et al.*, 2006), allerdings enthält ein Homologiemodell potentiell Fehler, die ein Docking erschweren. So ist die genaue Position der Proteinseitenketten ist nicht

bekannt und es liegen für HtrA auch keine Daten vor, in welchem Bereich der Bindetasche(n) das Protein starr oder flexibel ist. Eine komplett flexible Modellierung ist durch die Dockingsoftware unter Umständen nicht möglich. GOLD (Cole *et al.*, 2005) kann zum Beispiel nur maximal zehn manuell ausgewählte Aminosäureseitenketten als flexibel behandeln.

Es wird daher ein neues Modell zur Berechnung strukturabgeleiteter Pharmakophore vorgestellt. Potentielle Interaktionen des Proteins mit einem Liganden werden als Wahrscheinlichkeitsdichten modelliert, was die mögliche Flexibilität und Fehler des Homologiemodells kompensieren soll. Weiterhin wird das beschriebene Problem, dass oft zu viele potentielle Pharmakophoreigenschaften von der Proteinstruktur abgeleitet werden, auf eine neue Weise gelöst: Die Arbeitshypothese ist, dass, anstatt nur die Zahl der Aminosäuren oder Atome des Proteins zu begrenzen, die zur Kartierung potentieller Interaktionen benutzt werden, es zusätzlich auch sinnvoll ist den Raum zu begrenzen, in den die möglichen Interaktionen abgebildet werden. Hierfür wird eine automatisierte Methode zur Bindetaschenvorhersage herangezogen, was zusätzlich den Vorteil hat, dass eine möglicherweise voreingenommene manuelle Auswahl der Bindetasche vermieden wird.

Das neue Modell wurde zunächst in retrospektiven Studien validiert und schließlich für das virtuelle Screening nach Inhibitoren der Protease HtrA eingesetzt.

Parallel zu den informatischen Arbeiten wurde die Gene *hp1018* und *hp1019* kloniert (welche im Folgendem als *hp1018/19* bezeichnet werden, da sie für ein Genprodukt kodieren) und die Protease HtrA rekombinant exprimiert und gereinigt. Die proteolytische Wirkung konnte dann in *in-vitro* Experimenten gezeigt werden und die Protease steht für weitergehende Inhibitions- und Zellkulturexperimente zur Verfügung.

5. Material und Methoden

5.1 Verbrauchsmaterialien

5.1.1 Puffer, Lösungen und Reagenzien

DNA-Ladepuffer

- 80 % Glycerin
- 0,1 % Xylencyanol
- 0,1 % Bromphenolblau

TB-Medium (Medienküche Paul-Ehrlich-Institut)

- 1,2 % Trypton
- 2,4 % Hefeextrakt
- 0,4 % Glycerin
- 15,65 mM KH_2PO_4
- 72 mM K_2HPO_4

LB-Medium

- 1 % Pepton
 - 0,5 % Hefeextrakt
 - 1 % NaCl
- autoklaviert, pH 7,0

LB-Amp-Medium

- 1 % Pepton
 - 0,5 % Hefeextrakt
 - 1 % NaCl
- autoklaviert, pH 7,0
- 100 mg/l Ampicillin

LB-Amp-Agrose Platten

- 1 % Casein
- 0,5 % Hefeextrakt
- 1 % NaCl
- autoklaviert, pH 7,0
- 1,8 % Agarose
- aufkochen, unter Rühren auf ca. 35 – 40 °C abkühlen lassen
- 2 mg/l Ampicillin
- Platten gießen

PBS-Puffer (Medienküche Paul-Ehrlich-Institut)

- 0,8 % NaCl
- 0,02 % KCl
- 0,024 % KH_2PO_4
- 0,144 % Na_2HPO_4
- pH 7,4

TBST-Puffer (Medienküche Paul-Ehrlich-Institut)

- 10 mM Tris-HCl
- 150 mM NaCl
- 0,05 % Tween 20
- pH 8,0

TBE 10x Puffer

- 890 mM Tris-HCl
- 890 mM H_3BO_4
- 25 mM EDTA

TFB I-Puffer

100 mM KCl
50 mM MnCl_2
30 mM Kaliumacetat
10 mM CaCl_2
15 % Glycerin
pH 5,8; autoklaviert

TFB II-Puffer

10 mM KCl
75 mM CaCl_2
10 mM 3-(N-Morpholino)-Propansulfonsäure (MOPS)
15 % Glycerin
pH 7,0; autoklaviert

Waschpuffer

50 mM Tris-HCl
100 mM EDTA
0,1 % Tween 20
PIT nach Bedarf
pH 8,0; sterilfiltriert

Elutionspuffer

50 mM Tris-HCl
10 mM reduziertes Glutathion
pH 8,0; sterilfiltriert

Blockingpuffer

10 % Roti®-Block 100 (10x; Carl Roth)
10 μM Na_3VO_4

Proteasepuffer

50 mM Tris-HCl

150 mM NaCl

1 mM EDTA

1 mM DTT

pH 7,4

Coomassie-Lösung

0,5 % Coomassie G250

40 % Methanol

7 % konzentrierte Essigsäure

Coomassie-Entfärber

40 % Methanol

7 % konzentrierte Essigsäure

Stripping-Puffer

62,5 mM Tris-HCl

2 % SDS

100 mM β -Mercaptoethanol

pH 6,8

Western-Blot 5x Puffer (Medienküche Paul-Ehrlich-Institut)

25 mM Tris-HCl

192 mM Glycin

20 % Methanol

pH 8,3

SDS-PAGE 10x Laufpuffer (Medienküche Paul-Ehrlich-Institut)

25 mM Tris-HCl
192 mM Glycin
0,1 % SDS
pH 8,3

SDS-PAGE 4x Probenpufferpuffer

125 mM Tris-HCl
20 % Glycerin
6 % SDS
0,02 % Bromphenolblau
10 % β - Mercaptoethanol
pH 6,8

HEPES-Puffer (Medienküche Paul-Ehrlich-Institut)

1M 2-(4-(2-Hydroxyethyl)-1-piperazinyl)-ethansulfonsäure
(HEPES)
pH 7,4

Agarose-Gelelektrophorese Gel

0,8 % Agarose in 0,5x TBE-Puffer
oder 1,5 %

IPTG Stammlösung

1M IPTG
sterilfiltriert

X-Gal Lösung

2 % 5-Brom-4-chlor-3-indolyl- β -D-galactosid in N,N'-Dimethyl-
formamid
= 20 mg/ml

Ampicillin-Lösung

10 % Ampicillin
= 100 mg/ml

MgSO₄-Lösung (Invitrogen)

50 mM MgSO₄

MgCl₂-Lösung (Invitrogen)

50 mM MgCl₂

PCR-Puffer (Invitrogen)

200 mM Tris-HCl
500 mM KCl
pH 8,4

dNTP-Mix (Carl Roth GmbH)

10 µM dATP
10 µM dCTP
10 µM dGTP
10 µM dTTP

Restriktionsendonucleasepuffer BamH1, 10x (Fermentas)

10 mM Tris-HCl
5 mM MgCl₂
100 mM KCl
0,02 % Triton X-100
1 mM 2-Mercaptoethanol
0,1 mg/ml BSA
pH 8,0 bei 37°C

Restriktionsendonucleasepuffer EcoR1, 10x (Fermentas)

50 mM Tris-HCl

10 mM MgCl₂

100 mM NaCl

0,02 % Triton X-100

0,1 mg/ml BSA

pH 7,5 bei 37 °C

Restriktionsendonucleasepuffer Tango™, 10x (Fermentas)

33 mM Tris-acetat

10 mM Magnesiumacetat

66 mM Kaliumacetat

0,1 mg/ml BSA

pH 7,9 bei 37 °C

2x Ligationspuffer (Promega)

60 mM Tris-HCl

20 mM MgCl₂

20 mM DTT

2 mM ATP

10 % Polyethylenglycol

pH 7,8

Protease Inhibitor Tablets

(PIT, gemäß Herstelleranweisung dosiert; Roche)

Gluthation Spharose FastFlow (GE Healthcare)

PageRuler™ Proteinstandard (Fermentas)

Rekombinantes E-Cadherin

(0,1 µg/µl; E-Cadherin/Fc-Chimera; R&D Systems)

10x PCR_x Enhancer Lösung (Invitrogen)

DNA Standard (100 bp und 1k bp; Invitrogen)

5.1.2 Reaktionskits

Nucleospin® Plasmid (Macherey Nagel)
Nucleospin® Extract II (Macherey Nagel)
ECL Plus™ Western Blotting Detection System (GE Healthcare)
pGEM-T Easy Vector System (Promega)

5.1.3 Enzyme

Precission-Protease™ (2 u/μl; GE Healthcare)
Pfx DNA-Polymerase (2,5 u/μl; Invitrogen)
Taq DNA-Polymerase (2,5 u/μl; Invitrogen)
Restriktionsendonuclease BamH1 (10 u/μl; Fermentas)
Restriktionsendonuclease EcoR1 (10 u/μl; Fermentas)
T4 DNA-Ligase (3 u/μl; Promega)
RNase A (10 mg/ml; Fermentas)

5.1.4 Vektoren

pGEM-T Easy (50 ng/μl; Promega)
pGEX-6P-1 (GE Healthcare)

5.1.5 Bakterienstämme

Escherichia coli DH5α (Invitrogen)
fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96
recA1 relA1 endA1 thi-1 hsdR17
Escherichia coli BL21-Gold(DE3) (Stratagene)
E. coli B dcm ompT hsdS(r_B-m_B-) gal

5.1.6 Antikörper

Anti-E-Cadherin Sc-7870 (H108), Kaninchen, polyklonal (Santa Cruz),
1:200
Anti-HtrA 8918, Kaninchen (Dr. Steffen Backert, University College
Dublin), 1:1000
HRP-linked Anti-Rabbit (GE Healthcare), 1:5000

5.1.7 Polynukleotide

Vorwärtsprimer für das Gen *hp1018/19*:

5' -AAGGATCCGGCAATATCCAAATCCAGAGCATG-3'

BamH1 Gly18

Rückwärtsprimer für das Gen *hp1018/19*:

5' -AAGAATTCGACCCACCCCTATCAATTCACC-3'

EcoR1

Stop

5.1.8 Sonstige Materialien

Blottingpapier	Munktell & Filter GmbH, Bärenstein
Einmal-Handschuhe	Semperit Technische Produkte GmbH & Co. KG, Wien
Einmal-Feindosierungsspritze (1 ml, 20ml)	B.Braun Melsungen AG, Melsungen
Einwegskalpell	B.Braun Melsungen AG, Melsungen
Flüssigstickstoff	Messer, Griesheim
Impföse	Nunc GmbH & Co. KG, Wiesbaden
Kanüle (Größe 1, Größe 20)	B.Braun Melsungen AG, Melsungen
Kryogefäße	Nunc GmbH & Co. KG, Wiesbaden
Küvetten	Carl Roth GmbH & Co. KG, Karlsruhe
Messpipetten (2ml, 5ml, 10ml, 25ml)	Greiner Bio-One GmbH, Solingen
Nitrilhandschuhe	Carl Roth GmbH & Co. KG, Karlsruhe
Pipettenspitzen	Sarstedt AG & Co., Nümbrecht
PVDF-Membran	Roche Diagnostics GmbH, Mannheim
Reaktionsgefäße (1,5ml, 2ml)	Sarstedt AG & Co., Nümbrecht
Röntgenfilme Super RX	Fujifilm, Kisker Biotech, Steinfurt
Roti@-labo Spritzenfilter	Carl Roth GmbH & Co. KG, Karlsruhe
Schraubröhren (15ml, 50ml)	Sarstedt AG & Co., Nümbrecht

5.2 Geräte

Autoklav	MMM, Münchner Medizin Mechanik GmbH, München
Blotdokumentationssystem FUSION-FX7	Vilber Lourmat Deutschland GmbH
Einschweißgerät	Dual Electronic, Hulme Martin Heat Sealers Ltd., United Kingdom
Eismaschine	Ziegra-Eismaschinen GmbH, Isernhagen
Entwickler CP 1000	Agfa-Gevaert Gruppe, Köln
Gefrierschrank (-20°C)	Fa. Liebherr, Ochsenhausen
Gefrierschrank (-80°C) Hera Freeze	Heraeus, Hanau
Gelelektrophoresekammern	Peqlab Biotechnologie, Erlangen
Geldokumentationssystem Image Master VDS	Pharmacia Biotech, Wien
Geldokumentationssystem FUSION-FX7	Vilber Lourmat GmbH, Eberhardzell
Gel-/ Tank-Blot-Apparaturen Mini-PROTEAN® 3 Cell	Bio-Rad Laboratories GmbH, München
Heizblock Rotilabo®-Block-Heater H250	Carl Roth GmbH & Co. KG, Karlsruhe
Inkubatoren BBD 6220	Fa. Heraeus, Hanau
Kühlschrank (4°C)	Fa. Liebherr, Ochsenhausen
Leuchtpult Hobbylite 2	Kaiser Fototechnik GmbH & Co. KG, Buchen
Magnetrührer/ Heizplatte RCT basic	IKA® Werke GmbH & Co. KG, Staufen
Mikrowelle	Robert Bosch Hausgeräte GmbH, Gerlingen-Schillerhöhe
Netzgerät PowerPac 200	Bio-Rad Laboratories GmbH, München
Photometer Eppendorf BioPhotometer	Eppendorf
Pipetten Eppendorf Reference (0,5-10µl, 2-20µl, 10-100µl, 50-200µl, 100-1000µl)	Eppendorf
Pipettierhilfe Pipetus®	Fa. Hirschmann-Laborgeräte, Eberstadt
Rotator neoLab	Migge Laborbedarfs-Vertriebs GmbH, Heidelberg
Scanner CanoScan 8400F	Canon, Krefeld
Schüttel-Inkubator innova™4200	New Brunswick Scientific, Nürtingen
Schüttel-Inkubator Incubator Shaker G25	New Brunswick Scientific, Nürtingen
Spektrophotometer Ultrospec 2100 pro	GE Healthcare, München
Stickstofftank Chronos	Messer, Griesheim
Thermocycler Personal Cyclor	Biometra, Goettingen
Thermomixer Eppendorf Thermomixer comfort	Eppendorf
Ultraschall-Homogenisator Sonoplus	Bandelin electronic GmbH & Co. KG, Berlin
Vortexer MS2 Minishaker	IKA® Werke GmbH & Co. KG, Staufen

Laborwaage LP820	Sartorius AG, Göttingen
Analysenwaage A 200 S	Sartorius AG, Göttingen
Wasseraufbereitungsanlage Maxima USFP	Veolia Wasser Deutschland GmbH, Berlin
ELGA P	
Zentrifugen:	
Micro Centrifuge	Carl Roth GmbH & Co. KG, Karlsruhe
Biofuge 15R	Heraeus, Hanau
Eppendorff Centrifuge 5415D	Eppendorf
Eppendorff Centrifuge 5415R	Eppendorf
Eppendorff Centrifuge 5810R	Eppendorf
Sorvall® RC 26 plus	Thermo Fisher Scientific, USA

5.3 Klonierung des Gens *hp1018/19*

5.3.1 Polymerasekettenreaktion

Die Polymerasekettenreaktion (engl. *polymerase chain reaction*, PCR) dient zur Vervielfältigung bestimmter Abschnitte genomischer DNA. Grundlage ist dabei die spezifische Bindung kurzer Oligonucleotide, so genannter *Primer*, an die DNA-Vorlage vor und nach dem zu vervielfältigendem Bereich. Durch die Primer werden in dieser Arbeit zusätzliche Restriktionsendonuclease-schnittstellen für BamH1 und EcoR1 eingeführt (siehe Abschnitt 5.1.7). Die Pfx DNA-Polymerase bietet eine hohe Replikationsgeschwindigkeit und hohe Genauigkeit durch eine 3'→5'-Exonuklease-Aktivität.

Für die Amplifikation des Gens *hp1018/19* wurden folgende Reaktionsansätze vorbereitet:

1. (Ansatz A) 1,25 u Pfx DNA Polymerase; 1 mM MgSO₄; je 0,3 mM dATP, dCTP, dGTP, dTTP; je 0,3 µM Vorwärts- und Rückwärtsprimer; 5 µl PCR-Puffer (1x); 5 ng genomische DNA; 5 µl PCR-Enhancer (1x); H₂O ad 50 µl
2. (Ansatz B) 1,25 u Pfx DNA Polymerase; 1 mM MgSO₄; je 0,3 mM dATP, dCTP, dGTP, dTTP; je 0,3 µM Vorwärts- und Rückwärtsprimer; 5 µl PCR-Puffer (1x); 5 ng genomische DNA; 5 µl PCR-Enhancer (2x); H₂O ad 50 µl

Die Kontrolle der Reaktion und die Reinigung der Amplifikate erfolgten über Agarose-Gelelektrophorese.

Der Thermocycler wurde folgendermaßen programmiert:

1. 94 °C 120 Sekunden
2. 35 Zyklen
 94 °C 30 Sekunden Denaturierung
 56 °C 45 Sekunden Primerhybridisierung
 68 °C 120 Sekunden Elongation
3. 68 °C 10 Minuten
4. 4 °C 24 Stunden

5.3.2 Herstellung von chemokompetenten Zellen

Um für die Aufnahme von fremder DNA empfänglich, also kompetent, zu sein, müssen die *Escherichia coli* Zellen entsprechend präpariert werden. Dies geschah nach der so genannten Calciumchlorid-Methode (Hanahan, 1983).

100 ml LB-Medium wurden 1:100 mit einer *Escherichia coli* Flüssigkultur angeimpft, die über Nacht bei 37 °C bis zur Sättigung gewachsen war. Die neue Kultur wurde schüttelnd bei 37 °C bis zu einer optischen Dichte bei 660 nm Wellenlänge von 0,4 – 0,6 inkubiert. Die Bakterienzellen wurden bei 4 °C und 3000g für zehn Minuten zentrifugiert, der Überstand wurde verworfen und die Zellen in 33 ml kaltem TFB I-Puffer resuspendiert. Nach zehn Minuten Inkubation auf Eis wurde die Suspension erneut zentrifugiert (s.o.), der Überstand verworfen und die Zellen in 5 ml kaltem TFB II-Puffer resuspendiert. Die Suspension wurde in Portionen von 100 µl aliquotiert und in flüssigem Stickstoff eingefroren. Die Lagerung erfolgte bei -80 °C.

5.3.3 Klonierung in pGEM-T Easy

Der Plasmid pGEM-T Easy wird in linearer Form geliefert und besitzt an den 3'-Enden des Doppelstrangs eine überhängende Thyminbase. Eine zusätzlich Adenosinbase an den 3'-Enden des PCR-Amplifikats ermöglicht eine effiziente Klonierung. Zusätzlich verhindern die überhängenden Enden eine Rezirkularisierung des leeren Vektors. Um die zusätzliche Adenosinbase zu

binden, kann eine Eigenschaft der Taq DNA-Polymerase ausgenutzt werden, welche, eigentlich fehlerhaft, bei der Strangsynthese einen 3'-Adenin-Überhang erzeugt.

Die amplifizierte DNA wurde über Silicasäulen aus dem Agarosegel extrahiert; 20 µl des Extrakts wurden mit 5 µl Taq-Polymerasepuffer (1x), 1,5 mM MgCl₂, 0,2 mM dATP, 1,25 u Taq-Polymerase und H₂O ad 50 µl für 20 Minuten bei 72°C im Thermocycler inkubiert. Die DNA wurde danach aus dem Reaktionsansatz über Silicasäulen extrahiert und die DNA-Konzentration spektroskopisch bestimmt. Die Ligation in den pGEM-T Easy Vektor wurde nach Herstellerangaben mit ca. 23,6 ng DNA über Nacht bei 4°C durchgeführt.

5.3.4 Transformation und Blau-Weiß Selektion

Die Transformation, also die Aufnahme von fremder DNA durch die kompetenten Bakterien, wurde durch einen Hitzeschock stimuliert. Die folgende Blau-Weiß Selektion (Mühlhardt, 2006) basiert auf dem *lacZ* Gen, welches in beiden verwendeten Plasmiden enthalten ist und für das Enzym β-Galactosidase kodiert. Dieses spaltet ein in der X-Gal Lösung vorhandenes Substrat, so dass ein blauer Farbstoff entsteht. Eine erfolgreiche Klonierung zerstört allerdings das kodierende Gen und diese Klone können dann anhand der fehlenden Blaufärbung auf einer Agarplatte erkannt werden. Der Aufbau des pGEM-T Easy Vektors verhindert zwar eine solche Ligation ohne eingefügte DNA, bei dem Vektor pGEX-6P-1 ist dieses Ergebnis aber möglich.

Die chemokompetenten *Escherichia coli* Zellen wurden auf Eis aufgetaut und mit dem kompletten Reaktionsansatz der Ligation auf Eis für 30 Minuten inkubiert. Das Gemisch wurde danach für sieben Minuten bei 37°C einem Hitzeschock ausgesetzt und danach wiederum zehn Minuten auf Eis inkubiert. Anschließend wurden 800 µl LB-Medium zugegeben, die Zellen bei 37°C für 45 Minuten schüttelnd inkubiert und für eine Minute bei 4000g zentrifugiert. Der Überstand wurde verworfen und das Pellet wurde in 200 µl LB-Medium resuspendiert. Jede LB-Amp-Agrose wurde durch das Verteilen von 50 µl IPTG-Stammlösung und 50 µl X-Gal Lösung vorbereitet und die Zellsuspension wurde auf einer Platte bis zur Trockenheit verstrichen. Die Platten wurden über

Nacht bei 37°C inkubiert. Weiß gefärbte Kolonien wurden mit einem sterilen Zahnstocher aufgenommen und in fünf ml LB-Amp-Medium für 16h bei 37°C schüttelnd inkubiert.

5.3.6 Plasmidpräparation und -extraktion

Alle Plasmide wurden mit dem Nucleospin Plasmid Kit präpariert beziehungsweise mit dem Nucleospin Extract II Kit extrahiert.

5.3.7 Restriktionsanalysen und -verdaue

Um die erfolgreiche Klonierung in die Plasmide zu überprüfen, wurden 3 u EcoR1, 3 u BamH1, 4 µl Tango Puffer (1x), 20 µg RNase A und 15 µl präpariertes Plasmideluat für zwei Stunden bei 37°C inkubiert und auf ein Agraosegel aufgetragen.

Für die Präparation des rekombinanten Gens aus dem Vektor pGEM-T Easy_HtrA (Abbildung 9) und die Vorbereitung des Vektors pGEX-6P-1 wurden 10 u BamH1, 1,5 µl BamH1-Puffer (1x), 20 µg DNA und H₂O *ad* 15 µl für zwei Stunden bei 37°C inkubiert. Die DNA wurde extrahiert. Zu den ca. 30 µl Elutionsvolumen wurden 3 µl EcoR1-Puffer (1x) und 5 u EcoR1 zugeben und der Ansatz wurde zwei Stunden bei 37°C inkubiert.

5.3.8 Glycerinstock

Für die Lagerung von genetisch veränderten Bakterienkulturen wurde die jeweilige Zellsuspension 1:1 mit Glycerin versetzt und bei -80°C eingefroren.

5.3.9 Subklonierung in p-GEX-6P-1

Das geschnittene Gen und der geschnittene Plasmid pGEX-6P-1 wurden über Silicasäulen extrahiert. 2 µl geschnittener pGEX-6p-1, ca. 250 ng DNA-Extrakt, 3 µl Ligationspuffer (1x), 6 u DNA-Ligase und H₂O *ad* 30 µl wurden für zwei Stunden bei Raumtemperatur inkubiert (siehe auch Abbildung 9).

5.3.10 Sequenzierung

Für die Sequenzierung wurde die DNA (Vektor mit inseriertem Gen oder PCR-Amplifikat) an die Firma Genterprise in Mainz geschickt. Die korrekte Orientierung des eingefügten DNA Stücks in die Vektoren kann mir Hilfe der Sequenzdaten überprüft werden. Das eingefügte Element kann im Vektor pGEM-T Easy kann mit Primern für den T7 und SP6 Promotor sequenziert werden. Für die Sequenzierung von Vektoren der pGEX Familie gibt es spezifische Primer.

5.3.11 Vektorkarten

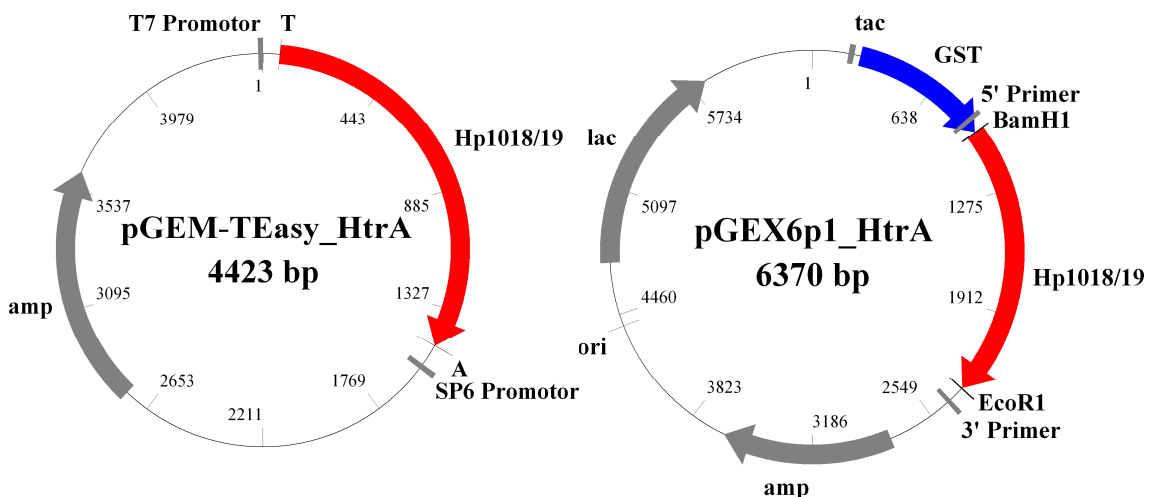


Abbildung 9. Vektorkarten der in dieser Arbeit verwendeten Konstrukte. Markiert sind die jeweiligen Hybridisierungsstellen der Sequenzierungsprimer (T7 Promotor und SP6 Promotor bzw. 5' Primer und 3' Primer) und die Stellen, an denen DNA eingefügt wurde (T und A bzw. BamH1 und EcoR1).

5.4 Expression und Reinigung der Protease HtrA

5.4.1 Expressionsbedingungen

Transformierte *Escherichia coli* BL21 Bakterien wurden in TB-Medium als Vorkultur bis zur Sättigung bei 30 °C inkubiert. TB-Medium wurde im Verhältnis 1:20 mit der Vorkultur angeimpft und bis zu einer optischen Dichte bei 300 nm von 0,6 bei 30 °C inkubiert. 0,1 mM IPTG wurde zugegeben um die Proteinexpression zu induzieren. Die Inkubation wurde drei Stunden fortgesetzt. Die Suspension wurde bei 4 °C und 4000 x g für 30 Minuten zentrifugiert und

das Zellpellet wurde in 4°C kaltem PBS mit PIT resuspendiert. Die Zellen wurden dreimal für 30 Sekunden mit dem Sonifikator lysiert, unlösliche Zellbestandteile wurden 30 Minuten bei 20.000 x g abzentrifugiert, der Überstand in flüssigem Stickstoff schockgefroren und bei -80°C aufbewahrt.

5.4.2 Reinigung

Der pGEX-6P-1 Vektor enthält ein Gen für das Enzym Glutathion-S-Transferase (GST), welche spezifisch an das Peptid Glutathion bindet. Ein in den Vektor eingefügtes Gen bildet mit dem Enzym GST nach der Expression ein Fusionsprotein, welches durch an eine stationäre Phase gebundenes Glutathion aus einem Zelllysate aufgereinigt werden kann. Die stationäre Phase wurde in diesem Fall aus Sepharose-Kügelchen gebildet, welche in einer Chromatographiesäule vorgelegt werden. Um das gereinigte Protein zu erhalten, kann entweder über eine Schnittstelle für die Prescission Protease das aus dem klonierten Gen entstandene Protein vom GST Enzym abgeschnitten oder das gesamte Fusionsprotein durch Zugabe von Glutathion in hoher Konzentration eluiert werden.

Das Bettvolumen Glutathion-Sepharose wurde gemäß den Herstelleranweisungen vorbereitet und mit einem entsprechenden Volumen des Zelllysateüberstandes über Nacht bei 4°C rotierend inkubiert.

5.4.3 Elution

Die Glutathion-Sepharose mit dem gebundenen Fusionsprotein wurde viermal mit dem zehnfachen Bettvolumen Waschpuffer gewaschen, anschließend wurde dreimal mit jeweils dem einfachen Bettvolumen an Elutionspuffer eluiert. Die Eluate wurden jeweils getrennt aufgefangen und analysiert.

5.4.4 Verdau mit Prescission-Protease

Die Glutathion-Sepharose mit dem gebundenen Fusionsprotein wurde dreimal mit dem zehnfachen Bettvolumen Waschpuffer und einmal mit dem zehnfachen Volumen an Proteasepuffer gewaschen. Der Ansatz wird mit einem einfachen

Bettvolumen an Proteasepuffer und 50U/ml Prescission-Protease für 16 Stunden bei 4 °C rotierend inkubiert.

5.4.5 Überwachung der Expression

Vor und nach der Induktion und während der Reinigung, Elution und dem Verdau wurden Proben genommen und durch SDS-PAGE mit anschließender Coomassie-Färbung analysiert. Für die SDS-PAGE (Laemmli, 1970) wurde ein zehnpromzentiges Trenngel und ein vierpromzentiges Sammelgel genutzt und eine Spannung von 50 (Sammelgel) oder 160 Volt (Trenngel) angelegt.

5.5 Inhibitionsassays der Protease HtrA

5.5.1 Reaktionsbedingungen

0,05 µg/µl der rekombinanten Protease HtrA (gelöst in HEPES-Puffer) wurden mit der gelösten Testsubstanz, 5% DMSO (v/v), 5 ng/µl rekombinatem E-Cadherin und HEPES-Puffer *ad* 20 µl für 2 bzw. 16 Stunden bei 37°C inkubiert. Die Reaktion wurde durch Zugabe von 10 µl 4x Probenpuffer und Inkubation für 5 Minuten bei 95°C gestoppt. Die Konzentration der Testsubstanz wurde für die verschiedenen Ansätze variiert. Bei den Positiv- und Negativkontrollen wurde reines DMSO ohne Testsubstanz zugegeben, bei den Negativkontrollen wurde die Proteaselösung durch HEPES-Puffer ersetzt. Die Reaktionsansätze wurden durch SDS-PAGE weiter analysiert.

5.5.2 Western-Blot

Die Proteine im SDS-Gel wurden im Wetblot-Verfahren und auf Eis für 2 Stunden bei 300 mA auf PVDF-Membranen geblottet.

5.5.3 Markierung

Die Blotmembranen wurden mit TBST-Puffer gewaschen, für 1h mit ca. 10 ml Blockingpuffer geblockt und mit 3 ml des mit Blockingpuffer verdünnten Antikörpers in Folie eingeschweißt und über Nacht bei 4°C inkubiert. Nach abermaligem Waschen wurden 10 ml Blockingpuffer mit dem verdünnten

Sekundärantikörper zugegeben und eine Stunde bei Raumtemperatur bei Raumtemperatur inkubiert.

5.5.4 Entwickeln

Die Chemolumineszenzreaktion mit der an den Sekundärantikörper gebundenen Peroxidase erfolgte nach Herstelleranweisung. Die Filme wurden in Fotokassetten für 5 bis 60 Sekunden belichtet, entwickelt und eingescannt. Alternativ wurde die Chemolumineszenz mit der FUSION-FX7 Kamera direkt digitalisiert, wobei die Dauer der Lichtakkumulation automatisch bestimmt wurde.

5.5.5 Stripping

Das als *Stripping* bezeichnete Verfahren entfernt die Antikörpermarkierungen von der Blotmembran und bereitet diese so auf eine zweite Markierung mit einem andern Primärantikörper vor.

Die Membran wurde in 100 ml Strippuffer im Wasserbad bei 52 °C für maximal 30 Minuten inkubiert und anschließend mit TBST für mindestens drei Stunden gewaschen, wobei das TBST mehrmals gewechselt wurde.

5.6 Informatische Methoden

5.6.1 PocketPicker

PocketPicker (Weisel *et al.*, 2007) ist ein Verfahren zur Vorhersage von Bindetaschen niedermolekularer Liganden auf Proteinoberflächen. Bei der Berechnung der Vorhersage wird ausschließlich auf geometrische Eigenschaften des Proteinmodells zurückgegriffen. Als Eingabe dient ein Proteinstrukturmodell mit expliziten Wasserstoffen im PDB-Format (Berman *et al.*, 2000).

Für die Berechnung wird zuerst ein dreidimensionales rechtwinkliges Gitter über das Proteinmodell gelegt, wobei das Gitter größer ist als das Proteinmodell und die einzelnen Gitterpunkte einen Abstand von einem Ångström haben. Proteinbindetaschen werden nahe der Oberfläche des Proteins erwartet, daher

werden alle Gitterpunkte entfernt, deren Abstand zum nächsten Atommittelpunkt größer als 4,5 Å ist oder die unterhalb der Proteinoberfläche liegen.

Für alle verbliebenen Gitterpunkte wird ein Vergrabenheitsindex berechnet, welcher die lokale Atomumgebung des Punkts wiedergibt. Wenn der Vergrabenheitsindex für einen Punkt einen großen Wert annimmt, befinden sich viele Proteinatome in der Nähe dieses Punkts und es wird davon ausgegangen, dass sich dieser Punkt tief in einer Proteinbindetasche befindet.

Um die Umgebung eines Gitterpunktes zu erfassen, werden für jeden Punkt 30 Vektoren mit einer Länge von 10 Å berechnet, die an dem Punkt beginnen und in 30 im Raum annähernd gleich verteilte Richtungen zeigen. Für jeden Vektor wird überprüft, ob sich mindestens ein Proteinatom im Abstand von maximal 0,9 Å zu diesem Vektor befindet. Ist dies der Fall, wird der Vergrabenheitsindex, welcher bei null beginnt, um eins erhöht und der nächste Vektor wird betrachtet. Der Vergrabenheitsindex kann so für jeden Gitterpunkt einen Wert zwischen null und 30 annehmen. Nur Gitterpunkte mit einem Vergrabenheitsindex zwischen 16 und 26 werden für das Clustering weiter verwendet. Durch dieses werden benachbarte Gitterpunkte mit ähnlichen Vergrabenheitsindices in disjunkte Gruppen zusammengefasst, die schließlich jeweils einzelne vorhergesagte Bindetaschen repräsentieren.

5.6.2 LUDI

LUDI (Böhm, 1992) ist ein Programm für das *de novo* Design von Proteinliganden. Die Berechnung läuft in drei Schritten ab. Zuerst werden ausgehend von dem Proteinstrukturmodell mögliche Interaktionspunkte in den Raum der Bindetasche abgebildet. Diese Punkte liegen in Bezug auf Winkel und Distanz zu den Atomen des Proteins günstig, um eine Wechselwirkung einzugehen. LUDI unterscheidet dabei aliphatische und aromatische Wechselwirkungen und Wasserstoffbrückenbindungen. Die bevorzugten Winkel und Distanzen wurden durch eine statistische Analyse der Cambridge Structural Database (Allen, 2002) bestimmt. Der zweite Schritt umfasst das Einpassen von Molekülfragmenten in die Bindetasche, wobei möglichst viele der

Interaktionspunkte mit passenden Atomen der Fragmente besetzt werden. Die Fragmente werden dabei aus einer Datenbank übernommen. Im dritten Schritt versucht das Programm, die Fragmente mit passenden Verbindungsfragmenten zu verbinden.

Tabelle 1. Für die Berechnung des virtuellen Liganden benutze Untermenge der LUDI-Regeln (Böhm 1992).

Regel Nr.	Atom am Enzym	Interaktionspunkt	Regel(n)
1	H-Donor N-H, O-H	H-Akzeptor A	Distanz $H...A = 1,9 \text{ \AA}$ Winkel $N/O-H...A = 180^\circ$
2	Sauerstoff C=O, R ₁ -C-R ₂	H-Donor D	Distanz $O...D = 1,9 \text{ \AA}$ Winkel $C-O...D = 120^\circ$
3	Stickstoff (unprotoniert)	H-Donor D	Distanz $N...D = 1,9 \text{ \AA}$
4	Kohlenstoff (aliphatisch)	Lipophil L	Distanz $C...L = 4 \text{ \AA}$

5.6.3 LIQUID

LIQUID (Tanrikulu *et al.*, 2007) ist eine Software zur Modellierung von dreidimensionalen Pharmakophoreigenschaften kleiner Moleküle. Diese Eigenschaften werden als trivariate Gauß-Verteilungen (Duda *et al.*, 2001) im Raum dargestellt. Der Wert dieser Verteilung an Punkt x wird mit Gleichung 1 berechnet.

$$trivG(\bar{x}) = \frac{1}{2\pi^{3/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\bar{x}-\bar{\mu})^T \Sigma^{-1}(\bar{x}-\bar{\mu})\right) \quad [1]$$

Hierbei ist μ der Mittelpunkt der Verteilung und Σ die Kovarianzmatrix der Standardabweichungen (Gleichung 2):

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}, \quad [2]$$

wobei σ_i die Standardabweichung in der i -ten Dimension ist. Weiterhin kann ein 120-dimensionaler Korrelationsvektor berechnet werden, welcher einen Fingerabdruck der Pharmakophoreigenschaften repräsentiert. Der Korrelationsvektor erlaubt es, ein LIQUID-Pharmakophormodell mit einer

Moleküldatenbank zu vergleichen, von welcher in derselben Weise berechnete Fingerabdrücke vorliegen. Der Rechenaufwand ist gering, da keine dreidimensionalen Überlagerungen von Molekülmodellen berechnet werden müssen, sondern pro Vergleich nur die Distanz zwischen zwei Vektoren.

Der erste Schritt für die Berechnung eines LIQUID-Pharmakophormodells ist die Zuweisung der Pharmakophortypen der einzelnen Nicht-Wasserstoffatome des Eingabemoleküls. Dies geschieht regelbasiert anhand der jeweiligen Atomtypen und –umgebung. LIQUID unterscheidet dabei lipophile Interaktionen und Wasserstoffbrückendonoren und –akzeptoren. LIQUID wurde entworfen, auch Überlagerungen mehrerer Moleküle als Eingabe zu akzeptieren, um ein Konsensmodell zu berechnen. Daher wird im nächsten Schritt für jedes Atom die lokale Dichte des jeweiligen Pharmakophortyps berechnet. Mit Hilfe dieser Dichtewerte werden in einem Union-Find (Galil und Italiano, 2001) Verfahren für jede der drei Pharmakophoreigenschaften Cluster berechnet, wobei deren Zentroide in der Nähe eines Dichtemaximums zu erwarten sind. Für jeden Interaktionstyp kann der Clusterradius als Parameter separat angegeben werden. Für jeden Cluster wird nun eine Hauptkomponentenanalyse (Duda *et al.*, 2001) ausgeführt. Die Clusterzentroide und die jeweiligen drei Hauptkomponenten definieren so die Erwartungswerte und die Standardabweichungen für eine trivariate Gauß-Verteilung pro Cluster und zusammen mit dem jeweiligen Pharmakophortyp einen so genannten potentiellen Pharmakophorpunkt (PPP).

Der Korrelationsvektor hat 120 Dimensionen, jeweils 20 für ein mögliches Paar an Pharmakophoreigenschaften (lipophil-lipophil, lipophil-Donor, lipophil-Akzeptor, Akzeptor-Donor, Akzeptor-Akzeptor, Donor-Donor). Für ein gegebenes Paar werden die Distanzen in 20 Partitionen von null bis 20 Ångström aufgeteilt. Der LIQUID-Korrelationsvektor enthält für jedes Pharmakophorpaar und jede Distanzpartition die Summe der jeweiligen Wahrscheinlichkeiten und wird durch Gleichung 3 berechnet:

$$CV_d^{A,B} = \frac{1}{\#Paare(A,B)} \sum_i^A \sum_j^B \frac{1}{2} \{trivG_i \cdot trivG_j\} \quad [3]$$

A und B sind Pharmakophortypen, i und j PPP und d ist das entsprechende Distanzintervall. Der Korrelationsvektor kann anschließend noch auf ein gesamtes Maximum von eins oder auf ein blockweises Maximum von eins skaliert werden. Ein Block entspricht dabei den 20 Vektorelementen für ein Interaktionspaar.

Jedes Element des Korrelationsvektors kann als die kummulative Wahrscheinlichkeit interpretiert werden, dass das betreffende Paar von Interaktionen in der gegebenen Distanz vorkommt.

5.6.4 Implementierung der Berechnung des virtuellen Liganden

Die Berechnung des strukturabgeleiteten Pharmakophormodells, dem so genannten *virtuellen Liganden* (Schüller *et al.*, 2006), erfolgt in folgenden Schritten (Abbildung 10):

1. Vorhersage des Protonierungszustands des Proteinstrukturmodells mit MOE Protonate3D (MOE, 2008) (pH 7.0, 300 K, 0,1 M Ionenkonzentration): Die Zuweisung von Interaktionstypen wie Wasserstoffbrückendonoren und -akzeptoren setzt die Kenntnis des Protonierungszustandes des betrachteten Atoms voraus. Dieser ist allerdings für Kristallstrukturmodelle, die durch Röntgenstrukturanalyse ermittelt wurden, nicht bekannt, da Protonen die gemessenen Elektronendichte kaum beeinflussen und daher ihre Position im Strukturbestimmungsverfahren nicht zugeordnet werden kann.
2. Vorhersage von Ligandenbindetaschen auf der Proteinoberfläche mit PocketPicker (Weisel *et al.*, 2007): Die Ausgabe von PocketPicker ist eine Gruppe von Punktmengen, wobei jede Menge eine mögliche Bindetasche repräsentiert. Eine solche Menge definiert also nicht nur die Aminosäuren des Proteins, die eine Bindetasche ausbilden, d.h. die Aminosäuren, die in der Umgebung der Punkte der Menge liegen. Gleichzeitig wird auch das ungefähre Volumen bestimmt, welches eine Bindetasche für einen Liganden bereitstellt. Eine oder mehrere

dieser Bindetaschenmodelle können manuell für die weitere Berechnung ausgewählt werden.

3. Bestimmen der umgebenden Aminosäuren: Ausgewählt wird für jede potentielle Bindetasche die Menge von Aminosäuremonomeren des Proteinmodells, welche mindestens ein Nicht-Wasserstoffatom mit minimaler Distanz zu einem der Punkte des Bindetaschenmodells haben.
4. Zuweisen von potentiellen Interaktionspunkten in der Bindetasche mit Hilfe einer Untermenge der LUDI-Regeln (Tabelle 1): Da das von PocketPicker definierte Volumen einer Bindetasche die Atome eines möglichen Liganden beinhaltet, können Interaktionen des Liganden mit den Atomen der umliegenden Aminosäuren als Vektoren von den Atomen in das Bindetaschenvolumen beschrieben werden. Die von LUDI übernommenen Regeln beschränken Länge und Winkel dieser Vektoren im Bezug auf Liganden und Proteinatome für eine Reihe von Interaktionstypen definieren. Da ein Bindetaschenmodell von PocketPicker aus einer diskreten Anzahl von Punkten besteht, kann im Programm effizient überprüft werden, ob ein Punkt der Bindetasche die korrekte geometrische Orientierung zu einem passendem Proteinatome hat, um eine der gegebenen Regeln zu erfüllen. Es wird nun für jedes Atom der ausgewählten Aminosäuren über alle Punkte der jeweiligen potentiellen Bindetasche iteriert und für jedes Paar überprüft, ob eine oder mehrere Regeln erfüllt werden. Dazu wird zuerst der Atomtyp des Proteinatoms bestimmt. Dies geschieht durch einen Regelsatz, der in Tabelle 2 gezeigt ist. Nun wird die Distanz d zwischen der optimalen Position eines Interaktionspunkts (Tabelle 1) und der Position des betrachteten Punktes der Bindetasche berechnet. Dazu wird der Kosinussatz (Gleichung 4) folgendermaßen eingesetzt:

$$d = \sqrt{D_{ber}^2 + D_{soll}^2 - 2 \cdot D_{ber} \cdot D_{soll} \cdot \cos(\alpha_{soll} - \alpha_{ber})} \quad [4]$$

D_{ber} und D_{soll} sind die berechneten beziehungsweise optimalen Distanzen und α_{ber} und α_{soll} die jeweiligen Winkel (Abbildung 11). D_{soll}

und α_{soll} werden aus Tabelle 1 entnommen. Im Optimalfall ist d gleich 0, allerdings wird dieser kaum eintreten. Daher wird eine Abweichung von 0,9 Ångström erlaubt. Dieser Wert ist etwas größer als der halbe maximale Abstand zwischen zwei Punkten (welcher gleich $(3^{1/2})/2$ ist) des PocketPicker Gitters und stellt sicher, dass auch im ungünstigsten Fall mindestens ein Punkt der Bindetaschenrepräsentation ausgewählt wird. Da die Regeln drei und vier (Tabelle 1) sich nur aus Distanzbeschränkungen zusammensetzen, wird in diesem Fall nur die Distanz zwischen Atom und Punkt der Bindetasche berechnet und mit dem Sollwert verglichen. Hier ist eine Toleranz von 0,5 Ångström erlaubt. Der jeweilige Punkt der Bindetasche wird mit dem jeweiligen Interaktionstyp für die weitere Verarbeitung markiert.

5. Clustering und Berechnung eines Korrelationsvektors mit LIQUID (Tanrikulu *et al.*, 2007): Mögliche Gruppierungen von Interaktionspunkten werden mit LIQUID zu Clustern zusammengefasst und ein auf diesen Clustern beruhender Korrelationsvektor wird berechnet. Dazu werden die mit Interaktionstypen markierten Punkte eines PocketPicker-Bindetaschenmodells als Pseudoatome in LIQUID eingegeben, wobei die Atomtypisierung von LIQUID übersprungen wird. Der Korrelationsvektor ermöglicht den Vergleich eines virtuellen Liganden mit einer Datenbank von potentiellen Liganden, welche auch mit LIQUID in Korrelationsvektoren kodiert sind.

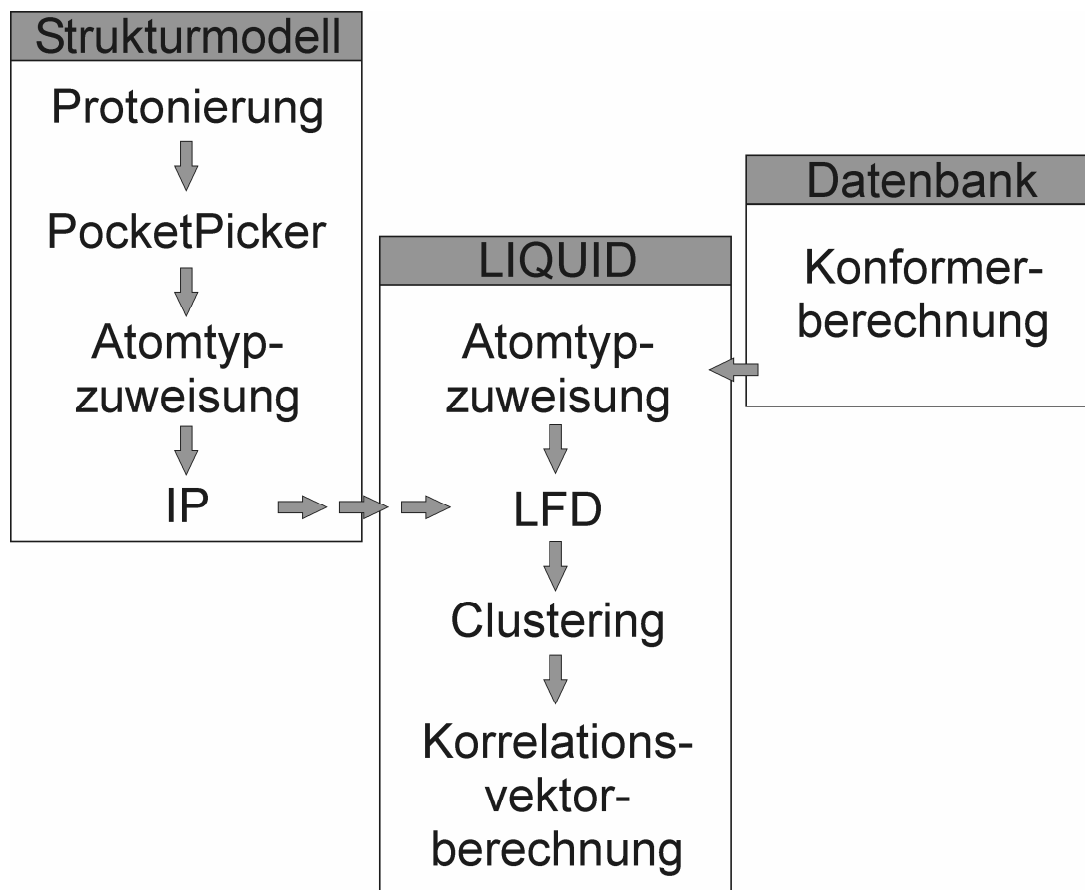


Abbildung 10. Schematische Darstellung der Berechnung des Virtuellen Liganden und der Vorbereitung der Strukturdatenbanken für das virtuelle Screening. IP steht für Interaktionspunkte und bezeichnet die Mengen der Punkte der Bindetaschenrepräsentation für die jeweiligen Interaktionstypen. LFD steht für die Berechnung der lokalen Eigenschaftsdichte in LIQUID (engl. *local feature density*).

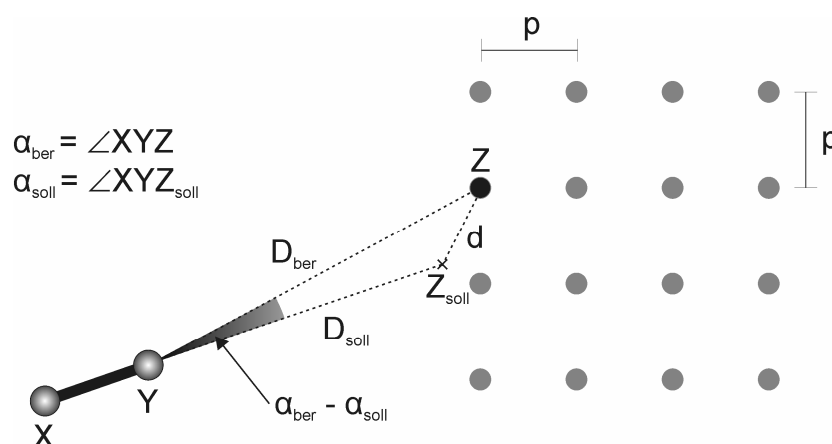


Abbildung 11. Erläuterung zu Gleichung 4. Links ist exemplarisch ein Molekül mit zwei Atomen X und Y zu sehen (z.B. eine Proteinseitenkette), rechts das *PocketPicker* Gitter, bei welchem ein Gitterabstand p vorliegt. Der Algorithmus betrachtet im vorliegenden Fall einen Punkt Z der Bindetaschenrepräsentation (schwarz) und das Proteinatom Y. Der Punkt Z_{soll} markiert die optimale Position des virtuellen Interaktionspartners des betrachteten Atoms.

Tabelle 2. Regeln zur Bestimmung von Atomtypen in Proteinen, gültig für die Seitenketten aller 20 Standardamino­säuren. Die Spalte „Nachbarschaft“ bezieht sich auf alle an das betrachtete Proteinatom kovalent gebundenen Atome, inklusive der Wasserstoffe.

Proteinatom	Nachbarschaft	zugeordneter Atomtyp
C (Tyr)	3 benachbarte Atome; maximal 1 Sauerstoff; kein Stickstoff	aromatisch
C (Phe, Trp, His)	3 benachbarte Atome; kein Sauerstoff; maximal 2 Stickstoffe	aromatisch
C (nicht aromatisch)		lipophil
N	2 Wasserstoffe	Donor
N	1 Wasserstoff	Donor
N	0 Wasserstoffe	Akzeptor
O	1 Wasserstoff	Akzeptor und Donor
O	0 Wasserstoffe	Akzeptor
S	1 Wasserstoff	-
S	0 Wasserstoffe	aromatisch

5.7 Virtuelles Screening

Für das retrospektive virtuelle Screening wurde die Strukturmodelle der Zielproteine (Tabelle 3) aus der PDB (Berman *et al.*, 2007) entnommen. Mit MOE Protonate3D (MOE, 2008) wurde der Protonierungszustand (pH 7.0, 300 K, 0,1 M Ionenkonzentration) vorhergesagt und der jeweilige virtuelle Ligand berechnet. Für jede Struktur in den Datenbanken (COBRA, UGI und MUV) wurden bis zu 250 Konformere mit MOE Conformation Import (MOE, 2008) berechnet.

Eine zehnfache *leave-group-out* Kreuzvalidierung (Kohavi, 1995) wurde durchgeführt. Dazu wurde zufällig die Hälfte der Korrelationsvektoren - wobei jeder Vektor eine Struktur repräsentiert - der jeweiligen Datenbank ausgewählt und mit dem virtuellen Liganden verglichen. Als (Un-)ähnlichkeitsmaß für zwei

zu vergleichende Vektoren x und y mit jeweils m Dimensionen diene dabei die Manhattan-Distanz (Gleichung 5; Black, 2006),

$$d_M = \sum_{i=1}^m |x_i - y_i| \quad [5],$$

die euklidische Distanz (Gleichung 6; Black, 2006),

$$d_E = \sqrt{\sum_{i=1}^m |x_i - y_i|^2} \quad [6]$$

und der Carbó Index, bzw. die Kosinusähnlichkeit (Gleichung 7; Carbó *et al.*, 1980).

$$s_C = \frac{\langle x, y \rangle}{\sqrt{\langle x, x \rangle \langle y, y \rangle}} \quad [7].$$

Die Datenbankeinträge werden dann nach den berechneten Ähnlichkeitswerten sortiert. Die Güte des virtuellen Screenings wurde mit der *Receiver operating characteristic area under curve* (ROCAUC; Gleichung 8; Fawcett, 2007)

$$ROCAUC = \frac{1}{nN} \sum_{k=2}^N F_a(k)(F_i(k) - F_i(k-1)) \quad [8]$$

und der *Boltzmann enhanced discrimination of receiver operating characteristic* (BEDROC)-Metrik (Gleichung 9; Truchon und Bayly, 2007)

$$BEDROC = \frac{\sum_{i=1}^n e^{-\alpha_i / N}}{\frac{n}{N} \left(\frac{1 - e^{-\alpha}}{e^{\alpha / N} - 1} \right)} \times \frac{R_a \sinh(\alpha / 2)}{\cosh(\alpha / 2) - \cosh(\alpha / 2 - \alpha R_a)} + \frac{1}{1 - e^{-\alpha(1-R_a)}} \quad [9]$$

bewertet. N und n sind die Anzahlen an Molekülen in der Datenbank bzw. als aktiv markierten Molekülen. $F_a(k)$ und $F_i(k)$ sind die Anzahlen an aktiven bzw. inaktiven Einträgen der Screeningdatenbank bis zum Rang k der sortierten Datenbank und R_a ist der relative Anteil an aktiven Einträgen der Datenbank. α ist ein Parameter der BEDROC Funktion und definiert den Bereich der sortierten Datenbank, der auf eine Anreicherung aktiver Substanzen untersucht wird. Ein Wert von $\alpha = 20$ wird für alle Experimente benutzt und entspricht den ersten 8% der sortierten Datenbank (Truchon und Bayly, 2007).

Tabelle 3. Proteinstrukturmodelle aus der PDB (Berman *et al.*, 2007), welche für das virtuelle Screening bzw. die Homologiemodellierung (PDB ID. 3CS0) verwendet wurden. Gegeben sind der volle Name als auch die gängige Abkürzung, soweit vorhanden.

PDB	Name	Abkürzung
1O86	Angiotensin Converting Enzyme	ACE
1CGH	Cathepsin G	
1UYL	Chaperon Hsp90	Hsp90
3PGH	Cyclooxygenase-2	COX-2
1KMV,	Dihydrofolatreduktase	DHFR
3CKH	EphA4 Rezeptor Tyrosin Kinase	Eph
1XPC	Estrogen Rezeptor-alpha	ER- α
1QKM	Estrogen Rezeptor-beta	ER- β
2BOK	Faktor Xa	fXa
1MP8	Focal Adhesion Kinase	FAK
2ZD1	HIV reverse Transkriptase	HIV-RT
1ZGY	Peroxisom-Proliferator-aktivierter Rezeptor Gamma	PPAR γ
3CS0	Protease DegP	DegP
2UZT	Proteinkinase A	PKA
2F2U	Rho-Kinase 2	
1DPO	Trypsin	
2FPZ	Tryptase	
2O8T	Urokinase Plasminogen Aktivator	uPA

Dieser Prozess wurde zehnmal wiederholt, die ROCAUC und BEDROC Werte wurden jeweils gemittelt.

Weiterhin wurde das retrospektive Screening für verschiedene Parametereinstellungen von LIQUID durchgeführt.

Für das prospektive virtuelle Screening wurden die Asinex und Specs Datenbanken (siehe Abschnitt 5.9) analog zum retrospektiven Screening vorbereitet. Der Protonierungszustand des Strukturmodells von HtrA wurde mit MOE Protonate3D vorhergesagt und der virtuelle Ligand berechnet.

5.8 Homologiemodellierung

Die Homologiemodellierung ist ein Verfahren, bei dem eine dreidimensionale Struktur eines Proteins anhand einer (oder mehrerer) durch Röntgenbeugung bekannten Proteinstruktur vorhergesagt wird. In MOE (MOE, 2008) wird dazu ein randomisierter, datenbankgestützter Ansatz genutzt. Ausgehend von einem möglichst hoch bewerteten Alignment mit einem Protein mit bekannter Struktur werden die Atomkoordinaten der im Alignment identischen Aminosäuren komplett für das Modell übernommen. Für homologe, aber nicht identische Aminosäuren werden nur die Koordinaten des Proteinrückgrates kopiert. Im nächsten Schritt werden für die Insertionen in der Modellsequenz ähnliche Sequenzen aus den hochauflösenden Kristallstrukturen der PDB gesucht. Diese dienen als mögliche Vorlagen für die, durch die Insertionen gebildeten Schleifen im Proteinmodell. Für die Seitenketten werden Rotamerbibliotheken aus PDB-Daten angelegt. Es werden nun mehrere unabhängige Zwischenmodelle erzeugt. Bei jedem werden zunächst in zufälliger Reihenfolge die Schleifen nacheinander modelliert, wobei für jede Schleife mit Hilfe einer Kontakt-Energie-Funktion eine der Vorlagen ausgewählt wird und die Atomkoordinaten übernommen werden. Alle fehlenden Seitenketten werden ausgehend von den Rotamerbibliotheken auf ähnliche Weise modelliert. Zum Schluss werden noch etwaige N- und C-terminal überstehende Peptide modelliert. Bei jedem dieser Modelle wird bewertet, in wie weit sich unpolare Aminosäureseitenketten im Proteininneren befinden und Möglichkeiten zur Wasserstoffbrückenbindung genutzt sind. Aufgrund dieser Bewertung wird das beste Zwischenmodell ausgewählt und mit einer nicht-linearen Kraftfeldoptimierungsmethode (Böhm *et al.*, 1996) eine Konformation mit lokal minimaler Energie gesucht. Diese stellt das endgültige Modell dar.

Ein Kraftfeld beschreibt die potentielle Energie eines Moleküls bzw. eines Systems von Molekülen mit folgender Gleichung (Gleichung 10):

$$U_{gesamt} = U_{Bindungslängen} + U_{Bindungswinkel} + U_{Torsionswinkel} + U_{vdW} + U_{Coulomb} \quad [10]$$

Hierbei ist U_{gesamt} die potentielle Energie, die sich aus den Teilenergien für Längen, Winkel und Torsionen der Atombindungen und den van der Waals

Kräften und elektrostatischen Interaktionen nicht gebundener Atompaaare zusammensetzt. Die Berechnung der Einzelenergien erfordert viele Parameter, die entweder empirisch bestimmt sind oder aus quantenmechanischen Rechnungen stammen. In der vorliegenden Arbeit wurde der AMBER99 Parametersatz benutzt (Wang *et al.*, 2000).

Die nicht-lineare Kraftfeldoptimierung von MOE läuft in vier Schritten ab: 1) Testen ob das Potential konvergiert, 2) Berechnung der Suchrichtung, 3) Berechnung der Schrittweite und 4) Berechnung des Potentials für die aktualisierten Atomkoordinaten; weiter mit Schritt 1. Für Schritt 2 werden nacheinander, je nach Größe des aktuellen Gradienten, die Methode der Sattelpunktsnäherung (großer Gradient), das Verfahren der konjugierten Gradienten und das inexakte Newton-Verfahren (engl. *Truncated Newton method*; sehr kleiner Gradient) benutzt (alle in Kelley, 1999).

5.9 Moleküldatenbanken

Folgende Datenbanken wurden für das prospektive bzw. retrospektive virtuelle Screening eingesetzt:

- COBRA (Version 6.1; Schneider und Schneider, 2003)
- UGI (Schüller *et al.*, 2006)
- MUV (Rohrer *et al.*, 2009)
- Asinex Gold (April 2007; Asinex Ltd, Moskau, Russland; 233420 Substanzen)
- Asinex Platinum (April 2007; Asinex Ltd, Moskau, Russland; 126584 Substanzen)
- Specs (April 2007; Specs, Delft, Niederlande; 196759 Substanzen)

Die Anzahlen an Verbindungen in den Datenbanken und die Auswahl der Wirkstoffziele für das retrospektive virtuelle Screening sind in den Tabellen 4 und 5 dargestellt.

Tabelle 4. Anzahlen von Verbindungen und ausgewählte Wirkstoffziele für die COBRA und UGI Datenbanken.

Datenbank	Version	# ¹⁾	Wirkstoffziel	Interaktion	Anzahl aktiver Verbindungen
COBRA	6.1	8140	Trypsin	Inhibitor	23
			fXa	Inhibitor	228
			DHFR	Inhibitor	64
			Tryptase	Inhibitor	17
			ACE	Inhibitor	52
			PPAR γ	Agonist	38
			uPA	Inhibitor	48
			Serinproteasen	Inhibitor	691
UGI		15840	COX-2	Inhibitor	136
			fXa	Inhibitor	1703
			Trypsin	Inhibitor	305
			Tryptase	Inhibitor	4726
			uPA	Inhibitor	1390

1) Anzahl an Verbindungen in der Datenbank

Tabelle 5. Anzahlen von Verbindungen und ausgewählte Wirkstoffziele für die MUV Datenbank.

Wirkstoffziel	Interaktion	Anzahl aktiver Verbindungen	Anzahl inaktiver Verbindungen
PKA	Inhibitor	30	15000
Rho-Kinase2	Inhibitor	30	15000
HIV RT-RNase	Inhibitor	30	15000
Eph rec. A4	Inhibitor	30	15000
HSP 90	Inhibitor	30	15000
ER- α -Coact. Bind.	Inhibitor	30	15000
ER- β -Coact. Bind.	Inhibitor	30	15000
FAK	Inhibitor	30	15000
Cathepsin G	Inhibitor	30	15000
FXIa	Inhibitor	30	15000

5.10 Software

5.10.1 GOLD

GOLD (Cole *et al.*, 2005) ist ein Programm zur Berechnung und Bewertung von Protein-Ligand Bindungsmodi (engl. *Docking*). Die Vorhersage von Bindungsposes wird durch einen genetischen Algorithmus vorgenommen, die Bewertung durch die GoldScore Bewertungsfunktion (Cole *et al.*, 2005).

5.10.2 gnuplot

Gnuplot (O'Boyle, 2008) ist ein freies, plattformunabhängiges Programm zum Erstellen von Daten- und Funktionsdiagrammen. Weiterhin bietet es die Möglichkeit, Funktionen an Daten anzupassen. Dazu wird die Methode der kleinsten Fehlerquadrate (Kelley, 1999) genutzt. In der vorliegenden Arbeit wurde eine generalisierte logistische oder sigmoide Funktion (Gleichung 11; Weisstein, 2009)

$$\text{sig}(x) = V_{\min} + \left(\frac{V_{\max} - V_{\min}}{1 + e^{k(a-x)}} \right) \quad [11]$$

an die Messdaten der Inhibitionsexperimente angepasst. V_{\max} und V_{\min} sind die obere beziehungsweise untere Asymptote und neben k und a freie Parameter für die Funktionsanpassung an den Wertebereich.

5.10.3 ImageJ

ImageJ (Abramoff *et al.*, 2004) ist ein Open-Source Bildbearbeitungsprogramm, welches besondere Funktionen für wissenschaftliche Zwecke bietet. In der vorliegenden Arbeit wird es für die densitometrische Auswertung von Blot- und SDS-Gelelektrophoresefotos genutzt.

5.10.4 Jalview

Jalview (Waterhouse *et al.*, 2009) ist ein freies Betrachtungsprogramm für (multiple) Alignments von biologischen Sequenzen. Die zusätzliche Funktion, paarweise globale Alignments zu berechnen, wurde in dieser Arbeit genutzt.

5.10.5 Java

Java (Gosling *et al.*, 2005) ist eine objektorientierte Programmiersprache. Besonderes Merkmal ist die weitestgehende Plattformunabhängigkeit von Java Programmen, da diese in einen Bytecode kompiliert werden, welcher dann in einer plattform-spezifischen Laufzeitumgebung ausgeführt werden. Diese ist für eine Vielzahl von Systemkonfigurationen kostenlos erhältlich.

Eine Vielzahl von wissenschaftlichen Programmbibliotheken wird für Java angeboten, so auch das Open-Source Projekt Chemistry Development Kit (CDK; Steinbeck *et al.*, 2003), welches Klassen für chemieinformatische Programme zur Verfügung stellt.

Die Berechnung des virtuellen Liganden wurde mit Java (Version 1.6) und dem CDK (Version 1.0) unter Linux (openSUSE Version 10.2) implementiert.

5.10.6 MOE

MOE (Molecular Operating Environment, Version 2007.09; MOE, 2008) ist ein interaktives Programm für bio- und chemieinformatische Anwendungen. In dieser Arbeit wurde es zur Homologiemodellierung von Proteinen und Vorhersage von Protonierungszuständen genutzt.

5.10.7 PyMOL

PyMOL (DeLano, 2002) ist ein frei verfügbares Programm zur Darstellung von Molekülstrukturen. Alle Darstellungen von Molekülmodellen in dieser Arbeit wurden mit PyMOL berechnet.

5.10.8 Python

Python (van Rossum, 1995) ist eine interpretierte Programmiersprache, welche objektorientierte, aspektorientierte und funktionale Programmierung unterstützt. Die Standarddistribution dieser Open-Source Programmiersprache enthält eine große Anzahl an Programmbibliotheken, wodurch eine schnelle Entwicklung verschiedenster Programme möglich ist. Auch die fehlende Notwendigkeit einer Kompilierung vor Programmausführung ermöglicht kurze Entwicklungszyklen.

Sämtliche Programme für das virtuelle Screening dieser Arbeit sind in Python (Version 2.6) implementiert.

6. Ergebnisse und Diskussion

6.1 Übersicht

Um die in Abschnitt 4.7 gesetzten Ziele zu erfüllen, wurden folgende Experimente und Arbeitsschritte durchgeführt:

1. Klonierung des Gens *hp1018/19* (Abschnitt 6.2)
2. Expression und Reinigung der Protease HtrA (Abschnitt 6.3)
3. *In-vitro* Versuche mit HtrA und E-Cadherin (Abschnitt 6.4)
4. Homologiemodellierung von HtrA (Abschnitt 6.5)
5. Retrospektive Validierung des virtuellen Ligandenmodells (Abschnitt 6.7)
6. Virtuelles Screening nach Inhibitoren von HtrA und *in-vitro* Tests mit den gefundenen Substanzen (Abschnitt 6.8)

6.2 Klonierung des Gens *hp1018/19*

Im Datenbankeintrag des Genoms von *H. pylori* 26695 (Tomb *et al.*, 1997) überlappen die Gene *hp1018* und *hp1019* (Abbildung 12). Es konnte von uns kürzlich gezeigt werden, dass diese Gene einen Leserahmen für die Protease HtrA bilden, da eine überschüssige Guanidinbase in der Datenbank verzeichnet ist (Löwer *et al.*, 2008). Das Gen, welches sich in diesem Leserahmen befindet, wird nachfolgend mit *hp1018/19* bezeichnet.

Für eine einfache Reinigung eines heterolog exprimierten Proteins ist die Konstruktion eines Fusionsproteins sinnvoll. Es wurde daher der Vektor pGEX-6P-1 für die Expression gewählt. Auf ihm ist 5' von der Stelle, an dem das Zielgen eingefügt wird (engl. *multiple cloning site*, MCS), das Enzym Glutathion-S-Transferase (GST) kodiert, welches spezifisch an das Peptid Glutathion bindet. Liegt dieses Peptid immobilisiert an der stationären Phase einer Chromatographiesäule vor, wird eine einfache Reinigung des Fusionsproteins ermöglicht.

Die Aminosäuren 1 bis 17 des Gens *hp1018/19* entsprechen mit hoher Wahrscheinlichkeit einem Signalpeptid für den *Sec*-abhängigen Proteinexport (von Heijne, 1985) und werden daher im Fusionsprotein nicht benötigt. Der entsprechende DNA-Abschnitt würde zwischen dem GST-Gen und dem

Genabschnitt der Protease liegen und wäre nicht funktional. Die Schnittstelle der Signalpeptidase könnte außerdem zu einer Abspaltung der GST Markierung während der Expression führen, was die Reinigung verhindert. Der Vorwärtsprimer für die Polymerasekettenreaktion wurde so gewählt, dass das Signalpeptid nicht mit exprimiert wird. Außerdem wurden mit den Primern Schnittstellen für die Restriktionsenzyme EcoR1 und BamH1 in 3' beziehungsweise 5' Position zum Gen eingeführt.

Abbildung 13 A zeigt eine Agarosegelelektrophorese der PCR Produkte. Die Reaktion zeigt erst bei Zugabe einer Enhancer-Lösung eine deutliche Ausbeute.

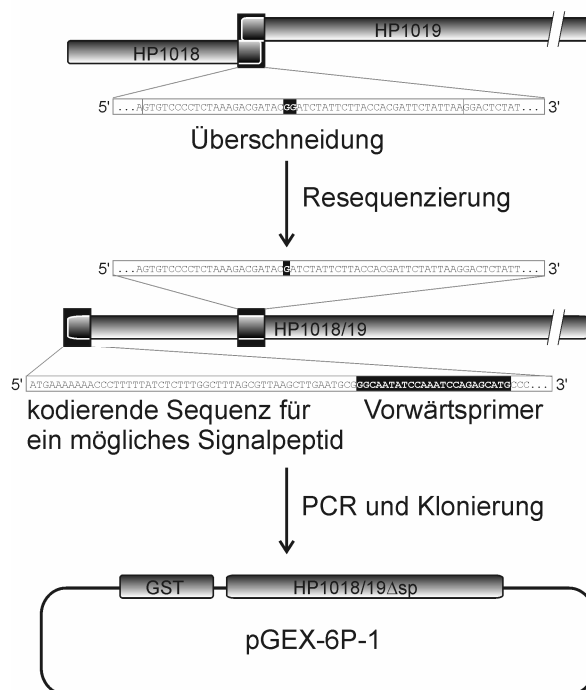


Abbildung 12. Expressionsstrategie für die Protease HtrA. Der 5' Anteil des Gens *hp1018/19* kodiert für ein mögliches Signalpeptid für den Sec-abhängigen Proteinexport. Der Vorwärtsprimer für die PCR bindet in 3' Richtung zu diesem Abschnitt damit das exprimierte Fusionsprotein kein solches Signalpeptid besitzt. Die zwischenzeitliche Klonierung in den Vektor pGEM-T Easy ist nicht dargestellt. Die Längen der Genabschnitte sind nicht maßstabsgerecht dargestellt.

Das lineare PCR-Produkt wurde mit überhängenden Adenosinbasen versehen und in den Vektor pGEM-T Easy kloniert. Der Erfolg wurde durch Blau-Weiß Selektion (Mühlhardt, 2006), Restriktionsanalyse (Abbildung 13B) und

Sequenzierung (siehe Anhang A) überprüft. Von den positiven Klonen wurden Glycerinstockkulturen angelegt.

Der leere Vektor pGEX-6P-1 wurde mit den Restriktionsenzymen EcoR1 und BamH1 verdaut, so dass das aus dem Vektor pGEM-T Easy geschnittene Genkonstrukt eingefügt werden konnte. Der Reaktionsverlauf wurde analog überprüft (Abbildung 13C), auch für diese zwei positiven Klone wurden Glycerinstockkulturen angelegt.

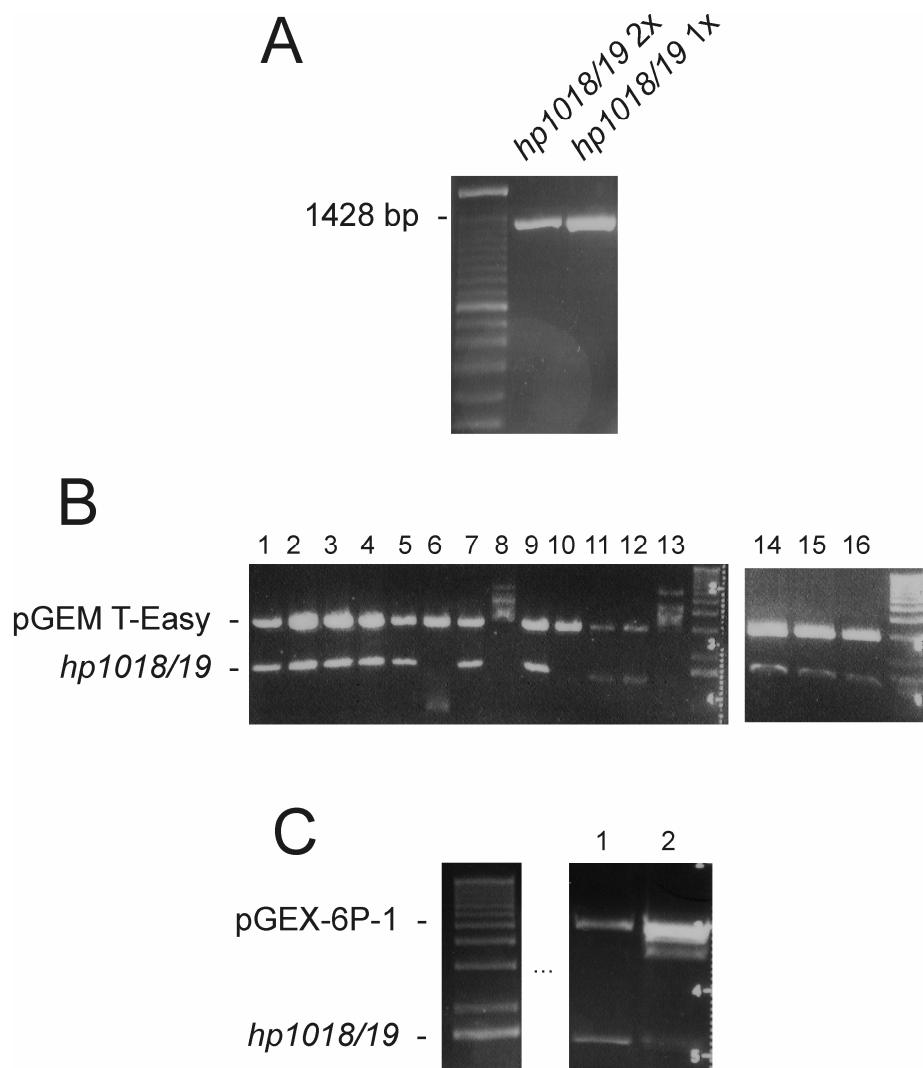


Abbildung 13. (A) Gelelektrophorese der PCR Produkte für das Gen *hp1018/19*. Die PCR-Enhancer Lösung wurde in einfacher (1X) und zweifacher (2X) Konzentration eingesetzt. Geringere Konzentrationen resultierten in keiner Reaktion (nicht dargestellt). (B) Gelelektrophorese der Restriktionsverdauung mit EcoR1 und BamH1 der präparierten pGEM-T Easy Plasmide aller Klone. Jede nummerierte Spur entspricht einem Klon. (C) Gelelektrophorese der Restriktionsverdauung mit EcoR1 und BamH1 der präparierten pGEX-6P-1 Plasmide der positiven Klone.

6.3 Expression der Protease HtrA

Das fertige Konstrukt pGEX-6P-1 *hp1018/19Δsp* kodiert für ein Fusionsprotein aus GST und der Protease HtrA, bei welcher allerdings das Signalpeptid fehlt.

Die nun folgende Reinigung des Fusionsproteins aus dem Zelllysate erfolgt über die Affinität des an HtrA fusionierten GST zur Glutathionsepharose. Für GST alleine ist diese Affinität sehr hoch (Dissoziationskonstante $k_d = 0,6 \cdot 10^{-6}$ M für ein GST-Fusionsprotein und immobilisiertes Glutathion; Forde, 2008) und spezifisch. Dies muss allerdings nicht auch für das Fusionsprotein gelten, da Teile des Proteins die Bindung erschweren könnten. Vor der eigentlichen Präzipitation (Lottspeich und Zorbas, 1998) wurde daher das optimale Verhältnis von Glutathionsepharose zu Zelllysate festgestellt (Abbildung 14A). Nach 2,5 Stunden Inkubation ist deutlich eine verstärkte Bande bei ungefähr 70 kDa zu sehen (Spuren 6 und 7 in Abbildung 14A), was den Erfolg der Expression des GST-HtrA Fusionsproteins zeigt. Weiterhin zeigt sich, dass schon geringe Mengen an Glutathionsepharose ausreichen, um eine hohe Ausbeute zu erzielen (Spuren 1 bis 5 in Abbildung 14A). Für die weiteren Versuche wurden daher Glutathionsepharose und Zelllysate im Verhältnis von zwei Teilen Sepharose zu 25 Teilen Lysate eingesetzt.

Um das Fusionsprotein von der Sepharosematrix zu lösen wurden zwei Methoden erprobt (Abbildung 14B).

Zum einen wurde das Fusionsprotein durch Glutathion dreimal hintereinander eluiert (Spuren 5, 6 und 7 in Abbildung 14B). Die Ausbeute ist im ersten Eluat am höchsten (Spur 5) und beim Waschen der Matrix entsteht nur ein geringer Verlust (Spur 4).

Andererseits wurde ein Verdau mit Precision Protease eingesetzt, um die GST Markierung zu entfernen. Eine entsprechende Schnittstelle ist auf dem pGEX-6P-1 Plasmid kodiert. Die GST Markierung bleibt an die Sepharose gebunden (Spur 14 in Abbildung 14B) und HtrA ist mit einem Molekulargewicht von ungefähr 45 kDa frei in Lösung (Spur 12). Ein geringer Anteil HtrA bleibt allerdings in der Sepharose als Verlust zurück (Spur 13 und 14).

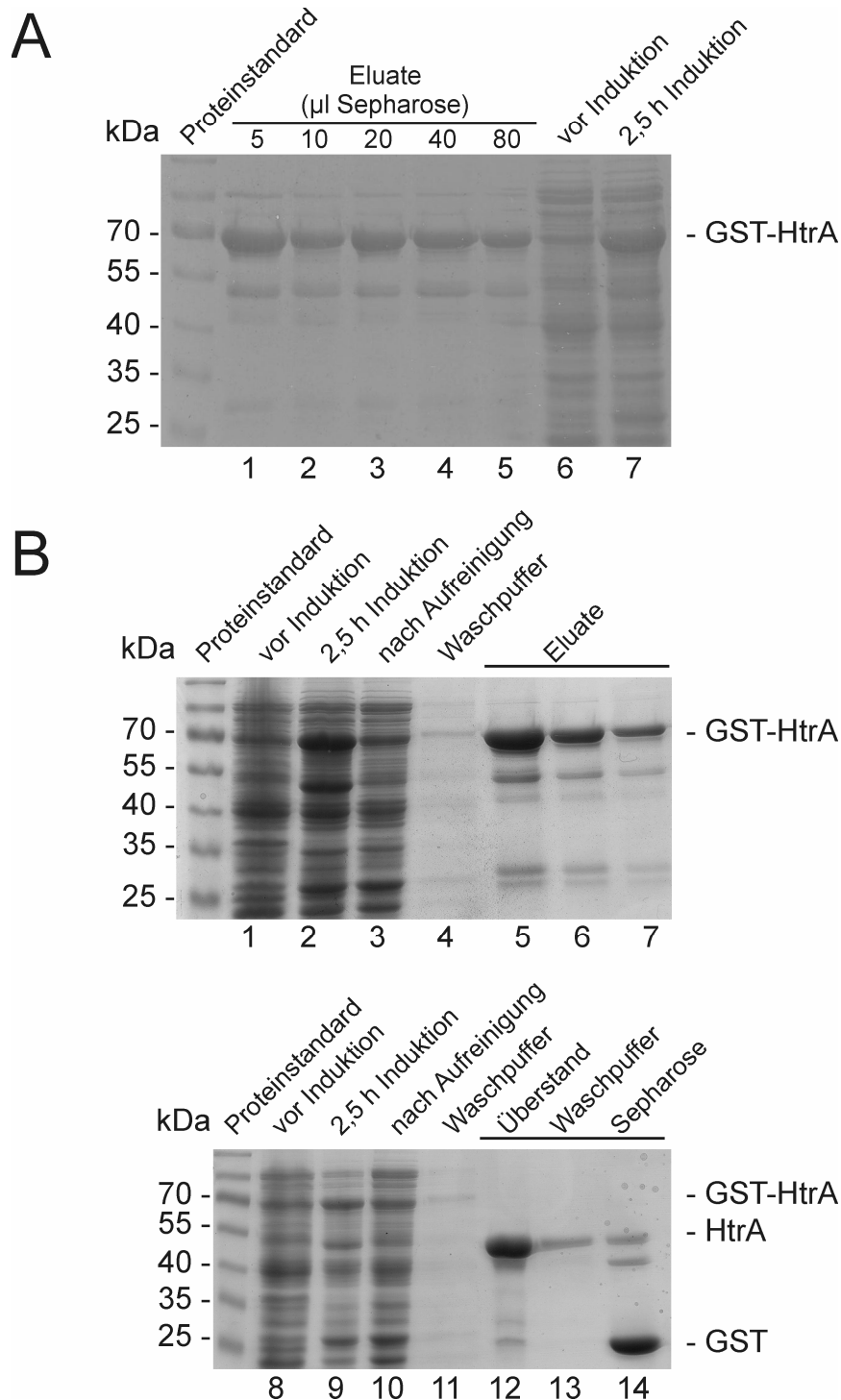


Abbildung 14. (A) SDS-Gelelektrophorese zur Bestimmung des besten Lysat/Sepharose Verhältnis. Es werden die gebundenen Mengen an Fusionsprotein bei verschiedenen Mengen von Sepharose gezeigt. Die Lysatmenge ist konstant (250 µl). Die beiden rechten Spuren zeigen den Expressionsverlauf vor und nach 2,5 Stunden Induktion. Deutlich zu sehen ist die Zunahme einer Bande bei ca. 70k Da. (B) Vergleich von Elution (oben) und Proteaseverdau (unten) zum Lösen von HtrA von der Chromatografiematrix. Zu sehen sind jeweils Zelllysate vor und nach der Induktion, und der Überstand und Waschlösung des Puldownassays. Rechts wird zusätzlich die Ausbeute von drei aufeinander folgenden Elutionsschritten gezeigt, links die Ausbeute des Proteaseverdaus (Überstand), der Verlust beim Waschen der Matrix (Waschlösung) und die an der Matrix verbleibende Menge GST (Sepharose).

Es zeigt sich, dass sowohl der Verdau mit Prescission Protease als auch die Elution mit reduziertem Glutathion mit hoher Ausbeute erfolgt (vgl. jeweils Spur 2 mit 5-7 und 9 mit 12 und 14). Weil nicht ausgeschlossen werden kann, dass die GST Markierung die HtrA Aktivität beeinflusst, wurde für den Pulldown-Assay dem Proteaseverdau der Vorzug gegeben.

Die so gereinigte Protease stand nun für die Experimente mit einem rekombinanten Substrat (Abschnitt 6.4) und mit den zu findenden Inhibitoren (Abschnitt 6.8.2) zur Verfügung.

6.4 HtrA schneidet E-Cadherin

Nach der Reinigung des überexprimierten HtrA wurde die proteolytische Aktivität mit einem rekombinanten Substrat getestet. Dieses Substrat ist ein Fusionsprotein aus der extrazellulären Domäne des humanen E-Cadherins und mit humanem Immunglobulin IgG₁ (Abbildung 15A). Die theoretische molekulare Masse des reifen Proteins beträgt 87,7 kDa, allerdings ist durch posttranslationale Modifikationen bei einer SDS-Gelelektrophorese eine Bande bei ungefähr 120 kDa zu beobachten (Abbildung 15B, Spuren 1 und 2). Das reife Protein beginnt mit der Aminosäure Asp 155.

Die rekombinante Protease HtrA schneidet das E-Cadherinsubstrat (Abbildung 15B, Spuren 3 und 4). Der verwendete Antikörper wurde mit den Aminosäuren 600-707 (Abbildung 15A) der extrazellulären Domäne von E-Cadherin als Antigen hergestellt und weiterhin erscheint nach der Proteolyse eine deutliche eine Bande zwischen 70 und 55 kDa. Es kann angenommen werden, dass eine HtrA-Schnittstelle nahe dem C-terminalen Ende des E-Cadherin existiert, da das längste mögliche E-Cadherinfragment des rekombinanten Substrats (Asp155 – Ile707) ein theoretisches Molekulargewicht von 60,35 kDa besitzt (Abbildung 15A). Innerhalb dieses Fragments scheint es noch weitere Schnittstellen zu geben, da auch kleinere Fragmente nachgewiesen wurden. Die geringere Stärke der entsprechenden Banden könnte allerdings ein Hinweis sein, dass diese weiteren Schnittstellen weniger spezifisch sind.

HtrA ist außerdem eine Serinprotease, da die S221A Mutante von HtrA (zur Verfügung gestellt von Christiane Weydig, Paul-Ehrlich Institut Langen) die

Das Protein DegP ist in drei Domänen aufgeteilt (Abbildung 16), eine N-terminale Proteasedomäne und zwei PDZ Domänen. Die PDZ Domänen vermitteln die Bildung von 3, 6, 12 und 24-meren (Krojer *et al.*, 2002; Krojer *et al.*, 2008). Diese unterscheiden sich in ihrer Aktivität im Bezug auf die doppelte Rolle von DegP als Chaperon und/oder Protease. So ist z.B. die Hexamerform (welche dem Strukturmodell 1KY9 entspricht) wahrscheinlich sowohl proteolytisch als auch als Chaperon inaktiv (Krojer *et al.*, 2002).

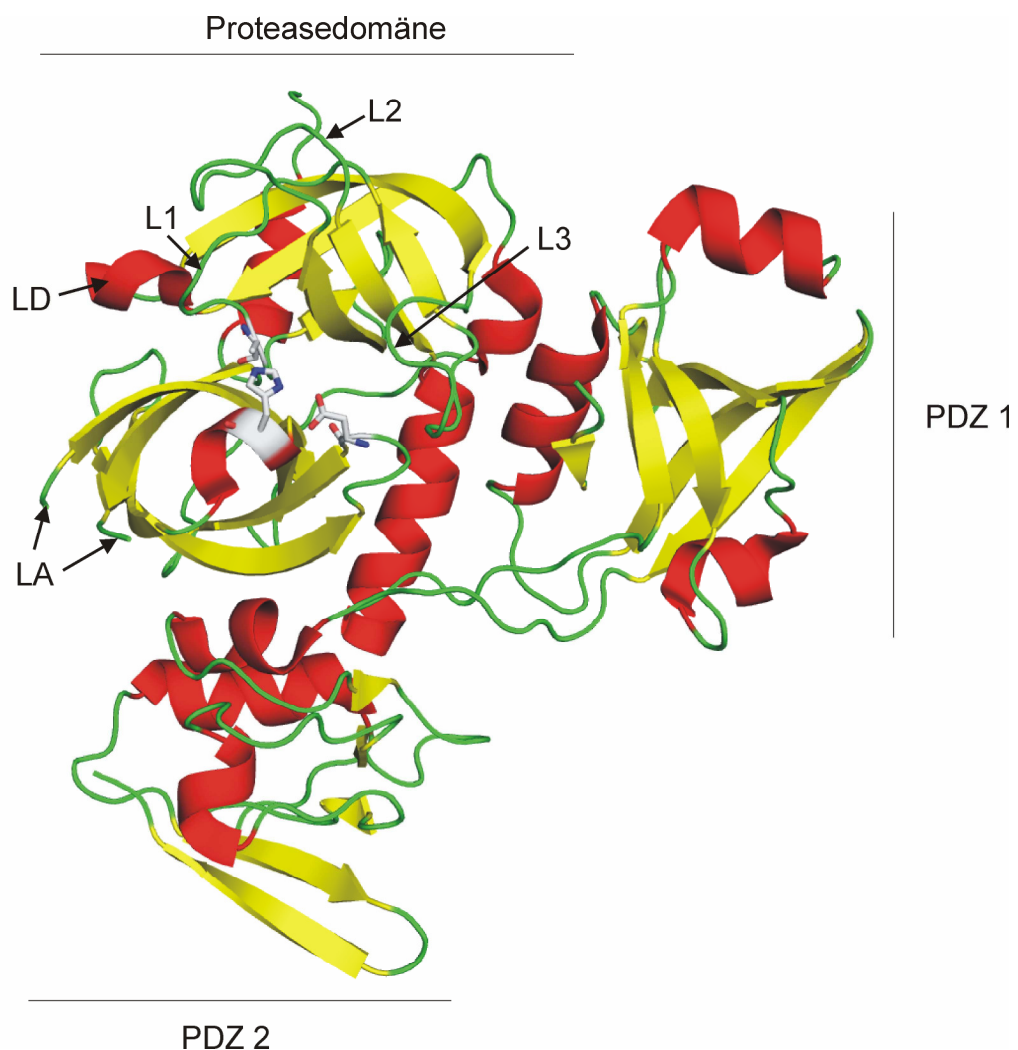


Abbildung 16. Strukturmodell der Protease DegP (PDB ID 3CS0). Die Benennung der Oberflächenschleifen der Proteasedomäne erfolgt nach Perona und Craik (Perona und Craik, 1995). Die Aminosäuren der katalytischen Triade (His105, Asp135 und Ser210) sind als Stabmodelle und grau dargestellt. Die Schliefe LA zeichnet sich durch eine hohe Flexibilität aus und ist daher im gezeigten Modell nicht vorhanden. Zu sehen sind nur die flankierenden Aminosäuren Thr35 und Glu82.

Ein Vergleich der Sequenzen von HtrA, DegP und Trypsin (Abbildung 17) zeigt, dass HtrA wahrscheinlich eine ähnliche Faltblattstruktur wie DegP hat, da die Sequenzabschnitte in diesen Bereichen konserviert sind. Neben der Homologie der katalytischen Triade rechtfertigt auch dies die Wahl der Vorlage.

Auch die Peptidschleifen LA, LD, L1, L2 und L3 von HtrA zeigen eine hohe Sequenzähnlichkeit zu den entsprechenden Bereichen von DegP (Abbildung 17), was auf eine ähnliche Substratspezifität schließen lässt.

Die Sequenzähnlichkeit nimmt im Bereich der PDZ Domänen ab (51,23% Sequenzidentität bis einschließlich Gly269; 29,81% Sequenzähnlichkeit ab Lys270). Dies ist für die weiteren Untersuchungen von Vorteil, da ein genaueres Modell der im Bereich der Proteasedomäne zu erwarten ist.

Das gezeigte Alignment diene auch als Grundlage für eine Homologie-modellierung von HtrA mit MOE.

Ein Ramachandrandiagramm (Ramachandran und Sasiskharan, 1968) wurde mit MOE berechnet, um die Qualität des fertigen Homologiemodells einzuschätzen (Abbildung 18). Insgesamt 15 Aminosäuren nehmen ungewöhnliche ϕ und ψ Winkel (Ramachandran und Sasiskharan, 1968) ein. Acht dieser Aminosäuren liegen in den C-terminalen PDZ Domänen und sind für die Ableitung eines Pharmakophormodells irrelevant, da dieses auf die Bindetaschen der Proteasedomäne abzielen soll. Von den weiteren sieben Aminosäuren ist nur Gly183 dem aktiven Zentrum zugewandt.

Abbildung 19A zeigt das vermutete aktive Zentrum des Homologiemodells von HtrA mit den umgebenden Peptidschleifen und den mit PocketPicker vorhergesagten Bindetaschen. Gly183 befindet sich in Schleife LD nahe der Tasche 38. Diese vorhergesagte Bindetasche hat lediglich ein Volumen* von ca. 15 \AA^3 . Im Vergleich zu Trypsin nimmt diese vorhergesagte Bindetasche den Platz der Bindetasche S2' ein (Perona und Craik, 1995).

Die vorhergesagte Bindetasche 12 wird vom C-terminalen β -Faltblatt der Schleife LA (Abbildung 17) und den Schleifen L1 und L2 gebildet und repräsentiert die Spezifitätstasche S1 (Perona und Craik, 1995). Die

* Da PocketPicker ein regelmäßiges Gitter mit 1 \AA Abstand für die Berechnungen nutzt, lassen sich die Volumina der vorhergesagten Bindetaschen durch Zählen der Mitglieder der einzelnen Cluster abschätzen.

vorhergesagte Bindetasche 11 ist durch die Schleifen LC und L3 und die α -Helix, welche His105 beinhaltet, begrenzt. Die Position dieser potentiellen Bindetasche gibt Anlass für die Annahme, dass sie die S3 Tasche repräsentieren könnte (Perona und Craik, 1995). Abbildung 20A zeigt die an den potentiellen Bindetaschen beteiligten Aminosäuren schematisch mit einem Proteinsubstrat.

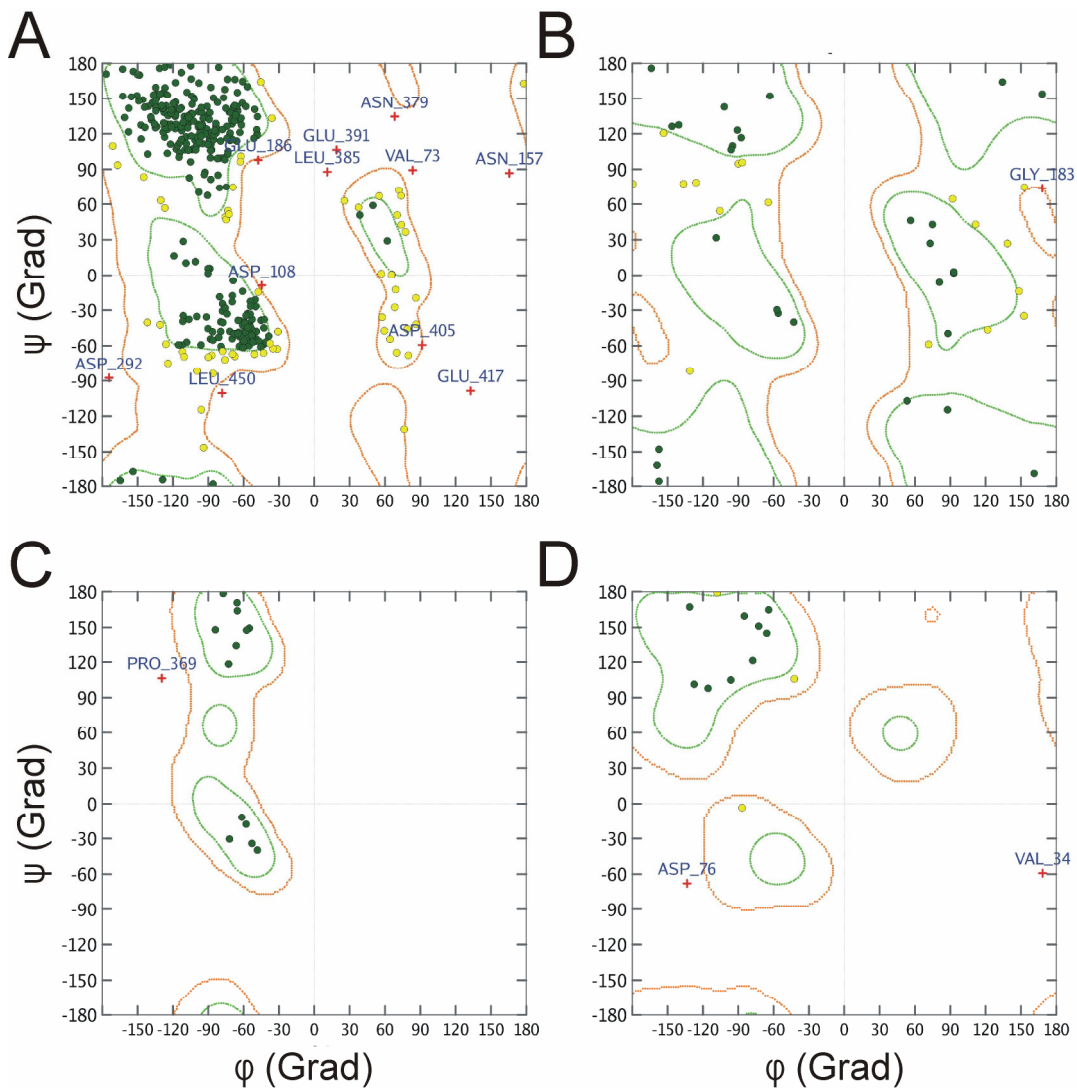


Abbildung 18. Ramachandrandiagramm für das Homologiemodell von HtrA. Grün umrandete Bereiche sind bevorzugt, orange umrandete Bereiche erlaubt (Ramachandran und Sasiskharan, 1968). Jeder grüne bzw. gelbe Punkt repräsentiert eine Aminosäure, Aminosäuren mit ungewöhnlichen Torsionswinkeln sind als rotes Kreuz dargestellt und beschriftet. (A) Allgemeine Torsionswinkel. (B) Torsionswinkel der Glycine. (C) Torsionswinkel der Proline. (D) Torsionswinkel der Aminosäuren vor einem Prolin.

Die Volumina der Taschen 11 und 12 werden durch das N-terminal von L2 liegende β -Faltblatt getrennt. An dieses konservierte Faltblatt (Abbildung 17) könnte das Peptidrückgrat des Substrats binden, was für andere Serinproteasen beobachtet wurde (Hedstrom, 2002). Dabei liegen die beiden Peptidketten antiparallel zu einander.

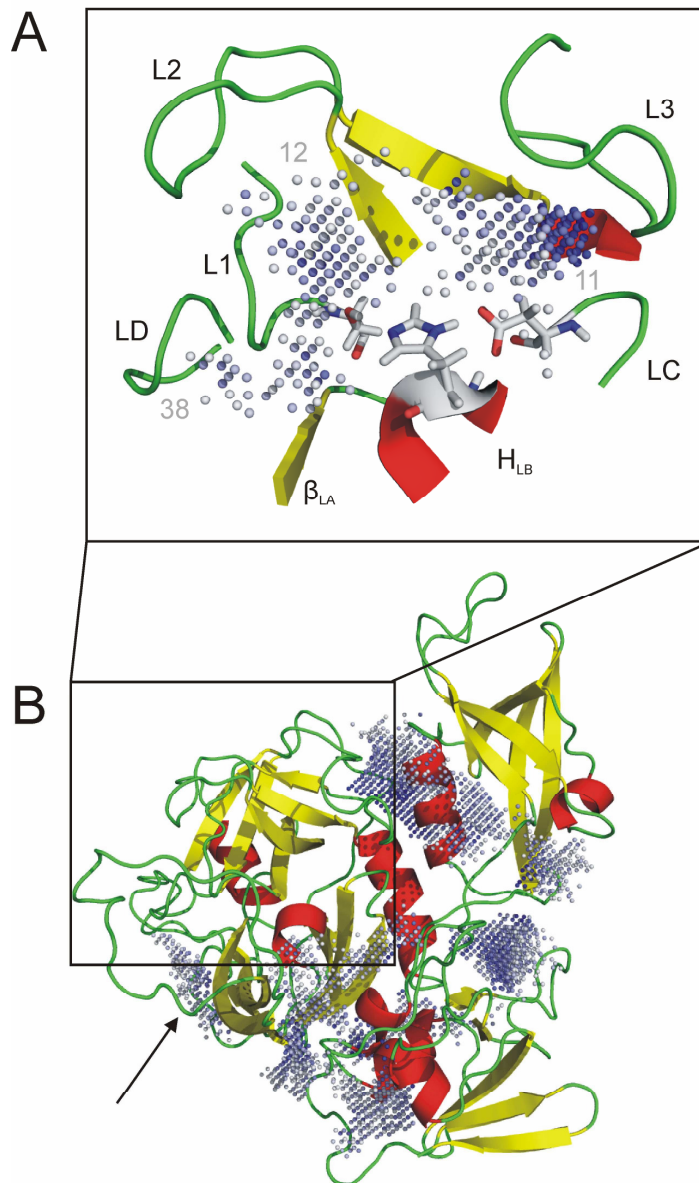


Abbildung 19. (A) Cartoondarstellung des aktiven Zentrums von HtrA und Kugelgittermodell der vorhergesagten Bindetaschen 11, 12 und 38. Die katalytische Triade ist als stabiles Modell dargestellt, die Wasserstoffatome entsprechen dem vorhergesagtem Protonierungszustand. Die Benennung der umgebenden Schleifen erfolgt gemäß Perona und Craik (Perona und Craik, 1995). β_{LA} bezeichnet das β -Faltblatt N-terminal zu Schleife LA und H_{LB} die α -Helix am C-terminalen Ende von Schleife LB. (B) Cartoondarstellung des Homologiemodells von HtrA und Kugelgittermodell der vorhergesagten Bindetaschen 1 bis 10. Tasche 8 (Pfeil) liegt im Gegensatz zu den anderen Taschen nicht zwischen den Proteindomänen.

Die Ausgabe von PocketPicker beinhaltet die vorhergesagten Bindetaschen nach Größe sortiert. Es wurde beobachtet, dass die größte vorhergesagte Bindetasche oftmals auch das Volumen der Ligandenbindung beinhaltet (Weisel *et al.*, 2007). Dies ist auch für die meisten Strukturen der retrospektiven Experimente wahr (Tabelle 3), allerdings liegen im Fall des Homologiemodells von HtrA die zehn größten Taschen nicht in der Nähe des aktiven Zentrums (Abbildung 19B). Diese Taschen liegen mit Ausnahme der Tasche 8 zwischen den drei globulären Domänen von HtrA und könnten bei der Multimerisierung von HtrA (Löwer *et al.*, 2008) beteiligt sein. Auch könnten *induced-fit* Effekte (Koshland, 1958) zu einer Vergrößerung der Taschen 11, 12 oder 38 führen, was auch die geringe Genauigkeit der Bindetaschenvorhersage erklären könnte. Weiterhin muss man bedenken, dass Proteasen mit Proteinen sehr große Substrate haben und nur die Seitenketten fest in tiefen Taschen gebunden werden. Der Algorithmus von PocketPicker ist gerade auf das Erkennen von tiefen Taschen ausgelegt, folglich wird die tatsächliche Größe der Bindetasche wahrscheinlich unterschätzt. Betrachtet man die Taschen 11, 12 und 38 als Einheit, ergibt sich ein kombiniertes Volumen von ungefähr 256 \AA^3 , wodurch diese Tasche die viertgrößte wäre.

Die Konformation des aktiven Zentrums einer Serinprotease ist wichtig für die proteolytische Aktivität, da ein Protonentransfer vom Serin zum Histidin und die Ladungsstabilisierung durch das Aspartat den nucleophilen Angriff des Seins auf die Carbonylgruppe des Substrats ermöglicht (Hedstrom, 2002; Abbildung 21).

Ohne Substrat existiert ein Netzwerk aus Wasserstoffbrückenbindungen im aktiven Zentrum (Abbildung 20B). Im Homologiemodell von HtrA sind die Längen der Wasserstoffbrücken zwischen Ser221, His116 und Asp147 (Abbildung 20B) mit denen von zwei Kristallstrukturen von Chymotrypsin vergleichbar (Tabelle 6). Die Winkel weichen allerdings deutlich von 180° als Optimum ab, was auf schwächere Wasserstoffbrücken hindeutet. Hier ist möglicherweise ein Fehler bei der Berechnung des Homologiemodells unterlaufen oder *induced-fit* Effekte sorgen bei Substratbindung für eine

korrekte Konformation. Auf der anderen Seite können natürlich nicht beide Wasserstoffbrücken vom Histidin zum Aspartat einen Winkel von 180° haben. Auch die Aminosäure Pro218 könnte einen ungünstigen Einfluss auf die vorhergesagte Konformation haben, da Prolin als Iminosäure eine andere bevorzugte Rückgratkonformation hat als andere Aminosäuren. An den homologen Positionen von DegP und Trypsin ist kein Prolin zu finden (Abbildung 17).

Es ist nicht bekannt ob diese Beobachtungen tatsächlich Fehler darstellen. Allerdings kann die intrinsische Unschärfe (*Fuzziness*; Tanrikulu et al. 2007) des virtuellen Liganden möglicherweise diese potentiellen Fehler kompensieren.

Bei Histidin wird im Gleichgewicht das $\text{N}\epsilon\text{-H}$ Tautomer bevorzugt, im aktiven Zentrum von Serinproteasen wird allerdings die $\text{N}\delta\text{1-H}$ Form angenommen (Abbildung 20C; Hedstrom, 2002). Diese Protonierungsform wurde von MOE nicht vorhergesagt und wurde manuell festgelegt um möglichst nahe an einer proteolytisch aktiven Konformation zu sein.

Das fertige Homologiemodell konnte nun als Grundlage für das virtuelle Screening nach Inhibitoren der Protease HtrA dienen (Abschnitt 6.8). Zuvor sollte allerdings überprüft werden, ob das neue strukturbasierte Pharmakophormodell für ein virtuelles Screening einsetzbar ist (Abschnitt 6.7). Abschnitt 6.6 gibt einen Überblick über die Berechnungen.

Tabelle 6. Winkel und Längen der Wasserstoffbrücken zwischen den Aminosäureseitenketten der katalytischen Triade von HtrA und Chymotrypsin. Die Längen beziehen sich auf die Heteroatome.

	Homologiemodell HtrA	Chymotrypsin (PDB ID 1ACB)	Chymotrypsin Acyl-Enzym- Zwischenprodukt (PDB ID 3GCT)
\angle Ser-OH... $\text{N}\epsilon\text{2-His}$	$156,62^\circ$		
Länge Ser-OH... $\text{N}\epsilon\text{2-His}$	3,2 Å	3,1 Å	3,0 Å
\angle Asp-O δ1 ... $\text{HN}\delta\text{1-His}$	$159,62^\circ$		
Länge Asp-O δ1 ... $\text{HN}\delta\text{1-His}$	2,64 Å	2,9 Å	2,6 Å
\angle Asp-O δ2 ... $\text{HN}\delta\text{1-His}$	$127,52^\circ$		
Länge Asp-O δ2 ... $\text{HN}\delta\text{1-His}$	3,32 Å	3,4 Å	3,6 Å

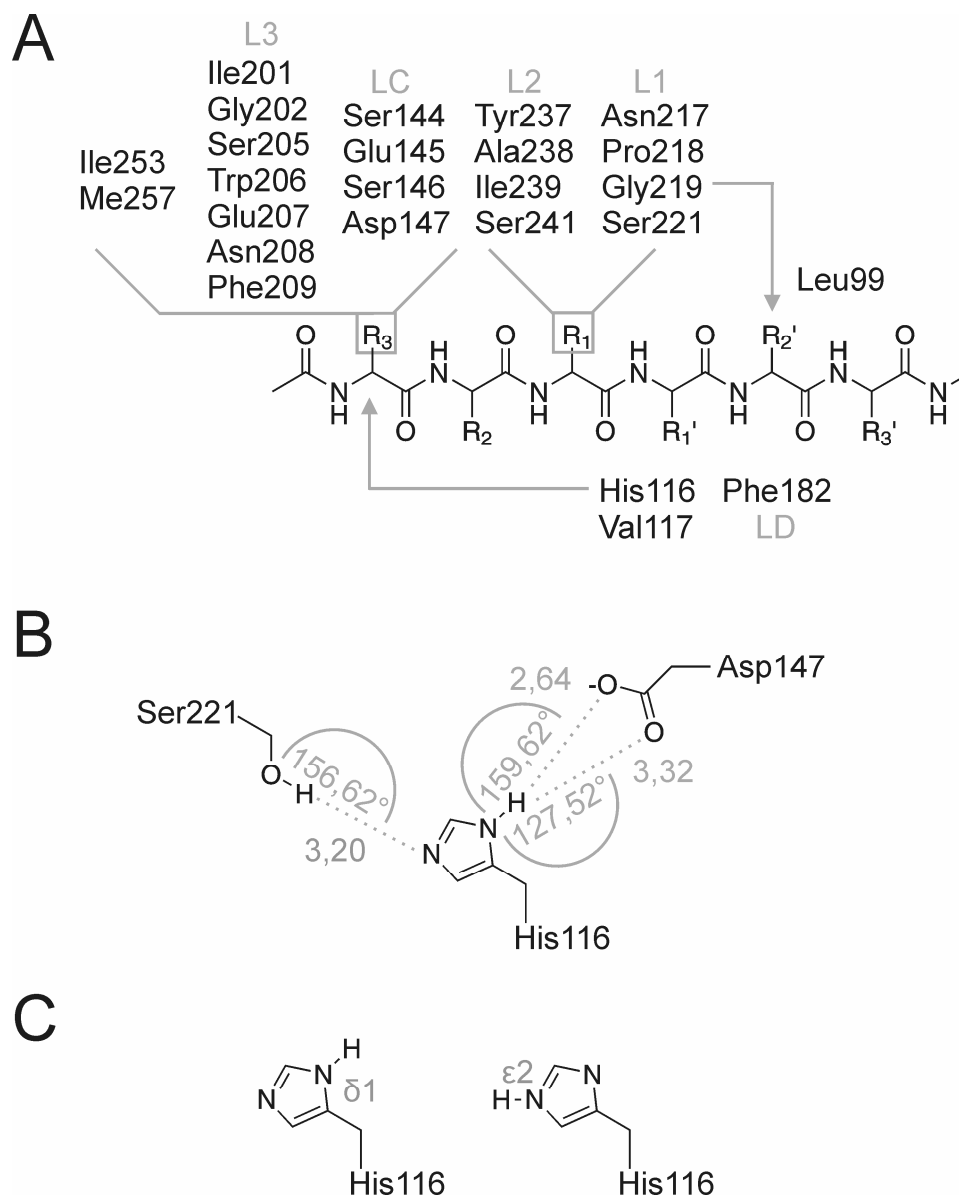


Abbildung 20. (A) An der Ausbildung der Bindetaschen beteiligte Peptidschleifen (grau) und Aminosäuren (schwarz) vom Homologiemodell von HtrA zusammen mit einem hypothetischen Proteinsubstrat. His116 liegt an der Öffnung der möglichen Bindetasche S3 (Pfeil) und Gly219 ist neben Leu99, Val117 und Phe182 an der potentiellen S2' Tasche beteiligt (Pfeil). Die Abbildung wurde von Hedstrom übernommen (Hedstrom, 2002) und an HtrA angepasst. Vergleiche auch Abbildung 17 für das zugrunde liegende Alignment. (B) Geometrie des aktiven Zentrums des Homologiemodells von HtrA. Die Distanzen beziehen sich auf die Heteroatome und sind in Ångström angegeben. Die gestrichelten Linien stellen mögliche Wasserstoffbrückenbindungen dar (Darstellung ist nicht vollständig). (C) Vergleich von zwei möglichen Protonierungszuständen von Histidin. Das N δ 1-H Tautomer wird links gezeigt, das N ϵ 2-H Tautomer rechts.

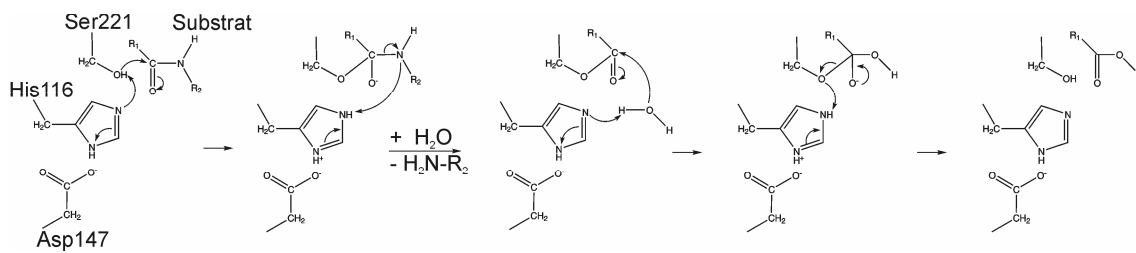


Abbildung 21. Möglicher Reaktionsmechanismus für die Serinprotease HtrA. An der katalytischen Triade beteiligt sind die Seitenketten der Aminosäuren His116, Asp147 und Ser221. Im ersten Schritt erfolgt ein nucleophiler Angriff des Hydroxysauerstoffs des Serins auf den Kohlenstoff der zu spaltenden Peptidbindung. Es entsteht ein tetraedischer Übergangszustand. Der Stickstoffteil der Bindung mit dem C-terminalen Teil des Substrats verlässt als Abgangsgruppe diesen Zustand. Ein Wassermolekül hydrolysiert das entstehende Acyl-Enzym-Zwischenprodukt über einen weiteren tetraedischen Übergangszustand. Die Aspartatseitenkette stabilisiert die temporäre positive Ladung am Histidin (nach Hedstrom, 2002).

6.6 Das virtuelle Ligandenmodell

Im Folgenden wird die Berechnung des neuen strukturabgeleiteten Pharmakophormodells für HtrA kurz dargestellt (die Details finden sich in den Abschnitten 5.6.1 bis 5.6.4). Das Modell kann als ein *virtueller Ligand* (Schüller *et al.*, 2006) verstanden werden, da es mögliche Wechselwirkungen mit dem Zielprotein wiedergibt, aber keiner tatsächlichen Molekülstruktur entspricht.

Der erste Schritt ist die Vorhersage von Bindetaschen mit *PocketPicker* (Weisel *et al.*, 2007). Die Ausgabe von *PocketPicker* ist nicht nur die Position der Bindetasche(n) sondern jeweils auch das Volumen und die Form (Abbildung 22). Dieser Teilraum wird als bevorzugter Ort der Ligandenbindung angesehen. Der zweite Schritt ist die Abbildung potentieller Interaktionen von den die Bindetasche(n) umliegenden Aminosäuren in den vorher definierten Raum (Abbildung 23). Es werden drei Interaktionstypen berücksichtigt: Lipophile Interaktionen, Wasserstoffbrückendonoren und -akzeptoren.

Als letztes werden die projizierten potentiellen Interaktionen durch ein Clustering zusammengefasst und in räumliche Wahrscheinlichkeitsverteilungen umgewandelt (Abbildung 24), was mit dem Programm *LIQUID* (Tanrikulu *et al.*, 2007) geschieht. Die Darstellung als Wahrscheinlichkeitsverteilung lässt eine gewisse Unschärfe oder *fuzziness* (Tanrikulu *et al.*, 2007) in das Modell mit

einfließen, welche zum einen Fehler der Proteinmodelle (Davis *et al.*, 2008) und zum anderen ansonsten nicht beachtete *induced-fit* Effekte kompensieren soll. LIQUID berechnet aus den Wahrscheinlichkeitsverteilungen einen Korrelationsvektor (Tanrikulu *et al.*, 2007), welcher für das eigentliche virtuelle Screening herangezogen wird. Der berechnete Korrelationsvektor eines virtuellen Liganden muss lediglich mit allen Vektoren einer vorbereiteten virtuellen Substanzbibliothek verglichen werden, was effizient möglich ist. Ein Element eines einzelnen Korrelationsvektors repräsentiert die Wahrscheinlichkeit, dass das ein Paar von potentiellen Interaktionen in einer gewissen Distanz zueinander vorkommt (Abbildung 25), sei es in der durch den virtuellen Liganden beschriebenen Bindetasche oder in einem Molekülmodell der Substanzbibliothek.

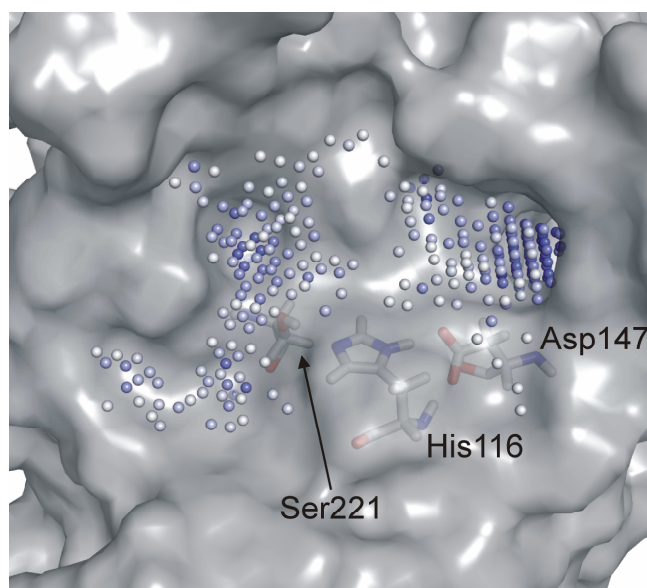


Abbildung 22. Bindetaschenvorhersage für HtrA als Beispiel für den Berechnungs-ablauf des virtuellen Liganden. Die Kugelgruppen repräsentieren die Bindetaschen und sind nach der Vergrabenheit eingefärbt, wobei eine dunklere Färbung eine höhere Vergrabenheit andeutet. Unter der Oberfläche sind die Aminosäuren der katalytischen Triade gezeigt (His116, Asp147 und Ser221).

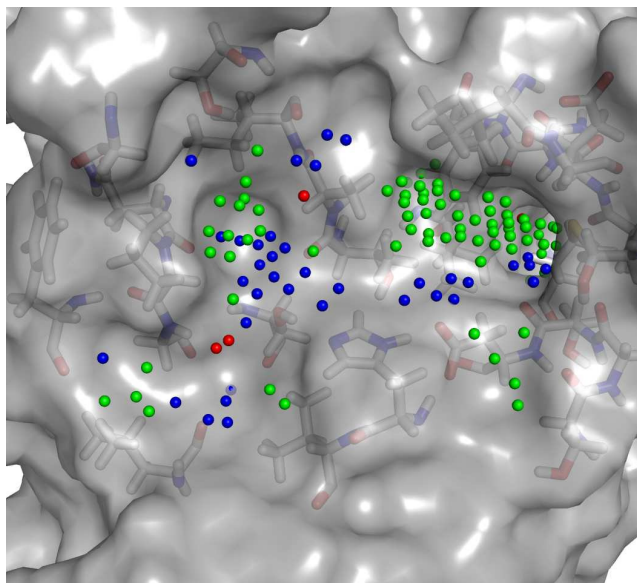


Abbildung 23. Zweiter Schritt der Berechnung des virtuellen Liganden. Die Punkte der Bindetaschenvorhersage sind nach den potentiellen Interaktionsabbildungen eingefärbt (grün: lipophil, rot: Wasserstoffbrückenakzeptor, blau: Wasserstoffbrückendonor). Die gezeigten Aminosäuren sind vom Algorithmus als Ausgangspunkt für die Abbildung herangezogen worden.

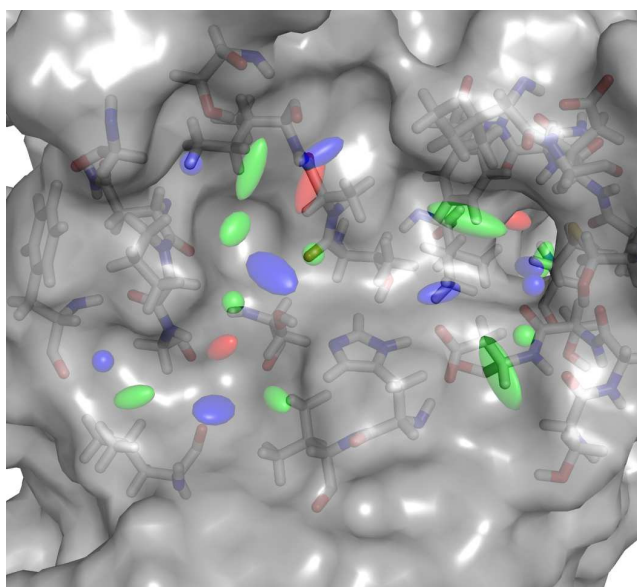


Abbildung 24. Letzter Schritt der Berechnung des virtuellen Liganden. Die von LIQUID berechneten Wahrscheinlichkeitsverteilungen sind als Ellipsoide dargestellt und gemäß dem jeweiligen Interaktionstyp eingefärbt (grün: lipophil, rot: Wasserstoffbrückenakzeptor, blau: Wasserstoffbrückendonor).

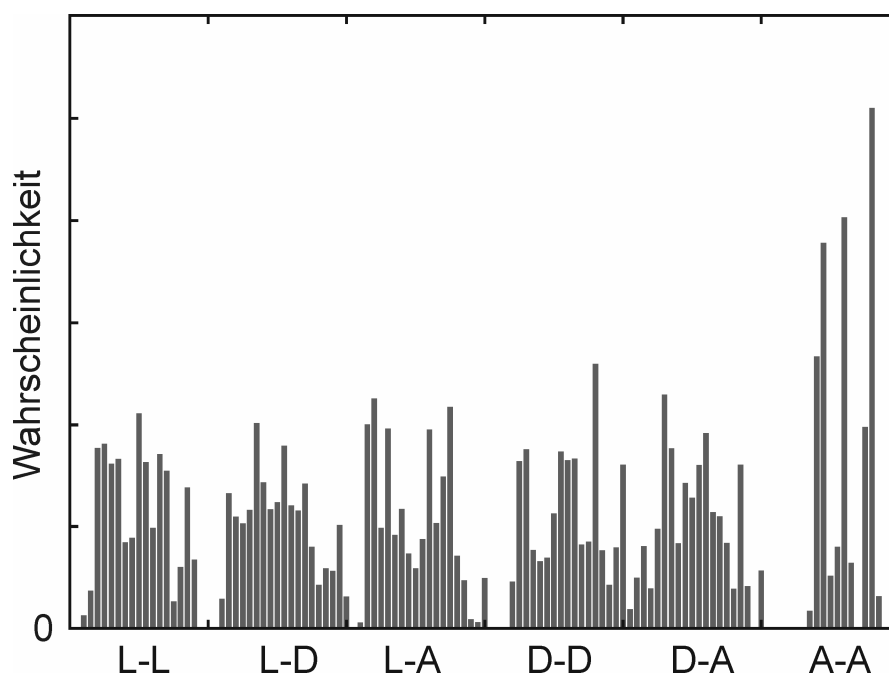


Abbildung 25. Beispiel eines LIQUID Korrelationsvektors. Für jedes Paar von Interaktionstypen (L: lipophil, D: Wasserstoffbrückendonor, A: Wasserstoffbrückenakzeptor) ist eine Wahrscheinlichkeitsverteilung über eine Distanz von 1-20 Å gegeben.

6.7 Retrospektive Validierung

6.7.1 Screeningdatenbanken

Drei verschiedene Datenbanken wurden für das retrospektive virtuelle Screening eingesetzt.

Die COBRA Datenbank (Schneider und Schneider, 2003) ist eine Datenbank von biologisch aktiven Substanzen, Leitstrukturen und bekannten Wirkstoffen.

Die UGI Datenbank (Schüller *et al.*, 2006) beinhaltet Produkte von Dreikomponenten-UGI-Reaktionen (Ugi *et al.*, 1959), deren inhibitorische Aktivität gegen die Serinproteasen Trypsin, Tryptase, uPA und Faktor Xa gemessen wurde.

Diese beiden Datenbanken können als Beispiele für eine prospektive Anwendung der vorgestellten Methode für das virtuelle Screening angesehen werden; die COBRA Datenbank ist eine Sammlung von in der jüngeren Literatur beschriebenen pharmakologisch aktiven Substanzen und Leitstrukturen, während die UGI Datenbank eine fokussierte Bibliothek für Serinproteaseinhibitoren darstellt. Es kann daher angenommen werden, dass

beide Datenbanken nur einen Teil des chemischen Raumes abdecken, da sie entweder durch die physikochemischen Eigenschaften oder das molekulare Grundgerüst (hier: das UGI Grundgerüst) in ihrer Diversität potentiell beschränkt sind. Auf der einen Seite ist dies ein realistisches Szenario, da zum Beispiel die Zugänglichkeit für die chemische Synthese oder das notwendige Anwenden von Filtern vor dem eigentlichen Screening (Seifert und Lang, 2008) den Umfang und die Abdeckung des chemischen Raums der Datenbank beschränken.

Um diese Beschränkung und die mögliche Verzerrung der Ergebnisse des retrospektiven Screenings zu umgehen, wurde auch die so genannte *maximum unbiased validation set* (MUV) Datenbank (Rohrer und Baumann, 2009) in die retrospektiven Experimente miteinbezogen.

Acht Wirkstoffziele wurden für das retrospektive virtuelle Screening mit der COBRA Datenbank ausgewählt: Angiotensin-konvertierendes Enzym (engl. *angiotensin-converting enzyme*, ACE), Cyclooxygenase 2 (COX-2), Dihydrofolatreduktase (DHFR), Faktor Xa (fXa), Peroxisom-Proliferator-aktivierter Rezeptor Gamma (PPAR γ), Urokinase-Typ Plasminogen Aktivator (uOA), Trypsin und Tryptase. Vier dieser Proteine sind Serinproteasen (fXa, uPA, Trypsin, Tryptase), für welche experimentelle Inhibitionsdaten in der UGI Datenbank vorliegen.

Die MUV Datenbank beinhaltet für jedes Wirkstoffziel 30 aktive Verbindungen (also Inhibitoren bzw. Agonisten, Modulatoren oder Potentiatoren) und 15000 inaktive Verbindungen. Die Zusammenstellung dieser Datensätze betont jeweils niedrige Ähnlichkeit der aktiven Verbindungen für ein Wirkstoffziel und eine hohe Ähnlichkeit der inaktiven Verbindungen zu den aktiven. Dies verhindert potentiell eine Verfälschung der Ergebnisse eines virtuellen Screenings (Rohrer und Baumann, 2009). Zusätzlich ist für die inaktiven Verbindungen experimentell bestätigt, dass sie keine Wirkung auf das gegebene Wirkstoffziel haben, was im Fall der COBRA Datenbank nicht gegeben ist. Die Datensätze sind für 17 Wirkstoffziele vorhanden, allerdings ist das eigentliche Einsatzgebiet der MUV Datenbank die Validierung von ligandenbasierten Methoden des virtuellen Screenings. Daher sind nicht für alle Wirkstoffziele

Proteinkristallstrukturen verfügbar, welche die Grundlage für die vorgestellte strukturbasierte Methode sind. Insgesamt wurden nur zehn Wirkstoffziele für das Screening benutzt.

6.7.2 Ergebnisse des retrospektiven Screenings

Die Ausgabe des Screenings ist die nach Ähnlichkeit zur Vorlage sortierte Screeningdatenbank. Ziel ist es, wichtige Eigenschaften der Rezeptor-Ligand Interaktion so zu kodieren, das viele aktive Substanzen auf die vorderen Ränge sortiert werden beziehungsweise im vorderen Teil der Datenbank angereichert werden. Dies wird mit Hilfe der *receiver operating characteristic area under curve* (ROCAUC; Fawcett, 2006) und *boltzmann enhanced discrimination of receiver operating characteristic* (BEDROC; Truchon und Bayly, 2007) Metriken untersucht.

Die Tabellen 7, 8 und 9 zeigen die Ergebnisse des retrospektiven virtuellen Screenings für die COBRA, UGI und MUV Datenbanken. Die Ergebnisse sind über alle untersuchten Parameterkombinationen des Algorithmus gemittelt. Eine Diskussion der Auswirkungen der verschiedenen Parameterwerte findet sich im weiteren Text, die gesamten Ergebnisse der *leave-group-out* Validierung ohne Mittelung im Anhang B.

Tabelle 7. Ergebnisse des retrospektiven virtuellen Screenings für die COBRA Datenbank. Die ROCAUC und BEDROC Werte sind über alle Parameterkombinationen gemittelt.

Enzym	PDB ID	Taschen(n) ¹⁾	ROCAUC (σ)	BEDROC (σ)
ACE	1O86	1	0,38 (0,07)	0,00 (0,00)
COX-2	3PGH	1	0,52 (0,12)	0,05 (0,06)
	3PGH	1 (Teil) ²⁾	0,62 (0,16)	0,23 (0,10)
DHFR	1KMV	1	0,65 (0,05)	0,15 (0,08)
fXa	2BOK	1	0,52 (0,14)	0,04 (0,03)
	2BOK	1, 6, 13, 17	0,74 (0,09)	0,14 (0,07)
PPAR γ	1ZGY	1	0,53 (0,04)	0,05 (0,02)
Trypsin	1DPO	1	0,56 (0,13)	0,03 (0,02)
Tryptase	2FPZ	2, 4, 17	0,72 (0,06)	0,18 (0,06)
UPA	2O8T	1	0,67 (0,07)	0,17 (0,08)

1) Die Nummerierung entspricht der Rheinform der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) Bindetaschenvorhersage wurde manuell bearbeitet (siehe Text).

Tabelle 8. Ergebnisse des retrospektiven virtuellen Screenings für die UGI Datenbank. Die ROCAUC und BEDROC Werte sind über alle Parameterkombinationen gemittelt.

Enzym	PDB ID	Taschen(n) ¹⁾	ROCAUC (σ)	BEDROC (σ)
fXa	2BOK	1, 6, 13, 17	0,58 (0,03)	0,15 (0,02)
Trypsin	1DPO	1	0,51 (0,04)	0,08 (0,02)
Tryptase	2FPZ	2, 4, 17	0,65 (0,06)	0,51 (0,11)
UPA	2O8T	1	0,61 (0,06)	0,19 (0,06)

1) Die Nummerierung entspricht der Rheifolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

Tabelle 9. Ergebnisse des retrospektiven virtuellen Screenings für die MUV Datenbank. Die ROCAUC und BEDROC Werte sind über alle Parameterkombinationen gemittelt.

Enzym	PDB ID	Taschen(n) ¹⁾	ROCAUC (σ)	BEDROC (σ)
Cathepsin G	1CGH	1	0,54 (0,09)	0,07 (0,05)
Eph	3CKH	2	0,52 (0,03)	0,05 (0,02)
ER-a	1XPC	1	0,60 (0,08)	0,13 (0,10)
ER-b	1QKM	1	0,55 (0,05)	0,07 (0,02)
FAK	1MP8	1	0,61 (0,03)	0,09 (0,02)
FXIa	1ZSJ	1	0,39 (0,09)	0,01 (0,01)
HIV-RT	2ZD1	1	0,52 (0,11)	0,08 (0,03)
Hsp90	1UYL	1	0,64 (0,05)	0,13 (0,06)
PKA	2UZT	1	0,51 (0,08)	0,05 (0,02)
Rho-Kinase 2	2F2U	1	0,50 (0,07)	0,04 (0,02)

1) Die Nummerierung entspricht der Rheifolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

In einer *receiver operating characteristic* (ROC) Kurve zur Auswertung eines virtuellen Screeningexperiments wird der Anteil an aktiven Verbindungen gegen den Anteil an inaktiven Verbindungen dargestellt für jeden Rang der sortierten Datenbank. Die Fläche unter der Kurve (ROCAUC) kann dann als Maß für die Güte des Experiments dienen. Abbildung 26 verdeutlicht den Zusammenhang zwischen dem ROCAUC Wert und der Verteilung von aktiven Substanzen in einer sortierten Datenbank.

Die benutzte strukturbasierte Methode für das virtuelle Screening ermöglicht es, eine signifikante Anreicherung von aktiven Verbindungen für die Mehrzahl der untersuchten Wirkstoffziele zu erreichen, was durch einen ROCAUC Wert von mehr als 0,5 angedeutet wird (Tabellen 7, 8 und 9).

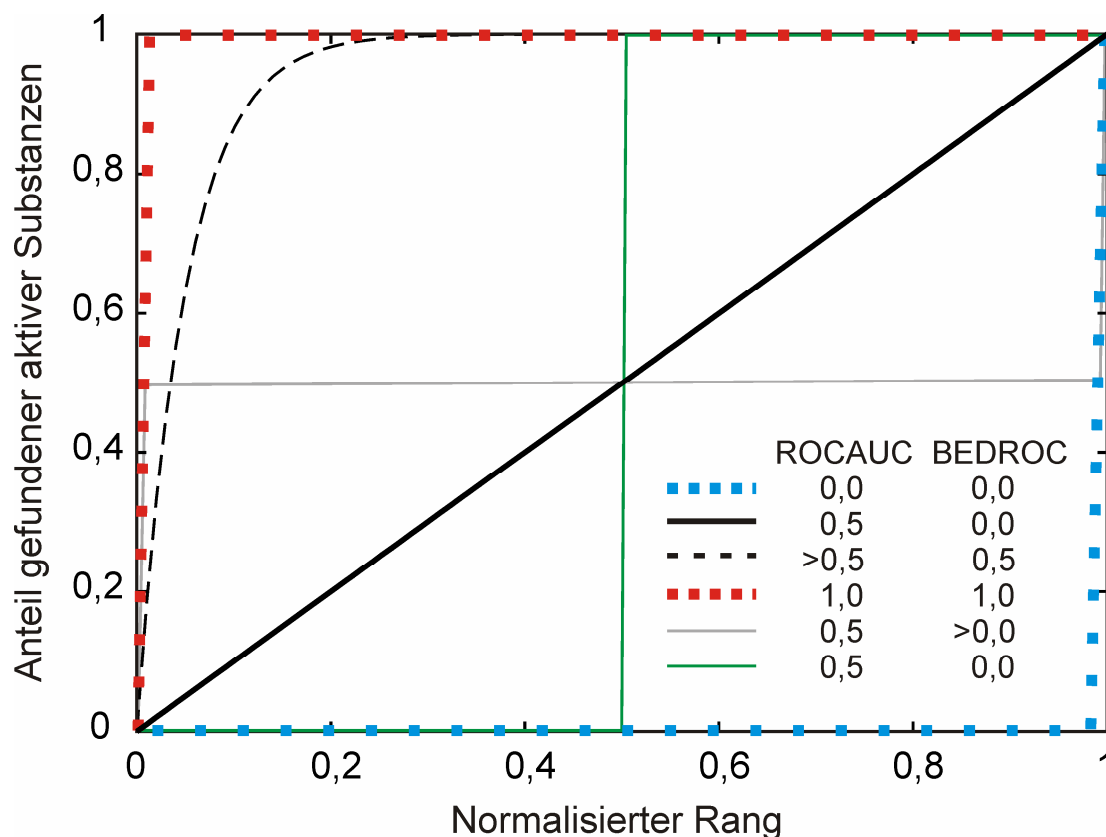


Abbildung 26. Mögliche Verteilungen von aktiven Substanzen einer Moleküldatenbank nach einem virtuellen Screening und der Zusammenhang mit den ROCAUC und BEDROC ($\alpha = 20$) Werten. Schwarz: uniforme Verteilung der aktiven Moleküle; rot: optimale Verteilung; blau: schlechteste Verteilung. Die gestrichelte Kurve entspricht einer exponentiellen Anreicherung. Die grüne und graue Kurve haben denselben ROCAUC Wert aber einen anderen BEDROC Wert, was die Nutzung beider Metriken für die Bewertung der Experimente motiviert.

Die ROCAUC Metrik bewertet die gesamte Sortierung, während der BEDROC Wert die Anreicherung von aktiven Verbindungen am Beginn der sortierten Datenbank bewertet. Dies ist im Hinblick auf das *early recognition*-Problem (Truchon und Bayly, 2007) wichtig, da in der Regel in prospektiven Anwendungen nur ein kleiner Teil der vorderen Ränge der sortierten Datenbank weiter bearbeitet wird. Ein BEDROC Wert von 0,5 oder mehr deutet eine exponentielle Anreicherung im vorderen Bereich der Sortierung an, ein Wert von null eine lineare Anreicherung (Abbildung 26). Die BEDROC Funktion besitzt einen Parameter α , welcher den betrachteten Bereich definiert. Der in dieser Arbeit benutzte Wert von $\alpha = 20$ entspricht den ersten 8% der sortierten Datenbank.

In den vorliegenden Ergebnissen (Tabellen 6, 7 und 8) findet sich nur in einem Fall eine exponentielle Anreicherung (Tryptase). Allerdings stellt diese den

Optimalfall da, der sicherlich nicht mit einer Methode in allen denkbaren Fällen erreicht werden kann. Weiterhin erzeugt möglicherweise die hohe Zahl an aktiven Verbindungen für das Enzym Tryptase in der UGI Datenbank eine Überschätzung des tatsächlichen BEDROC Werts (Truchon und Bayly, 2007).

Die retrospektiven Screening Experimente mit der UGI Datenbank zeigen eine Anreicherung für alle Targets. Der Grund dafür könnte sein, dass die Möglichkeit verschiedene Bindetaschen eines Proteins in ein Modell zu integrieren, speziell die Bindungseigenschaften von Serinproteasen mit ihren Substraten repräsentiert. Dies ist besonders im Hinblick auf die prospektive Anwendung der Methode auf die Serinprotease HtrA eine positive Eigenschaft. Auf der anderen Seite basiert die Einteilung der Verbindungen in der UGI Datenbank in aktiv und inaktiv nur auf den gemessenen Aktivitätswerten auf den gegebenen Proteasen. Sobald eine Substanz auch nur eine geringe Aktivität zeigt, wird er in den vorliegenden Experimenten als aktiv gewertet, daher ist es möglich, dass viele nur schwach aktive Substanzen durch das virtuelle Screening angereichert werden. In einer realistischen Anwendung ist dies natürlich kein wünschenswertes Ergebnis. Im Gegensatz dazu enthalten die COBRA und MUV Datenbanken nur Verbindungen, die gut an das jeweilige Zielprotein binden.

6.7.3 Einfluss der Bindetaschenvorhersage auf die Screeningergebnisse

Keine Anreicherung wurde bei den Screenings für ACE und fXla Inhibitoren erreicht. Eine eingehende Betrachtung der vorhergesagten Bindetasche für die ACE Struktur 1O86, welche für das Screening verwendet wurde, zeigt jedoch, dass die Größe der Ligandenbindetasche für diese Struktur überschätzt wird (Abbildung 27). 3525 Punkte des Gitters von PocketPicker werden zu der Bindetasche gezählt, die den Liganden enthält, was einem Volumen von ungefähr 3525 \AA^3 entspricht. Auch für andere Strukturen von ACE können solche großen Werte beobachtet werden (1O8A: 3238 \AA^3 , 1UZE: 3572 \AA^3). Die Fehleinschätzung der Größe der Bindetasche könnte das schlechte Ergebnis beim virtuellen Screening erklären.

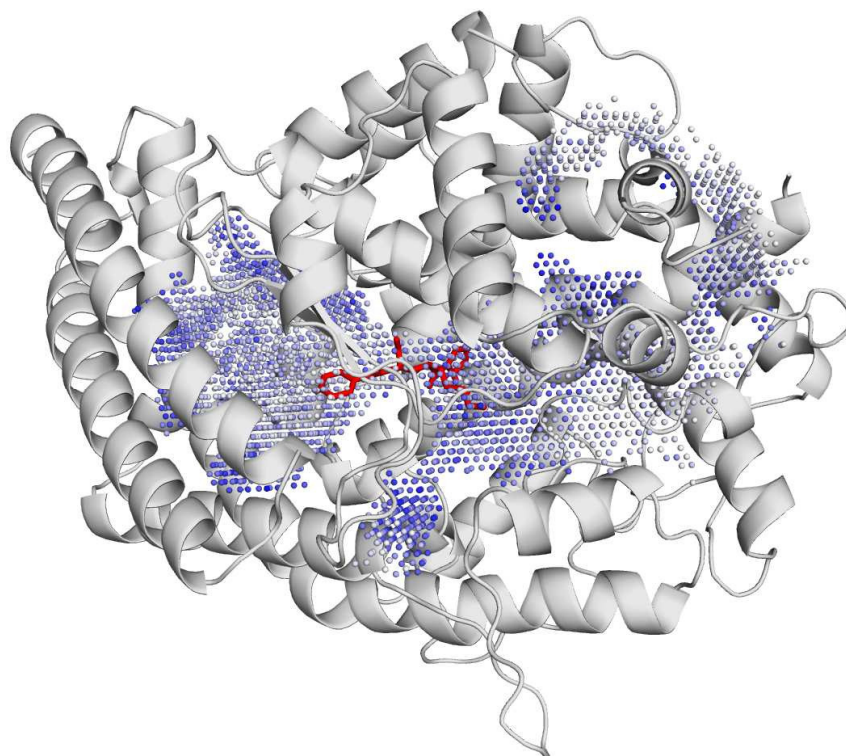


Abbildung 27. Modell des Angiotensin konvertierenden Enzyms (PDB Struktur 1O86) mit der größten vorhergesagten Bindetasche (blaue Kugeln) und gebundenem Ligand (Lisinopril, rot).

Für COX-2 und die entsprechende Kristallstruktur 3PGH konnte eine ähnliche falsche Vorhersage beobachtet werden. Eine Bindetasche mit einer Größe von ungefähr 2332 \AA^3 wurde vorhergesagt und der entsprechende mittlere ROCAUC Wert war nur geringfügig größer als 0,5. Wurde die vorhergesagte Bindetasche allerdings manuell auf die Proportionen des gebundenen Liganden der Kristallstruktur (Flurbiprofen; Kurumbail *et al.*, 1996) verkleinert, stieg der mittlere ROCAUC Wert auf 0,62 (Tabelle 7), wobei das Volumen nur noch ungefähr 225 \AA^3 beträgt (Abbildung 28). Dieses Ergebnis hebt die Wichtigkeit einer korrekten Vorhersage der Bindetasche und deren Geometrie hervor.

Die in dieser Arbeit vorgestellte Methode für strukturbasiertes virtuelles Screening erlaubt es, mehrere Bindetaschen in ein Modell zu integrieren. Dies ist unter anderem für Serinproteasen sinnvoll, welche mehrere Bindetaschen für die Seitenketten des Proteinsubstrats besitzen (Hedstrom, 2002). In der retrospektiven Studie wird dies am Beispiel von Fxator Xa und der Kristallstruktur 2BOK gezeigt (Tabelle 7). Das Hinzufügen der vorhergesagten

Taschen 6, 13 und 17 zum Modell (Abbildung 29) resultierte in einem Anstieg des mittleren ROCAUC Werts von 0,52 auf 0,74. Diese Taschen umgeben das aktive Zentrum dieser Protease und sind daher Kandidaten für die Substratbindung und -erkennung (Hedstrom, 2002). Ein Hinzufügen von weiteren Bindetaschen zum Pharmakophormodell der Faktor XIa Struktur 1ZSJ und der Trypsin Struktur 1DPO könnte auch hier eine Verbesserung der Screening Ergebnisse bewirken.

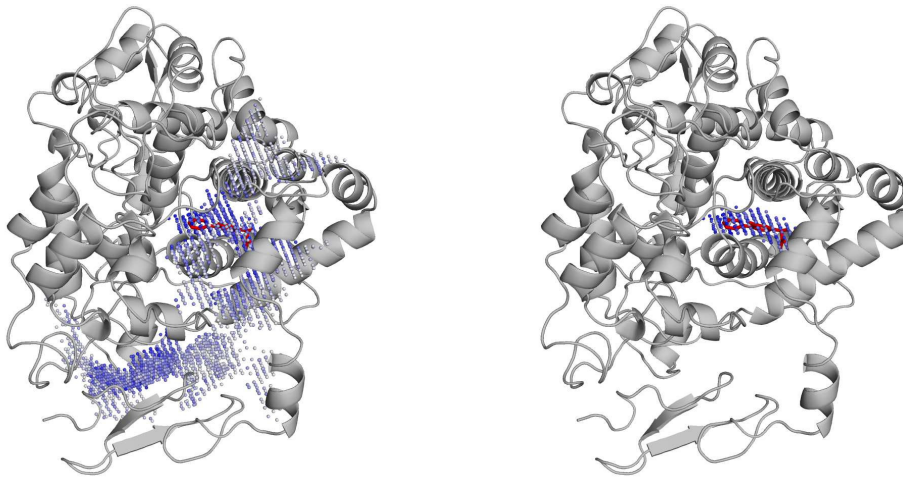


Abbildung 28. Modell der Cyclooxygenase 2 (PDB Struktur 3PGH) mit der größten vorhergesagten Bindetasche (blaue Kugeln, links) beziehungsweise der manuell verkleinerten Tasche (blaue Kugeln, rechts). Der gebundene Ligand ist rot dargestellt.

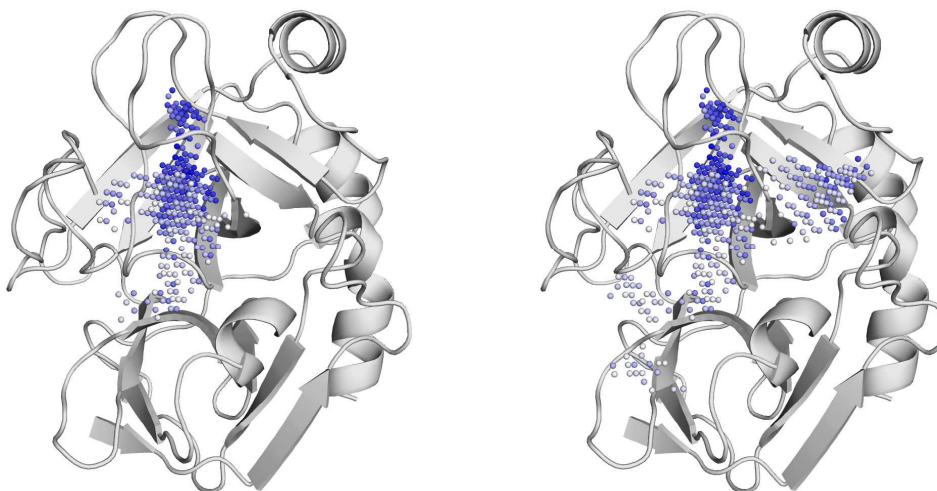


Abbildung 29. Modell der Protease Faktor Xa (PDB Struktur 2BOK) mit der größten vorhergesagten Bindetasche (blaue Kugeln, links) beziehungsweise den Taschen 1, 6, 13 und 17 (blaue Kugeln, rechts).

Die Struktur 1DPO repräsentiert zusätzlich eine Ser195Met Mutante des Trypsins. Ser195 liegt an der S1 Tasche von Trypsin (Perona und Craik, 1995) und eine mögliche Wasserstoffbrückendonator- oder Akzeptorfunktion könnte hier vorhergesagt werden und die Qualität des Pharmakophormodells verbessern, was mit Methionin nicht möglich ist.

Die Wahl von PocketPicker als Methode zur Vorhersage der Ligandenbindetaschen wird durch die Ergebnisse gerechtfertigt. In den meisten Fällen ist die erste und damit größte Bindetasche die tatsächliche Ligandenbindetasche und das darauf basierende Pharmakophormodell ist ausreichend eine Anreicherung im virtuellen Screening zu erzielen.

6.7.4 Parameterauswahl für das prospektive Screening

Da LIQUID für das Clustering und die Berechnung des Korrelationsvektors benutzt wurde, ist es möglich, dass die Auswahl der Parameterwerte der Berechnung Auswirkungen auf den Erfolg des virtuellen Screenings hatte. LIQUID hat vier variable Parameter: Clusterradien für jeden der drei Interaktionstypen und die Skalierungsart des Berechneten Korrelationsvektors. Der Vektor oder jeder Block eines Interaktionstypenpaars kann auf ein Maximum von eins skaliert werden oder es kann auf eine Skalierung verzichtet werden.

Weiterhin ist die Wahl des (Un-)Ähnlichkeitsmaßes für den Vergleich von Vektoren ein möglicherweise entscheidender Faktor. Rupp *et al.* (Rupp *et al.*, 2009) empfehlen, ein virtuelles Screening mit mehreren (Un-)ähnlichkeitsmaßen durchzuführen und dabei mindestens eine Minowskimetrik und einen Ähnlichkeitskoeffizienten zu verwenden. Es wurden daher die hier euklidische Metrik, die Manhattanmetrik (Minowskimetriken basierend auf der L^2 bzw. L^1 Norm) und der Carbóindex (Kosinusähnlichkeit basierend auf dem inneren Produkt; Carbó *et al.*, 1980) ausgewählt.

Die Wahl der Clusterradien wird durch den Aufbau des PocketPicker Gitters und die daraus abgeleiteten diskreten Bindetaschenrepräsentationen beeinflusst, welches einen Abstand von einem Ångström zwischen horizontal oder vertikal benachbarten Gitterpunkten hat. Ein Wert von mindestens der

Quadratwurzel aus drei, also dem maximalen Abstand zweier benachbarter Punkte, und von weniger als zwei scheint angebracht zu sein. So werden immer zwei direkt benachbarte Punkte in einem Cluster vereinigt, aber nicht ein Punkt mit seinem übernächsten Nachbarn (Abbildung 30). Ein Wert von $1,9 \text{ \AA}$ wurde daher für alle Clusterradien ausgewählt.

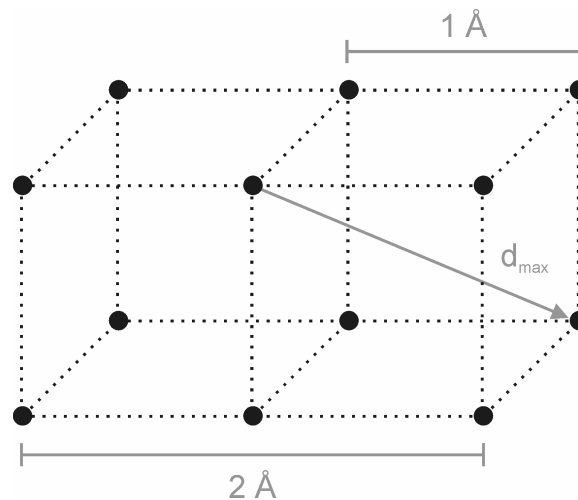


Abbildung 30. Distanzen im regelmäßigen Gitter, welches von PocketPicker zur Vorhersage von Bindetaschen genutzt wird. d_{\max} entspricht der Quadratwurzel aus Drei, also ungefähr $1,73 \text{ \AA}$.

Im Fall der lipophilen Interaktionen konnte für viele Modelle beobachtet werden, dass eine große Anzahl von Punkten der Bindetaschenrepräsentation mit einem lipophilen Interaktionstyp markiert wird. So sind zum Beispiel bei Trypsin (1dpo) 160 der 163 Punkte der Repräsentation der größten vorhergesagten Bindetasche als lipophil markiert. Dies ist wahrscheinlich das Resultat der ungerichteten Natur und der großen Reichweite der lipophilen Interaktionen nach dem benutzten Modell der LUDI Regeln (siehe Tabelle 1). Ein Clustering führt hier möglicherweise zu großen und unspezifischen Clustern. Um diesen potentiellen Effekt zu studieren, wurde der Clusterradius der lipophilen Interaktionen neben $1,9 \text{ \AA}$ auf $1,5 \text{ \AA}$ und 4 \AA gesetzt.

Tabelle 10 zeigt die Ergebnisse des retrospektiven Screenings, gemittelt über alle betrachteten Wirkstoffziele und Modelle. Betrachtet man die Unterschiede im Bezug auf verwendete Metrik und Clusterradius sind nur geringe Unterschiede zu erkennen, jedoch sind die Mittelwerte für die beiden kleineren

Clusterradien oft geringfügig besser. Im Unterschied dazu gibt es bei der Wahl der Skalierungsmethode teilweise große Unterschiede, wobei die Blockskalierung oft am schlechtesten abschneidet. Für das prospektive Screening nach Inhibitoren für die Protease HtrA wurde daher auf eine Skalierung verzichtet und ein Clusterradius für lipophile Interaktionen von 1,5 Å und die Manhattanmetrik verwendet. Mit dieser Auswahl werden für alle drei Testdatenbanken durchschnittliche ROCAUC Werte von mindestens 0,6 erzielt.

Tabelle 10. Ergebnisse des retrospektiven Screenings, gemittelt über die Einzelexperimente aller Wirkstoffziele (Standartabweichung in Klammern). Die jeweils höchsten ROCAUC und BEDROC Werte für jede Screeningdatenbank sind hervorgehoben.

Datenbank	Metrik ¹⁾	Cluster-radius [Å] ²⁾	keine Skalierung				Blockskalierung				Skalierung auf Eins			
			ROCAUC	Standart-abweichung	BEDROC	Standart-abweichung	ROCAUC	Standart-abweichung	BEDROC	Standart-abweichung	ROCAUC	Standart-abweichung	BEDROC	Standart-abweichung
COBRA	E	1,5	0,60	(0,13)	0,09	(0,05)	0,55	(0,11)	0,05	(0,03)	0,59	(0,19)	0,11	(0,09)
		1,9	0,61	(0,11)	0,11	(0,08)	0,58	(0,11)	0,08	(0,08)	0,60	(0,19)	0,14	(0,13)
		4	0,58	(0,12)	0,10	(0,07)	0,59	(0,09)	0,08	(0,09)	0,57	(0,19)	0,13	(0,13)
	M	1,5	0,60	(0,14)	0,08	(0,06)	0,57	(0,13)	0,06	(0,06)	0,58	(0,21)	0,10	(0,09)
		1,9	0,62	(0,14)	0,12	(0,10)	0,63	(0,11)	0,11	(0,11)	0,59	(0,18)	0,13	(0,13)
		4	0,58	(0,18)	0,10	(0,11)	0,62	(0,14)	0,10	(0,12)	0,55	(0,21)	0,12	(0,15)
	C	1,5	0,59	(0,14)	0,10	(0,08)	0,52	(0,10)	0,04	(0,02)	0,59	(0,14)	0,10	(0,08)
		1,9	0,62	(0,14)	0,14	(0,12)	0,57	(0,12)	0,07	(0,08)	0,62	(0,14)	0,14	(0,12)
		4	0,61	(0,14)	0,13	(0,10)	0,57	(0,14)	0,08	(0,09)	0,61	(0,14)	0,13	(0,11)
MUV	E	1,5	0,61	(0,08)	0,11	(0,07)	0,49	(0,11)	0,04	(0,03)	0,53	(0,09)	0,08	(0,06)
		1,9	0,56	(0,09)	0,11	(0,09)	0,51	(0,13)	0,07	(0,07)	0,52	(0,12)	0,07	(0,06)
		4	0,59	(0,10)	0,11	(0,09)	0,55	(0,10)	0,07	(0,07)	0,53	(0,12)	0,08	(0,06)
	M	1,5	0,61	(0,07)	0,09	(0,03)	0,49	(0,11)	0,05	(0,03)	0,55	(0,09)	0,07	(0,04)
		1,9	0,54	(0,08)	0,10	(0,09)	0,52	(0,12)	0,06	(0,03)	0,51	(0,09)	0,06	(0,04)
		4	0,60	(0,11)	0,10	(0,10)	0,55	(0,13)	0,08	(0,07)	0,53	(0,11)	0,07	(0,05)
	C	1,5	0,53	(0,08)	0,06	(0,04)	0,50	(0,07)	0,04	(0,02)	0,53	(0,07)	0,06	(0,04)
		1,9	0,53	(0,08)	0,06	(0,04)	0,51	(0,09)	0,05	(0,04)	0,53	(0,08)	0,06	(0,04)
		4	0,54	(0,09)	0,07	(0,05)	0,52	(0,08)	0,07	(0,05)	0,54	(0,09)	0,07	(0,04)
UGI	E	1,5	0,61	(0,05)	0,25	(0,23)	0,52	(0,03)	0,18	(0,11)	0,61	(0,08)	0,30	(0,25)
		1,9	0,61	(0,06)	0,25	(0,23)	0,52	(0,02)	0,18	(0,14)	0,61	(0,08)	0,29	(0,24)
		4	0,58	(0,08)	0,24	(0,23)	0,56	(0,02)	0,16	(0,12)	0,59	(0,11)	0,27	(0,22)
	M	1,5	0,62	(0,03)	0,26	(0,22)	0,57	(0,04)	0,22	(0,15)	0,61	(0,05)	0,29	(0,25)
		1,9	0,61	(0,04)	0,25	(0,24)	0,56	(0,05)	0,19	(0,16)	0,61	(0,06)	0,29	(0,26)
		4	0,56	(0,08)	0,21	(0,21)	0,56	(0,05)	0,15	(0,12)	0,56	(0,09)	0,25	(0,23)
	C	1,5	0,61	(0,10)	0,26	(0,22)	0,51	(0,04)	0,17	(0,10)	0,61	(0,10)	0,26	(0,22)
		1,9	0,64	(0,09)	0,26	(0,21)	0,54	(0,03)	0,17	(0,13)	0,64	(0,09)	0,26	(0,21)
		4	0,62	(0,09)	0,24	(0,20)	0,56	(0,03)	0,16	(0,14)	0,62	(0,09)	0,24	(0,20)

1) E – euklidisch, M – Manhattan, C – Carbóindex

2) LIQUID Clusterradius für lipophile Interaktionen.

6.8 Prospektives Screening

6.8.1 Virtuelles Ligandenmodell der Protease HtrA

Die Wahl der Parameterwerte für die Berechnung des virtuellen Ligandenmodells von HtrA beruht auf den Erkenntnissen aus dem retrospektiven Screening. Es wurde die Manhattan-Distanz für Vektorvergleiche in Verbindung mit einem kleinen Clusterradius (1,5 Å) für lipophile Interaktionen gewählt, da diese Einstellung oft die besten bzw. gleichwertige Screening-ergebnisse erzeugte. Außerdem wurde auf eine Skalierung der Korrelationsvektoren verzichtet.

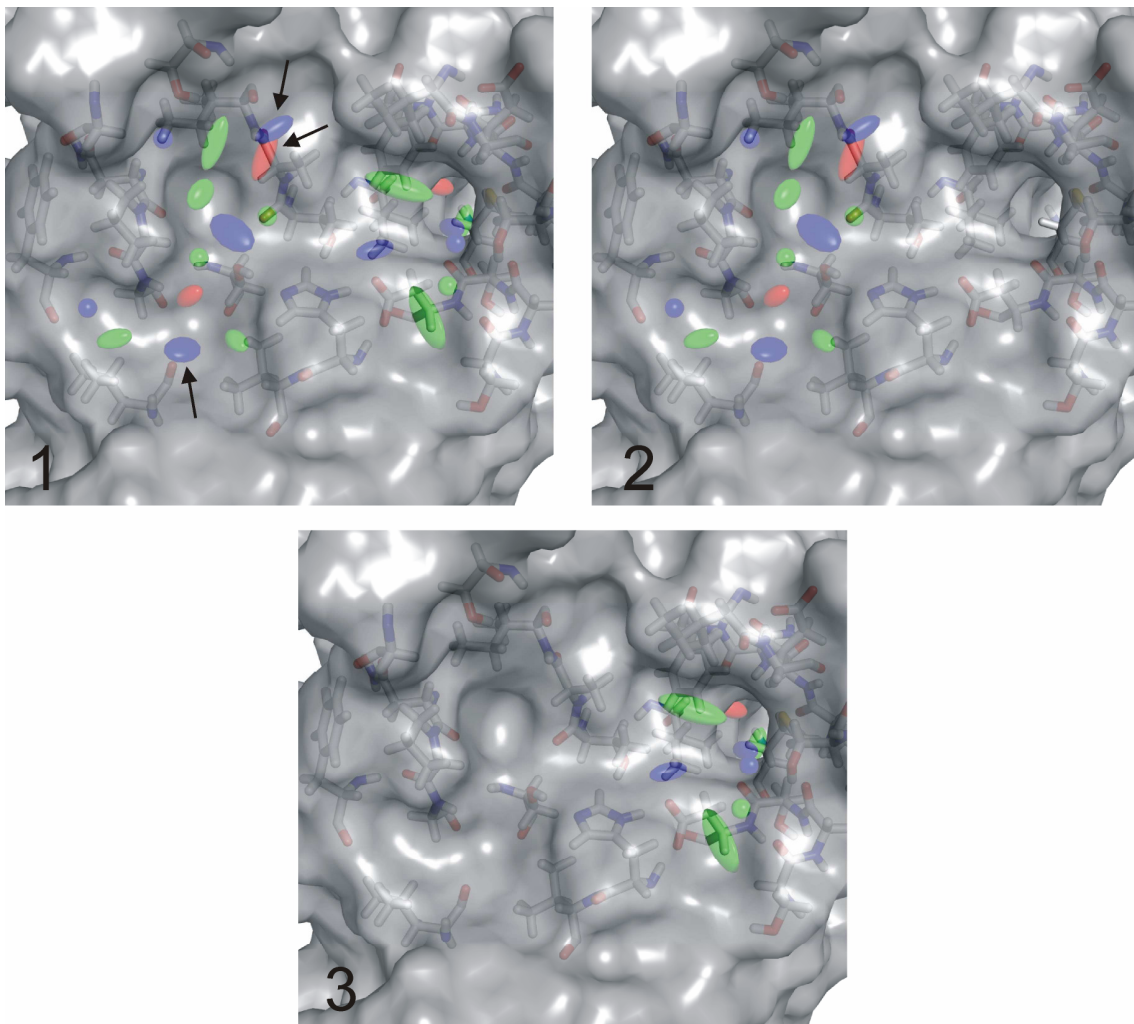


Abbildung 31. Schematische Darstellung der drei virtuellen Ligandenmodelle (1, 2 und 3) für das prospektive Screening (grün: lipophil, blau: Wasserstoffbrückendonor, rot: Wasserstoffbrückenakzeptor). Die Pfeile markieren Substrat-Protease Interaktionen, die bei Serinproteasen häufig vorkommen und korrekt vorhergesagt wurden (Perona und Craik, 1995).

Da die Auswahl der Bindetaschen als Grundlage für den virtuellen Liganden einen hohen Einfluss auf das Screeningergebnis hat, wurden drei Modelle erstellt (Abbildung 31). Modell 1 benutzt alle drei vorhergesagten Bindetaschen (11, 12, 38), Modell 2 die Taschen 12 und 38 und 3 nur die Tasche 11.

Eine Beobachtung unterstreicht die Plausibilität des Modells bzw. der vorhergesagten Interaktionen: Es ist bekannt, dass das Proteinrückgrat von Serinproteasen an bestimmten Stellen das Rückgrat des Substrats über Wasserstoffbrücken vor und nach der Schnittstelle bindet (Perona und Craik, 1995). Involviert sind die Carbonylsauerstoffe von Phe41 und Gly216 und die Amidgruppe von Gly216 (Nummerierung von Chymotrypsin; Hedstrom, 2002). Diese Interaktionen werden auch im virtuellen Ligandenmodell korrekt abgebildet (Abbildung 31).

6.8.2 Inhibitionsexperimente mit HtrA

Jedes der drei Modelle wurde mit den Asinex und SPECS Datenbanken verglichen und insgesamt 26 Substanzen (Tabelle 11) wurden aus den jeweils 100 Molekülen mit der geringsten Manhattandistanz zum Modell ausgewählt, bestellt und getestet (die kompletten Listen werden in Anhang E gezeigt). Abbildung 32 zeigt die mit einem E-Cadherin Antikörper markierten Western-Blots des ersten Experiments (die Ergebnisse der Kontrollexperimente werden in Anhang C gezeigt). Jede Substanz wurde bei einer Konzentration von 100 μM auf die Inhibition der Protease HtrA getestet. Nach der Inkubationszeit ist das komplette rekombinante E-Cadherin der Positivkontrolle (Spur +) verdaut, was durch die fehlende Bande gezeigt wird. Die Negativkontrolle zeigt das unverdaute Substrat, wobei gleiche Mengen in jede Spur geladen wurden. Dementsprechend ist eine verbleibende E-Cadherin Bande in den weiteren Spuren das Resultat einer Hemmung der Proteaseaktivität von HtrA.

Tabelle 11 zeigt die Strukturformeln der bestellten Substanzen zusammen mit der gemessenen mittleren relativen Inhibition der Protease HtrA bei 100 μM Substanzkonzentration. Von den 26 bestellten Substanzen waren nur 22 löslich in DMSO und sechs inhibieren die Aktivität von HtrA. Die Standardabweichung der relativen Inhibition ist in einigen Fällen sehr hoch (bis zu 102% des

Mittelwerts; Tabelle 11, Zeile 25). Die Messwerte sind daher nur für einen groben Überblick geeignet. Trotzdem wird das Verhältnis der Messwerte durch den visuellen Eindruck der markierten Western-Blots bestätigt, wobei Substanz **12** eine Ausnahme ist. Im Folgenden wird angenommen, dass die relative Inhibition bei Verbindung **3** am geringsten ist und über Verbindung **14**, **5** und **25** bis zu Nummer **8** ansteigt.

Für die diese Substanz wurde ein IC_{50} Wert (Steinhilber *et al.*, 2005) bestimmt (Abbildung 33; die Blots der Kontrollexperimente werden in Anhang C gezeigt). Die berechnete Dosis-Wirkungs-Kurve (Abbildung 34) zeigt, dass Verbindung Nummer **8** ein partieller kompetitiver Inhibitor ist mit einer theoretischen maximalen Hemmung von 77%. Die halbmaximale Hemmung von HtrA durch Verbindung **8** wird daher bei einer Konzentration von 26 μ M erreicht. Durch die Ausreißer in den Daten (Abbildung 34) ist der Fehler für die Anpassung der logistischen Funktion an die Daten sehr hoch ($\chi^2 = 6,103$). Der IC_{50} ist daher nur als grobe Annäherung zu betrachten.

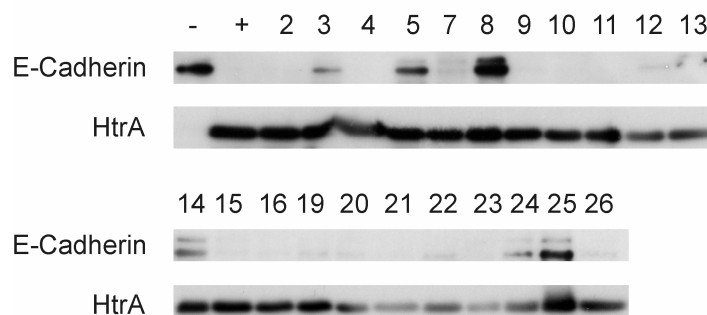


Abbildung 32. Mit E-Cadherin bzw. HtrA markierte Western-Blots des Screenings nach Inhibitoren für die Protease HtrA. Die Spur - ist eine Negativkontrolle ohne HtrA, die Spur + die Positivkontrolle ohne Inhibitor. Die anderen Spuren sind entsprechend der benutzten Substanz nummeriert (vgl. Tabelle 11).

Die Substanzen **3**, **5**, **8**, **12** und **14** hemmen die Aktivität von HtrA und wurden mit dem virtuellen Ligandenmodell 1 gefunden (Abbildung 31); die HtrA inhibierende Verbindung **25** wurde mit Modell 3 gefunden. Alle Substanzen, die von der Ergebnisliste des zweiten Modells ausgewählt wurden (**15-23**), haben keinen Einfluss auf die Aktivität von HtrA. Dieses Modell beinhaltet nicht die vorhergesagte Bindetasche 11 bzw. die vermutete S3 Tasche, was eine wichtige Rolle dieser Bindetasche für die Inhibition von HtrA und Substraterkennung und -bindung andeutet.

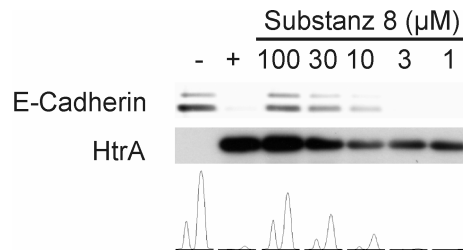


Abbildung 33. Mit E-Cadherin bzw. HtrA markierte Western-Blots für die Bestimmung eines IC_{50} der Substanz **8** (vgl. Tabelle 11). Unter den Blots sind die mittleren Farbdichten der Pixelzeilen der E-Cadherinbanden aufgetragen. Die Integrale dieser Kurven sind die Grundlage für die Berechnung der relativen Inhibition.

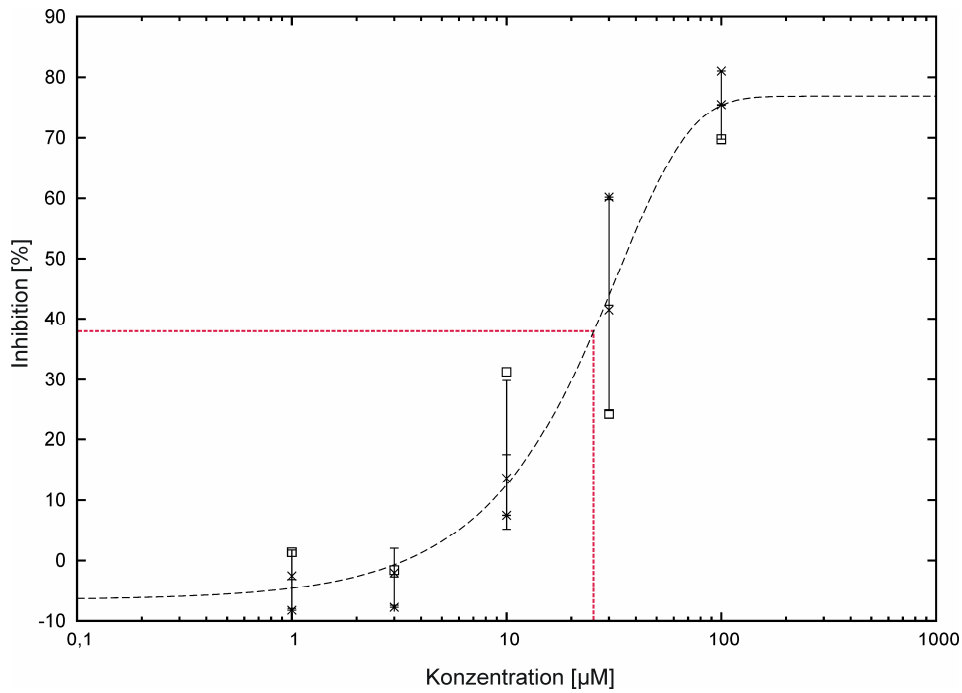
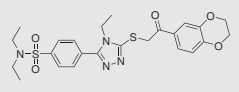
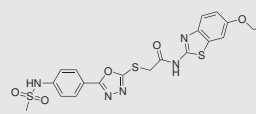
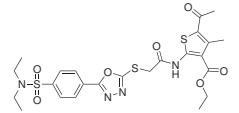
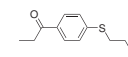
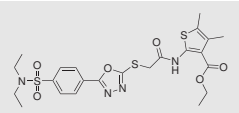
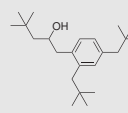
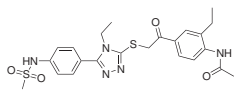
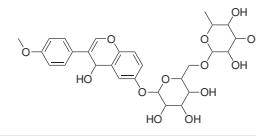
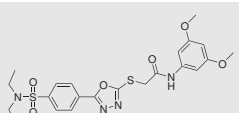
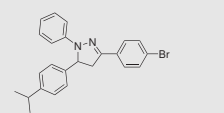
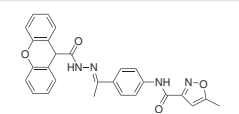
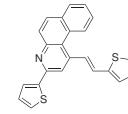
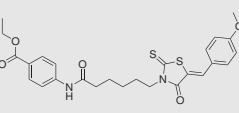
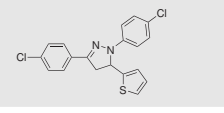
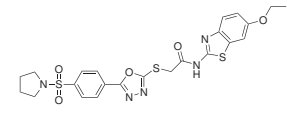
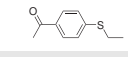
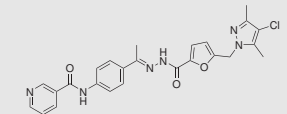
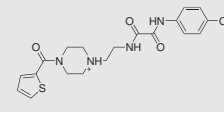
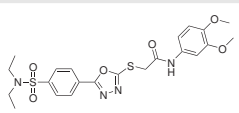
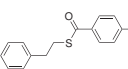
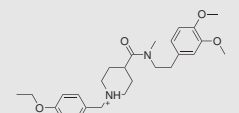
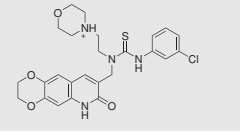
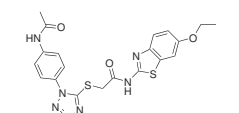
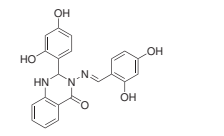
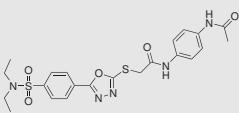
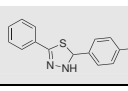


Abbildung 34. Berechnete Dosis-Wirkungs-Kurve für Substanz **8**. Die Fehlerbalken geben die Standardabweichung von drei unabhängigen Messungen wieder. Die individuellen Messwerte sind als Quadrate, Sterne bzw. Kreuze dargestellt. Die rote Linie markiert die Substanzkonzentration bei halbmaximaler Inhibition (= IC_{50}).

Tabelle 11. Strukturen der bestellten Substanzen und relative Inhibition von HtrA.

Nr.	Struktur	I ¹⁾	σ ²⁾	M ³⁾	Nr.	Struktur	I ¹⁾	σ ²⁾	M ³⁾
1		n.f. ⁴⁾		1	14		39,0	31,4	1
2		inaktiv		1	15		inaktiv		2
3		19,9	8,7	1	16		inaktiv		2
4		inaktiv		1	17		n.f. ⁴⁾		2
5		52,9	32,9	1	18		n.f. ⁴⁾		2
6		n.f. ⁴⁾		1	19		inaktiv		2
7		inaktiv		1	20		inaktiv		2
8		66,2	27,8	1	21		inaktiv		2
9		inaktiv		1	22		inaktiv		2
10		inaktiv		1	23		inaktiv		2
11		inaktiv		1	24		inaktiv		2
12		59,8	52,7	1	25		56,5	57,5	3
13		inaktiv		1	26		inaktiv		3

1) Inhibition von HtrA in % bei einer Konzentration von 100 μ M.

2) Standardabweichung der gemessenen Inhibition.

3) Modellnummer.

4) Aktivität nicht feststellbar, da die entsprechende Substanz in DMSO nicht löslich ist.

6.8.3 Docking und Struktur-Wirkungs-Beziehung

Um potentielle Bindemodi der gefundenen inhibierenden Substanzen zu untersuchen, wurde mit *GOLD* (Cole *et al.*, 2005) die Konformation der Verbindungen, gebunden an das aktive Zentrum von HtrA, vorhergesagt. Die Aminosäuren des aktiven Zentrums von HtrA wurden dabei vom Algorithmus zur Berechnung des virtuellen Liganden übernommen (Abbildung 35; gesamte Parametereinstellungen in Anhang D). Für die Vorhersage wurden die Strukturmodelle der potentiellen Liganden als flexibel und die Bindetasche von HtrA als starr betrachtet.

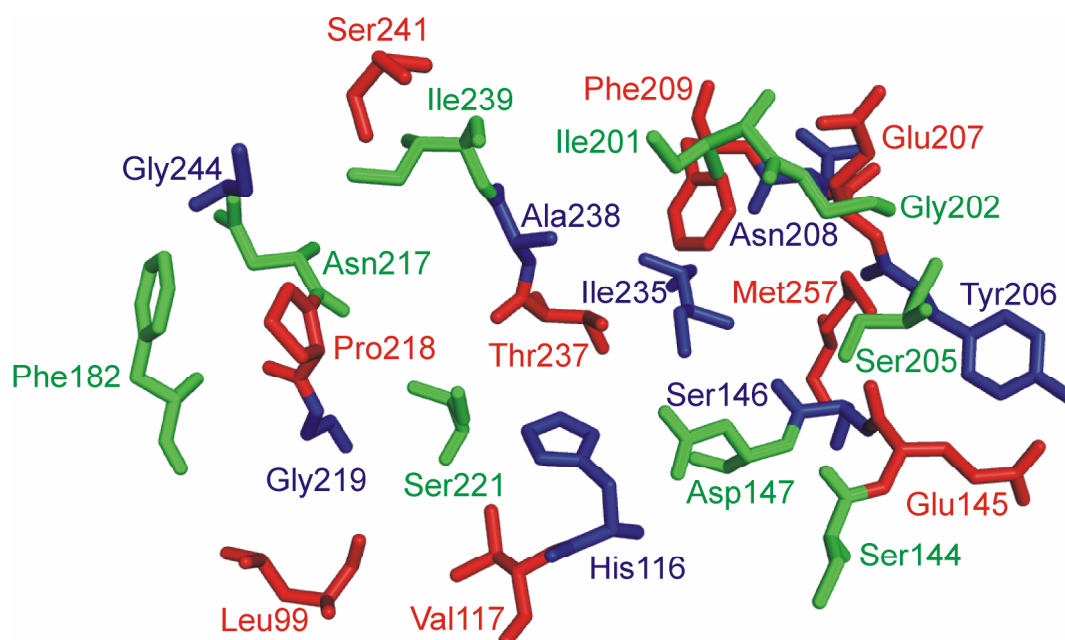


Abbildung 35. Das aktive Zentrum von HtrA umgebende Aminosäuren. Für eine bessere Unterscheidung sind die einzelnen Aminosäuren rot, grün oder blau gefärbt und die Wasserstoffatome wurden entfernt. Die Ansicht ist analog zu den Abbildungen 12, 17 und 18.

Abbildung 37 und 38 zeigen die höchstbewerteten vorhergesagten Konformere und jeweiligen *GoldScore* Werte (Cole *et al.*, 2005). Die Verbindungen **3**, **5**, **8** und **14** haben ein gemeinsames Strukturgerüst (Abbildung 36), welches an zwei Stellen variable Seitenketten hat (R_1 und R_2 in Abbildung 36). Diese Anreicherung an ähnlichen Strukturen ist ein weiterer Hinweis auf den Erfolg des virtuellen Screenings.

Die Vorhersagen zeigen, dass nur die Verbindungen **8** und **14** die Bindetasche 11, also die potentielle S3 Tasche, mit der Seitenkette an R_1 -Position ausfüllen.

Das Ringsystem der R₁-Gruppe von Verbindung **8** interagiert demnach mit Phe209 und der terminale Methylrest wird in die lipophile Tasche 11 platziert. Die Interaktionen werden dort von den Seitenketten der Aminosäuren Ile253 und Met257 vermittelt. Dies kann auch bei Verbindung **14** beobachtet werden, aber nicht bei den Verbindungen **3** und **5**, welche eine sperrigere R₁-Gruppe haben. Da diese nicht in Tasche 11 passt, wird eine Konformation, bei der das Strukturgerüst in umgekehrter Richtung vorliegt, vom Docking-Algorithmus mit einer höheren Bewertung versehen. Die jeweiligen großen R₁-Gruppen der Verbindungen **3** und **5** werden in der Nähe der Tasche 38 platziert, welche offener als Tasche 11 ist. Die Stickstoffatome der Oxadiazolgruppe des Grundgerüsts können keine Wasserstoffbrücke zur Amidgruppe des Proteinrückgrats der Aminosäure Ile239 ausbilden, was allerdings bei den Verbindungen **8** und **14** vorhergesagt wird. Diese Beobachtungen könnten die geringere Aktivität der Verbindungen **3** und **5** im Gegensatz zu Nr. **8** erklären. Im vorhergesagten Bindungsmodell werden die R₂-Gruppen der Verbindungen **8** und **14** in Tasche 38 platziert. Die Sulfongruppe von Verbindung **8** kann so mit der Amidgruppe der Aminosäure Gly219 eine Wasserstoffbrücke bilden. Diese günstige Position der Sulfongruppe ist bei Verbindung **14** nicht zu beobachten, da sie nicht direkt an den Phenylring gebunden ist. Weiterhin kann die Pyrrolidgruppe von Verbindung **8** mit der hydrophoben Umgebung der Tasche 38 interagieren. Insgesamt ist dies eine mögliche Erklärung für die geringere Inhibition von HtrA durch Verbindung **14** im Vergleich zu Verbindung **8**.

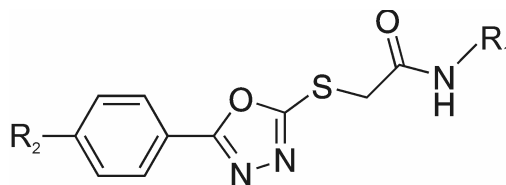


Abbildung 36. Gemeinsames Strukturgerüst der Testsubstanzen **2**, **3**, **5**, **8**, **10**, **13** und **14**. R₁ und R₂ bezeichnen die variablen Seitenketten.

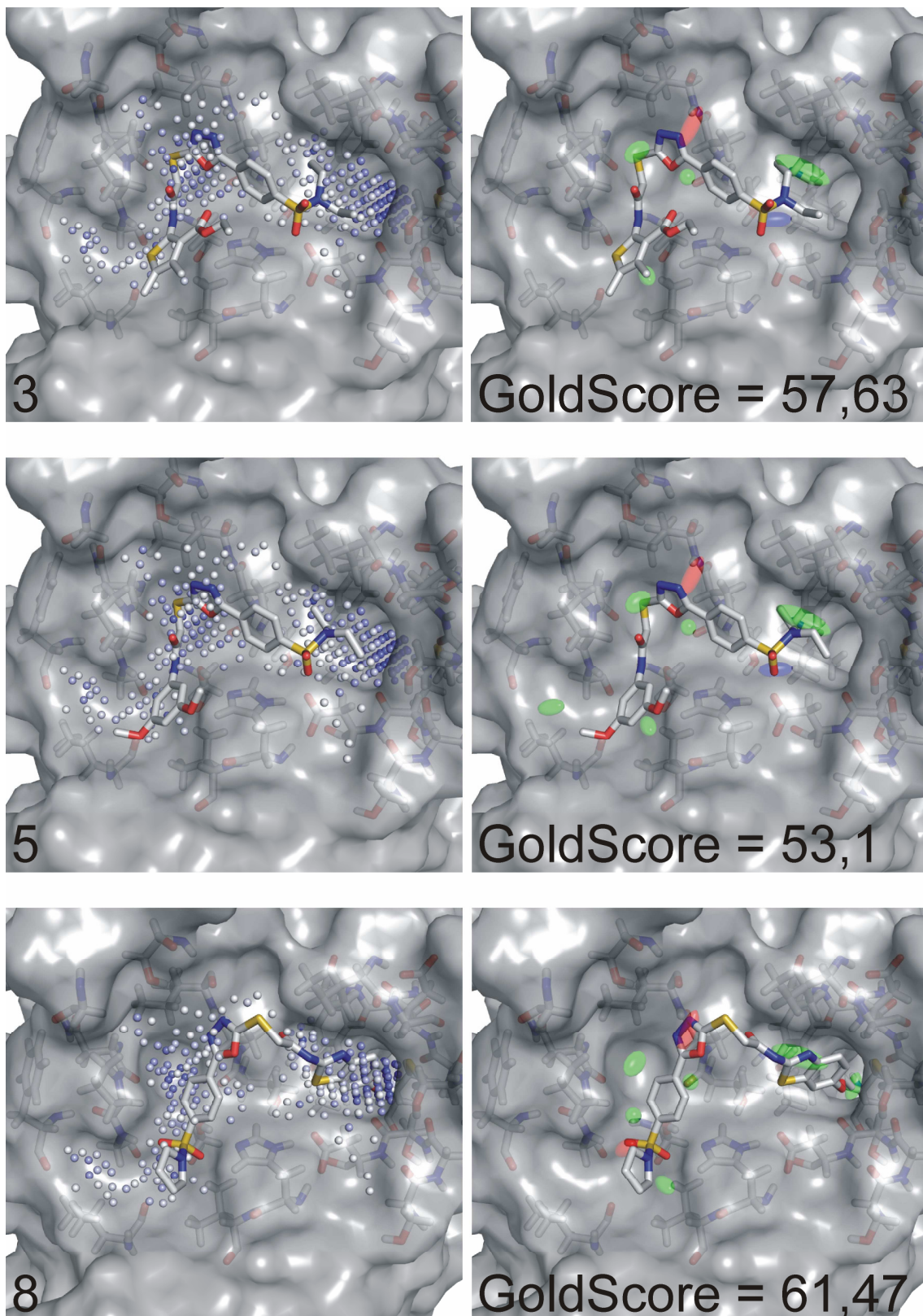


Abbildung 37. Ergebnisse der Bindemodivorhersage für die Verbindungen **3**, **5** und **8** mit den vorhergesagten Bindetaschen (links), den umliegenden potentiellen Interaktionen (rechts, grün: lipophil, blau: Wasserstoffbrückendonor, rot: Wasserstoffbrückenakzeptor) und der Bewertung der Pose als GoldScore.

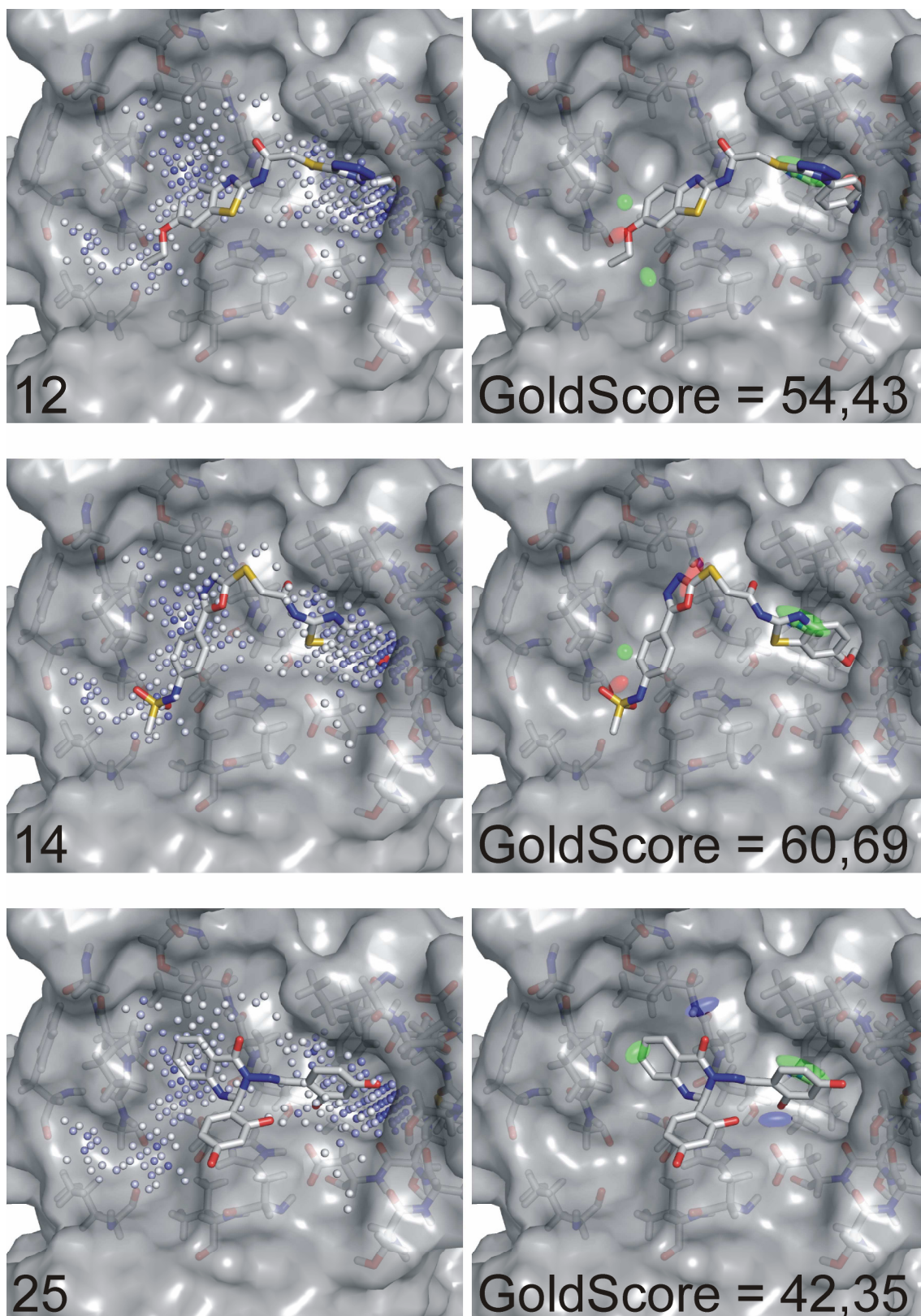


Abbildung 38. Ergebnisse der Bindemodivorhersage für die Verbindungen **12**, **14** und **25** mit den vorhergesagten Bindetaschen (links), den umliegenden potentiellen Interaktionen (rechts, grün: lipophil, blau: Wasserstoffbrückendonor, rot: Wasserstoffbrückenakzeptor) und der Bewertung der Pose als GoldScore.

Obwohl die Verbindungen **2**, **10** und **13** auch die in Abbildung 36 gezeigte Grundstruktur haben, inhibieren sie nicht die Aktivität von HtrA. Verbindung **2** hat die größte R₁-Gruppe, was die fehlende Wirkung erklären könnte. Verbindung **10** und **13** sind ähnlich zu Verbindung **5**, wobei diese HtrA inhibiert. Verbindung **10** unterscheidet sich nur durch eine 3,4 Konfiguration der Dimethoxybenzolgruppe von Nummer **5**, welche 3,5 konfiguriert ist. Diese geringe Änderung der Struktur hat anscheinend eine große Auswirkung auf die Wirkung im Experiment, welche komplett verschwindet, was ein Anzeichen für eine steile Aktivitätslandschaft sein könnte (Maggiora, 2006). Bei Verbindung **13** findet sich analog zu Verbindung **10** auch ein Substituent in *para* Stellung am Phenylring der R₁-Gruppe. Diese Stellung scheint demnach nicht vorteilhaft für eine Bindung an HtrA zu sein.

Im Gegensatz dazu haben die Verbindungen **12** und **25** ein anderes Strukturgerüst und inhibieren die Protease HtrA. Verbindung **12** besetzt im vorhergesagten Bindungsmodus auch Tasche 11 mit einer Acetamidgruppe, während Verbindung **25** einen im Vergleich zu den anderen Substanzen einzigartigen Bindungsmodus zu haben scheint.

Insgesamt wird jeweils nur ein Teil der 21 vorhergesagten Interaktionen durch die möglichen Bindungen der einzelnen Substanzen ausgebildet. Es bleibt allerdings offen, ob die Vorhersage des Bindungsmodus tatsächlich korrekt ist. Auf der anderen Seite ist es ein bekanntes Problem strukturabgeleiteter Pharmakophormodelle, dass oft zu viele Interaktionen vorhergesagt werden (Barillari *et al.*, 2008). Für ligandenbasierende Pharmakophormodelle werden oft nur drei oder vier Interaktionspunkte benutzt (Baroni *et al.*, 2007) und auch bei den vorhergesagten Bindungsmodi sind maximal sechs Übereinstimmungen mit dem virtuellen Liganden zu finden. Da trotzdem sowohl die retrospektive als auch die prospektive Anwendung erfolgreich waren, scheint die Kodierung der Informationen in einen Korrelationsvektor, welcher als Wahrscheinlichkeitsverteilung interpretiert werden kann, ausschlaggebend zu sein.

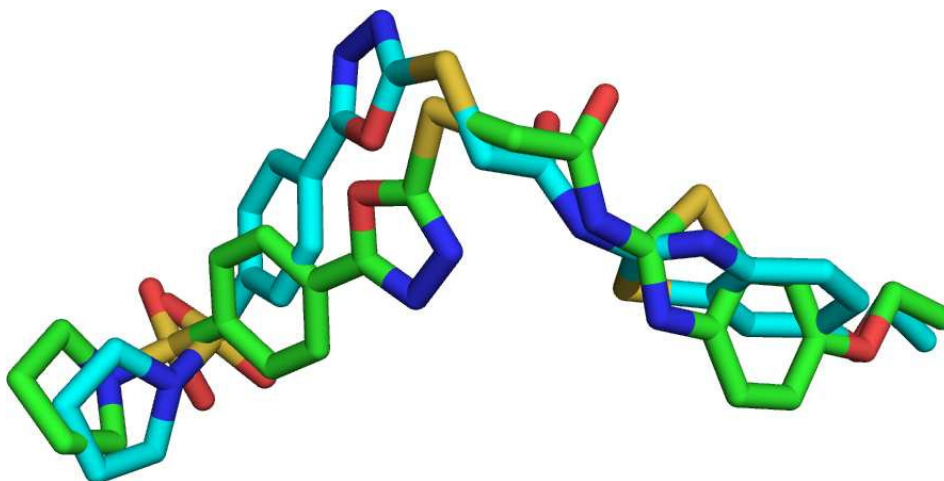


Abbildung 39. Überlagerung der von GOLD vorhergesagten und an HtrA gebundenen Konformation von Verbindung **8** (cyan) und der Konformation in der Screeningdatenbank (grün).

Diese Wahrscheinlichkeitsverteilung kodiert nur die Distanzverhältnisse der potentiellen Pharmakophorpunkte der jeweiligen Bindetasche bzw. des jeweiligen Moleküls. Konformationsisomere, die durch Rotationen entstehen, können daher ähnliche Korrelationsvektoren haben. Eine Überlagerung der Konformation von Verbindung **8**, welche GOLD berechnet hat (Abbildung 38) mit der Konformation, welche mit dem virtuellen Screening gefunden wurde (Abbildung 39), zeigt ein Beispiel für ein solches Verhalten. Die Ringsysteme der Konformationen der R₁-Gruppen stehen senkrecht zueinander, trotzdem wurde eine hohe Ähnlichkeit der zugehörigen Korrelationsvektoren berechnet. Die generelle Struktur der Bindetasche wurde allerdings erfolgreich in den Korrelationsvektor kodiert, da beide Konformationen nicht gestreckt sind sondern eine ähnliche Biegung im Grundgerüst haben; die R₁- und R₂- Gruppen liegen nahe beieinander (RMSD = 1,37 Å).

Die erwähnten Beobachtungen beruhen auf der Annahme, dass GOLD die korrekten Bindemodi vorhergesagt hat. Die Verbindungen **3**, **5**, **8** und **14** wurden daher mit einem flexiblen Alignment mit MOE überlagert (Abbildung 40) und mit LIQUID ein Konsenspharmakophormodell berechnet. Diese Substanzen inhibieren HtrA und haben die gleiche Grundstruktur, allerdings wird nun angenommen, dass aufgrund dieser Grundstruktur ein gleicher Bindungsmodus vorliegt. Hier zeigt sich, dass die Substituenten des Benzolrings der R₁-Gruppe von Verbindung **3** und **5** (Carboxylethylestergruppe

bzw. Methoxygruppen) raumfüllend sind und so möglicherweise durch sterische Hinderungen die geringere relative Inhibition verursachen.

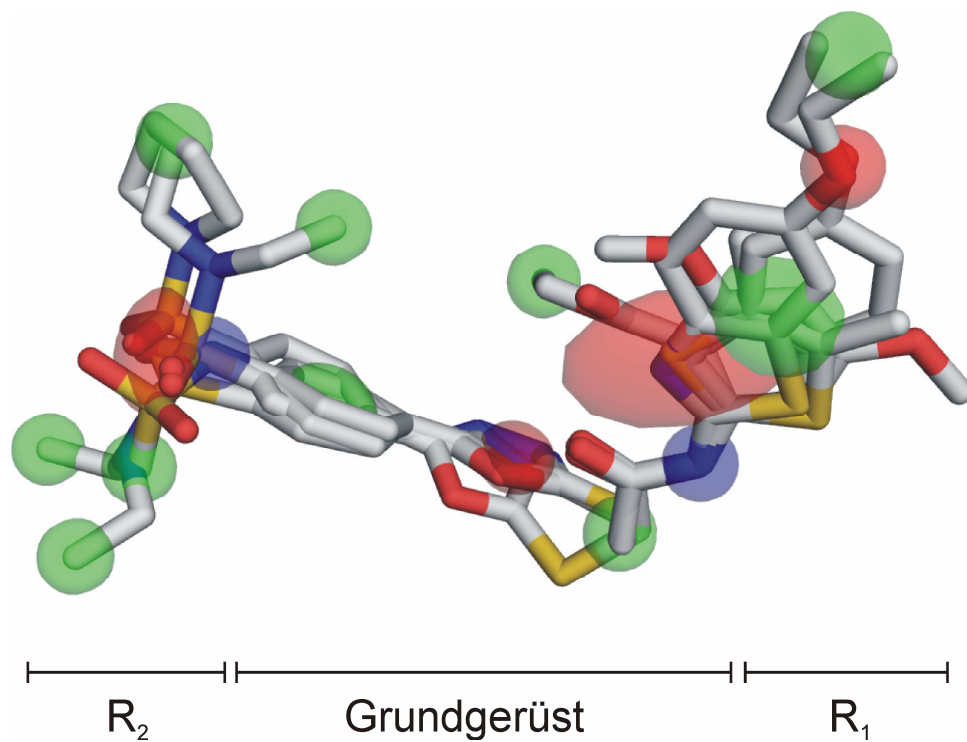


Abbildung 40. Flexibles Alignment und LIQUID Konsenspharmakophormodell der Verbindungen **3**, **5**, **8** und **14** (grün: lipophil, blau: Wasserstoffbrückendonor, rot: Wasserstoffbrückenakzeptor).

7. Ausblick

Das wichtigste Ergebnis der vorliegenden Arbeit ist die erstmalige Vorstellung eines niedermolekularen Inhibitors für die Protease HtrA von *H. pylori*. Dieser Inhibitor wird als wichtiges Werkzeug für die weitere Erforschung von HtrA und der Rolle dieses Enzyms in der Pathogenese von *H. pylori* dienen. Die entsprechenden Experimente werden in Kooperation mit dem Paul Ehrlich Institut in Langen durchgeführt, eine Publikation ist in Vorbereitung.

Sollte HtrA wirklich die vermutete Rolle in der Pathogenese von *H. pylori* haben, könnte eine Inhibition von HtrA eine neue Strategie sein, eine Infektion von Menschen mit *H. pylori* und die dadurch ausgelösten Krankheiten zu behandeln. Die in dieser Arbeit vorgestellte Serie von Inhibitoren wäre ein möglicher Startpunkt für eine weitergehende Wirkstoffentwicklung. Folgende Schritte könnten dazu durchgeführt werden:

1. Die Etablierung eines sensitiveren Experiments zur Bestimmung der relativen Inhibition von HtrA.
2. Die Verifikation des vorhergesagten Bindemodus, beispielsweise durch Kernspinresonanzspektroskopie oder durch Mutationsexperimente der Aminosäuren, die potentiell an der Ligandenbindung beteiligt sind.
3. Das Erstellen eines Modells zur quantitativen Beschreibung der Beziehung zwischen Aktivität und Struktur (engl. *quantitative structure activity relationships*, QSAR). Dies könnte die gezielte Synthese von neuen Substanzen ermöglichen.
4. Ein erneutes virtuelles Screening mit den bekannten Inhibitoren als Vorlage. Das in Abbildung 40 gezeigte Pharmakophormodell könnte als Grundlage dienen.

Auch das in dieser Arbeit vorgestellte Modell zur Berechnung von struktur-abgeleiteten Pharmakophormodellen bietet Potential für Verbesserungen. Zum einen könnten mehr Interaktionstypen wie zum Beispiel gerichtete aromatische Interaktionen berücksichtigt werden. Weiterhin könnte untersucht werden, ob eine Gewichtung der Interaktionstypen, die den Punkten der Bindetaschenrepräsentation zugewiesen werden, sich positiv auf die

retrospektiven Ergebnisse auswirkt. Die Gewichtung könnte zum Beispiel von den Vergrabenheitswerten von PocketPicker abgeleitet werden. Die Idee wäre, dass eine tief in einer Tasche vergrabene Interaktion wichtiger für eine Ligandenbindung und damit wahrscheinlicher ist.

8. Literaturverzeichnis

A

Abramoff MD, Magelhaes PJ, Ram SJ (2004) Image Processing with ImageJ. *Biophotonics International* 11:36-42.

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc Natl Acad Sci USA* 101:17837-17842.

Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst* 58:380-388.

Andersen LP (2007) Colonization and infection by *Helicobacter pylori* in humans. *Helicobacter* 12:12-15.

Aspholm M, Kalia A, Ruhl S *et al.* (2006) *Helicobacter pylori* Adhesion to Carbohydrates. *Methods Enzymol* 417:293–339.

Assay Drug Dev Technol (2008) Current trends in high-throughput screening. Roundtable discussion. *Assay and Drug Development Technologies* 6:491-504.

B

Backert S, Moese S, Selbach M, Brinkmann V, Meyer TF (2001) Phosphorylation of tyrosine 972 of the *Helicobacter pylori* CagA protein is essential for induction of a scattering phenotype in gastric epithelial cells. *Mol Microbiol* 42:631–644.

Backert S, Selbach M (2008) Role of type IV secretion in *Helicobacter pylori* pathogenesis. *Cell Microbiol.* 10:1573-1581.

Barillari C, Marcou G, Rognan D (2008) Hot-Spots-Guided Receptor-Based Pharmacophores (HS-Pharm): A Knowledge-Based Approach to Identify Ligand-Anchoring Atoms in Protein Cavities and Prioritize Structure-Based Pharmacophores. *Journal of Chemical Information and Modeling* 48:1396-1410.

Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J Chem Inf Model* 47:279-294.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research* 28:235-242.

Bingle LEH, Bailey CM, Pallen MJ (2008) Type VI secretion: a beginner's guide. *Curr Opin Microbiol* 11:3-8.

Black PE (2006) Dictionary of Algorithms and Data Structures [online]. U.S. National Institute of Standards and Technology. (<http://www.itl.nist.gov/div897/sqg/dads/>, Stand vom 26.5.2009).

Böhm HJ (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput-Aided Mol Des* 6:61–78.

Böhm HJ, Klebe G, Kubinyi H (1996) *Wirkstoffdesign*. Spektrum, Heidelberg.

Brown LM (2000) *Helicobacter pylori*: Epidemiology and Routes of Transmission. *Epidemiologic Reviews* 22:283-297.

Bumann D, Aksu S, Wendland M, Janek K, Zimny-Arndt U, et al. (2002) Proteome analysis of secreted proteins of the gastric pathogen *Helicobacter pylori*. *Infect Immun* 70: 3396–3403.

C

Carbó R, Leyda L, Arnau M (1980) How Similar is a Molecule to Another? *International Journal of Quantum Chemistry* 17:1185-1189.

Chen J, Lai L (2006) Pocket v.2: further developments on receptor-based pharmacophore modeling. *J Chem Inf Model* 46:2684-2691.

Cole JC, Nissink JWM, Taylor R (2005) Protein-Ligand Docking and Virtual Screening with GOLD. In: *Virtual Screening in Drug Discovery*. Herausgeber: Shoichet B, Alvarez J. Taylor & Francis CRC Press, Boca Raton, Florida, USA.

Cover TL, Blanke SR (2005) *Helicobacter pylori* VacA, a paradigm for toxin multifunctionality. *Nat Rev Microbiol* 3:320-32.

Cremonini F, Gasbarrini A (2003) Atopy, *Helicobacter pylori* and the hygiene hypothesis. *European Journal of Gastroenterology & Hepatology* 15:635–636.

D

Davis A, St-Gallay S, Kleywegt G (2008) Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discovery Today* 13:831-841.

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

Delaney B, McColl K (2005) Review article: *Helicobacter pylori* and gastro-oesophageal reflux disease. *Aliment Pharmacol Ther* 22:32-40.

Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. John Wiley & Sons, New York.

E

Ehrlich P (1909) Über den jetzigen Stand der Chemotherapie. *Dtsch Chem Ges* 42:17.

Eppinger M, Baar C, Linz B, Raddatz G, Lanz C, Keller H, Morelli G, Gressmann H, Achtman M, Schuster SC (2006) Who ate whom? Adaptive *Helicobacter* genomic changes that accompanied a host jump from early humans to large felines. *PLoS Genet* 2:e120.

F

Fara DC, Oprea TI, Prossnitz ER, Bologna CG, Edwards BS, Sklar LA (2006) Integration of virtual and physical screening. *Drug Discovery Today: Technologies* 3:377-385.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27:861-874.

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

Forde GM (2008) Preparation, analysis and use of an affinity adsorbent for the purification of GST fusion protein. *Methods Mol Biol* 421:125-136.

G

Galán JE, Wolf-Watz H (2006) Protein delivery into eukaryotic cells by type III secretion machines. *Nature* 444:567-573.

Galil Z, Italiano GF (2001) Data structures and algorithms for disjoint set union problems. *ACM Computing Surveys* 23:319 – 344.

Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* 28:849-857.

Gosling J, Joy B, Steele G, Bracha G (2005) *The Java Language Specification. Third Edition.* Addison-Wesley, München. <http://java.sun.com>

Graham DY, Shiotani A (2008) New concepts of resistance in the treatment of *Helicobacter pylori* infections. *Nature Clinical Practice Gastroenterology & Hepatology* 5:321-331.

H

Hanahan D (1983) Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* 166:557-80.

Hawkins P, Skillman G, Nicholls A (2006) Comparison of shape-matching and docking as virtual screening tools. *Journal of Medicinal Chemistry* 50:72-82.

Hedstrom L (2002) Serine Protease Mechanism and Specificity. *Chem Rev* 102:4201-4523.

von Heijne G (1985) Signal sequences. The limits of variation. *J Mol Biol* 184:99-105.

Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala'Aldeen D (2004) Type V protein secretion pathway: the autotransporter story. *Microbiol Mol Biol Rev* 68:692-744.

Hynes RO (2002) Integrins: bidirectional, allosteric signaling machines. *Cell* 110:673-87

I

IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (1994) Schistosomes, Liver Flukes and *Helicobacter Pylori*. Volume 61.

J

Jalaie M, Shanmugasundaram V (2006) Virtual screening: are we there yet? *Mini Rev Med Chem* 6:1159-1167.

Jiang J, Zhang X, Chen Y, Wu Y, Zhou ZH, Chang Z, Sui SF (2008) Activation of DegP chaperone-protease via formation of large cage-like oligomers upon binding to substrate proteins. *Proc Natl Acad Sci USA* 105:11939-11944.

Johnson M, Maggiora G (Herausgeber) (1990) Concepts and Applications of Molecular Similarity. Wiley, New York.

K

Kairys V, Fernandes MX, Gilson MK (2006) Screening drug-like compounds by docking to homology models: a systematic study. *J Chem Inf Model* 46:365-379.

Kavermann H, Burns BP, Angermüller K, Odenbreit S, Fischer W, et al. (2003) Identification and characterization of *Helicobacter pylori* genes essential for gastric colonization. *J Exp Med* 197: 813–822.

Kelley CT (1999) Iterative Methods for Optimization. Society for Industrial and Applied Mathematics, Philadelphia.

Kim DY, Kim KK (2005) Structure and function of HtrA family proteins, the key players in protein quality control. *J Biochem Mol Biol* 38:266-74.

Kohavi R (1995) A study of cross-validation and boot-strap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence 1137-1145.

Koshland DE (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci USA* 44:98-104.

Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T (2002) Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature*. 416:455-459.

Krojer T, Sawa J, Schäfer E, Saibil HR, Ehrmann M, Clausen T (2008) Structural basis for the regulated protease and chaperone function of DegP. *Nature* 453:885-890.

Kurumbail RG, Stevens AM, Gierse JK, McDonald JJ, Stegeman RA, Pak JY, Gildehaus D, Miyashiro JM, Penning TD, Seibert K, Isakson PC, Stallings WC (1996) Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. *Nature* 384:644-648.

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

L

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227:680-685.

Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, Yamaoka Y, Graham DY, Perez-Trallero E, Wadstrom T, Suerbaum S, Achtman M (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445:915-918.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* 46:3-26.

Liu H, Prugnolle F, Manica A, Balloux F (2006) A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* 79:230-237.

Lottspeich F, Zorbas H (Hrsg.) (1998) *Bioanalytik*. Spektrum Akademischer Verlag, Heidelberg.

Löwer M, Weydig C, Metzler D, Reuter A, Starzinski-Powitz A, et al. (2008) Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA. PLoS ONE 3:e3510.

Löwer M, Schneider G (2009) Prediction of Type III Secretion Signals in Genomes of Gram-negative Bacteria. PLoS ONE 4:e5917.

Lu H, Yamaoka Y, Graham DY (2005) *Helicobacter pylori* virulence factors: facts and fantasies. Curr Opin Gastroenterol 21:653—659.

M

Maggiora GM (2006) On outliers and activity cliffs - why QSAR often disappoints. J Chem Inf Model 46:1535.

Marshall BJ, Warren JR (1984) Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet 8390:1311-1315.

Mbulaiteye SM, Hisada M, El-Omar EM (2009) *Helicobacter Pylori* associated global gastric cancer burden. Front Biosci 14:1490-1504.

McGhie EJ, Brawn LC, Hume PJ, Humphreys D, Koronakis V (2009) Salmonella takes control: effector-driven manipulation of the host. Curr Opin Microbiol 12:117-24.

MOE (2008) The Molecular Operating Environment, Version 2008b. Chemical Computing Group, Montreal, Canada.

Moese S, Selbach M, Kwok T, Brinkmann V, König W, Meyer TF, Backert S (2004) *Helicobacter pylori* induces AGS cell motility and elongation via independent signaling pathways. Infect Immun 72:3646-3649.

Mühlhardt C (2006) Der Experimentator: Molekularbiologie/Genomics. Spektrum Akademischer Verlag, Heidelberg. S.144.

Murzin AG, Lesk AM, Chothia C (1994) Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol 236:1369-1381.

N

Nishimori I, Onishi S, Takeuchi H, Supuran CT (2008) The alpha and beta classes carbonic anhydrases from *Helicobacter pylori* as novel drug targets. *Curr Pharm Des.* 2008;14:622-630.

O

O'Boyle NM (2008) Book Review of Gnuplot in Action. *Journal of Chemical Information and Modeling* 48:2095-2095. <http://www.gnuplot.info/>

Oprea TI (2005) Chemoinformatics in Lead Discovery. In: *Chemoinformatics in Drug Discovery*. S. 25ff. Herausgeber: Tudor I Opera. Wiley VCH, Weinheim, New York.

P

Pallen MJ, Matzke NJ (2006) From The Origin of Species to the origin of bacterial flagella. *Nat Rev Microbiol* 4: 784-790.

Perona JJ, Craik CS (1995) Structural basis of substrate specificity in the serine proteases. *Protein Science.* 4:337-360.

R

Ramachandran GN, Sasiskharan V (1968) Conformation of polypeptides and proteins. *Advan Prot Chem* 23:283-437.

Rarey M, Lemmen C, Matter H (2005) Algorithmic Engines in Virtual Screening. In: *Chemoinformatics in Drug Discovery*. S. 25ff. Herausgeber: Tudor I Opera. Wiley VCH, Weinheim, New York.

Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN (2007) Virtual screening in drug discovery -- a computational perspective. *Curr Protein Pept Sci* 8:329-351.

Rechenberg I (1973) *Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart-

Rohrer SG, Baumann K (2009) Maximum Unbiased Validation (MUV) Datasets for Virtual Screening Based on PubChem Bioactivity Data. *J Chem Inf Model* 49:169–184.

Rokkas T, Pistiolas D, Sechopoulos P, Robotis I, Margantinis G (2007) Relationship between *Helicobacter pylori* infection and esophageal neoplasia: a meta-analysis. Clin Gastroenterol Hepatol. 5:1413-1417.

Roses AD (2008) Pharmacogenetics in drug discovery and development: a translational perspective. Nature Reviews Drug Discovery 7:807-817.

van Rossum G (1995) Python Reference Manual. CWI Report CS-R9525, Amsterdam.

Ryan KA, Karim N, Worku M, Penn CW, O'Toole PW (2005) *Helicobacter pylori* flagellar hook-filament transition is controlled by a FliK functional homolog encoded by the gene HP0906. J Bacteriol. 187:5742-50.

van Roy F, Berx G (2008) The cell-cell adhesion molecule E-cadherin. Cell Mol Life Sci 65:3756-88.

Rupp M, Schneider P, Schneider G (2009) Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. Journal of Computational Chemistry, accepted.

S

Sandkvist M (2001) Biology of type II secretion. Molecular Microbiology 40:271-283.

Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-Hopping" by topological pharmacophore search: A contribution to virtual screening. Angew Chemie Int Ed 38:2894-2896.

Schneider P, Schneider G (2003) Collection of bioactive reference compounds for focused library design. QSAR Comb Sci 22:713-718.

Schneider G, Baringhaus KH (2008) Molecular Design - Concepts and Applications. Wiley VCH, Weinheim, New York.

Schüller A, Fechner U, Renner S, Franke L, Weber L, Schneider G (2006) A Pseudo-Ligand Approach to Virtual Screening. Combinatorial Chemistry & High Throughput Screening 9:359-364.

Seifert MH, Lang M (2008) Essential factors for successful virtual screening. *Mini Rev Med Chem* 8:63-72.

Shao F (2008) Biochemical functions of Yersinia type III effectors. *Curr Opin Microbiol* 11:21-29.

Sheng YH, Nomura K, Whittam TS (2004) Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim Biophys Acta* 1694: 181-206.

Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen EL (2003) The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences* 43:493-500.

Steinhilber D, Schubert-Zsilavec M, Roth HJ (2005) *Medizinische Chemie*. Deutsche Apotheker Verlag, Stuttgart.

T

Tanrikulu Y, Nietert M, Scheffer U, Proschak E, Grabowski K, Schneider P, Weidlich M, Karas M, Goebel M, Schneider G (2007) Scaffold hopping by "fuzzy" pharmacophores and its application to RNA targets. *ChemBioChem* 8:1932-6.

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388: 539–547.

Truchon JF, Bayly CI (2007) Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling* 47:488-508.

U

Ugi I, Meyr R, Fetzer U, Steinbrückner C (1959) Versuche mit Isonitrilen. *Angew Chem* 71:386.

W

Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comp Chem* 21:1049-1074.

Warren LG, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A Critical Assessment of Docking Programs and Scoring Functions. *J Med Chem* 49:5912-5931.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189-1191.

Weisel M, Proschak E, Schneider G (2007) PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chemistry Central Journal* 1:7.

Weisgerber D (1997) Chemical abstracts service chemical registry system: History, scope, and impacts. *Journal of the American Society for Information Science* 48:349-360.

Weisstein EW (2009) Sigmoid Function. *MathWorld - A Wolfram Web Resource*. <http://mathworld.wolfram.com/SigmoidFunction.html>

Wermuth GC, Ganellin CR, Lindberg P, Mitscher LA (1998) Glossary of Terms used in Medicinal Chemistry. *Pure Appl Chem* 70:1129-1143.

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

Windle HJ, Kelleher D (1997) Identification and characterization of a metalloprotease activity from *Helicobacter pylori*. *Infect Immun* 65: 3132–3137.

Winkler DA, Burden FR (2002) Application of Neural networks to Large Dataset QSAR, Virtual Screening, and Library Design. In: Combinatorial Library Methods and Protocols: Methods and Protocols. S. 326. Herausgeber: Lisa Bellavance English. Humana Press, New York.

Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160-9.

9. Danksagung

Ich bedanke mich herzlich bei allen, die mich bei der Erstellung dieser Arbeit unterstützt haben. Mein besonderer Dank geht an:

- Professor Dr. Gisbert Schneider für die ausgezeichnete Betreuung dieser Arbeit als mein Doktorvater und die vielen Ideen und motivierenden Kommentare. Vielen Dank für die Zeit und alles was ich lernen durfte!
- PD Dr. Silja Wessler, für die Betreuung des experimentellen Teils dieser Arbeit und dafür, dass mir Laborarbeit (fast) doch noch Spaß macht.
- Professor Dr. Anna Starzinski-Powitz, für das Ermöglichen der Hospitation in ihrem Labor.
- Martin Weisel für PocketPicker und das Korrekturlesen dieser Arbeit.
- Yusuf Tanrikulu für die Bereitstellung und Hilfe mit LIQUID.
- Christiane Weydig für die viele, viele Hilfe und Geduld im Labor. Ich hoffe, das andauernde Stirn runzeln gibt keine Falten.
- Jan Alexander Hiß für die ganzen kleinen Nebenprojekte und Ideen für meine Vorhersagen.
- Dr. Alexander Schreiner für die Einführung in die Molekularbiologie.
- Swetlana Derksen, Kristina Grabowski und Dr. Ewgenij Proschak für die gute Stimmung und das angenehme Arbeitsklima im Büro.
- Norbert Dichter für die Betreuung der Hard- und Software.
- Dr. Andreas Reuter für die Massenspektroskopie.

- Das Modlab-Team für viele Antworten, die gemeinsamen Mittagessen und die phantastische Arbeitsatmosphäre. Namentlich und sicher lückenhaft: Uli Fechner, Lutz Franke, Tim Geppert, Alireza Givehchi, Volker Hähnke, Markus Hartenfeller, Bettina Hofmann, Alexander Klenner, Björn Krüger, Felix Reisen, Matthias Rupp, Brigitte Scheidemantel-Geiß, Michael Schmuker, Andreas Schüller, Tim Werner.

- Die Nachwuchsforschungsgruppe NG3 des Paul-Ehrlich-Instituts für das freundliche willkommen heißen und alle Unterstützung. Vielen Dank an Olivia Knauer, Gert Carra, Benjamin Hoy, Nadine Binai, Stefan Reese und Sabine Schneider.

- Das Beilstein-Institut zur Förderung der chemischen Wissenschaften und dem Center for Membrane Proteomics für die finanzielle Förderung.

- Meine Familie und meine Freunde für das Ertragen meiner trüben Stimmung und meiner Abwesenheit.

- Simone Färber (bald Löwer). Du weißt ja am besten für was und warum.

10. Anhang

Anhang A: Sequenzierungsdaten

Die sequenzierten Teile der in dieser Arbeit benutzten Nukleotidkonstrukte sind im Folgenden aufgeführt. Die komplementären Bereiche der Stopcodons bzw. das erste Codon des Gens *hp1018/19* sind grau hinterlegt.

pGEM-T easy hp1018/19 T7-Primer

CTCCcGGCCGCCATGGCGGGCCGGGAATTCGATTAAGGATCCGGCAATATCCAAATCCAGAGCATGCCCAAAGTTAAAGAGCGAGTG
AGTGTCCCTCTAAAGACGATACGATCTATCTTACCACGATTCTATTAAGGACTCTATTAAGGCGGTGGTGAATATCTCCACTGAAA
AGAAGATTAACAAATTTTATAGGTGGCGGTGTGTTAATGACCCCTTTTTCCAACAATTTTTGGGGATTGGGTGGCATGATTCC
TAAAGAAAGAAATGGAAAGGCTTTAGGCAGCGCGTAATCATTTCTAAAGACGGCTATATTGTAATAATAACCATGTGATTGATGGC
GCGGATAAGATTAAAGTTACCATTCCAGGGAGCAATAAAGAATATCCGCCACTCTAGTAGGCACCGATTCTGAAAGCGATTAGCGG
TGATTCGCATCACTAAAGACAATCTGCCACGATCAAATCTCTGATTCTAATGATATTTAGTGGGCGATTGGTTTTGCGATTGG
TAACCCTTTTGGCGTGGGCGAAAGCGTTACGCAAGGCATTGTTTCAGCGCTCAATAAAAGCGGGATTGGGATCAACAGCTATGAGAA
TTCATTCAAACAGACGCTTCCATCAATCCTGGAAATTCGGCGGGCGCTTTAATTGATAGCCGTGGAGGGTTAGTGGGGATTAATACCG
C

pGEM-T easy hp1018/19 SP6-Primer

TcCaACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCAGGCGCGCGAATTCAGTGTGATTAAGAATTCGACCCACCCCTATCAT
TTCACCAAAATGATCCTATAACCTTGATTCAGTCTAAAACCTAAGAATCGTTTGGGTTTGCCTTTATACTTTCTAAAGCATGGTTAA
AATCCGCAACGCTTTTAACTTCAACCTCTCAATTTTTGTGATAATGTTACCTTGCCATAATCCGGCTTGCTCTGCTGGGGATTTTT
ATTCAGTTGAGAGACTAAAACCCCTTGAACATCATCGCTCAAACGCATAGACCTTTGGTTTCTTGAGTTAAATCTTCACTTGAAGC
CCGTTCAATTTGGCCTTGGCGCCGTTTTGAGCAGAAATGGTTTTCTTTTTGTTAGGGTTTTCTTTCAGCTAGAGTGAGGGTGAAG
CGGTTCTTTTTGTCTCTAATGACTTTTAAAGTTACTCTTTGATTGGGTAGCATGGAGCCGATTAGATTTCTTAATCATTGCTGTT
TTTTAACCTTTTTCCCATTTGACTTCGGTGATCAAATCCACACCAAAATCCCTGCTTTTTAGCCGGAGAGTCTTTTTCTACGCTAATG
ACTACCGCCCTTCTTTGTTGTCTATAAGAATTTTGCAATCGCCACTCAAATCTTGCAAGCCCACGCCAAGTAACCTCTTTCATCT
TAC

pGEX-6P-1 hp1018/19 5'-Primer

AATCGGATCTGGAAGTTCTGTTCCAGGGGCCCTGGGATCCGGCAATATCCAAATCCAGAGCATGCCCAAAGTTAAAGAGCGAGTGAG
TGTCCCTCTAAAGACGATACGATCTATCTTACCACGATTCTATTAAGGACTCTATTAAGGCGGTGGTGAATATCTCCACTGAAAAG
AAGATTAACAAATTTTATAGGTGGCGGTGTGTTAATGACCCCTTTTTCCAACAATTTTTGGGGATTGGGTGGCATGATTCTTA
AAGAAAGAAATGGAAAGGCTTTAGGCAGCGCGTAATCATTTCTAAAGACGGCTATATTGTAATAATAACCATGTGATTGATGGCGC
GGATAAGATTAAAGTTACCATTCCAGGGAGCAATAAAGAATATTCGCCACTCTAGTAGGCACCGATTCTGAAAGCGATTTAGCGGTG
ATTCGCATCACTAAAGACAATCTGCCACGATCAAATCTCTGATTCTAATGATATTTAGTGGGCGATTGGTTTTGCGATTGGTA
ACCCTTTTGGCGTGGGCGAAAGCGTTACGCAAGGCATTGTTTCAGCGCTCAATAAAAGCGGGATTGGGATCAACAGCTATGAGAA
CATTCAAACAGACGCTTCCATCAATCCTGGAAATTCGGCGGGCGCTTTAATTGATAGCCGTGGAGGGTTAGTGGGGATTAATACCGCT
A

pGEX-6P-1 hp1018/19 3'-Primer

CGAaACGCGCGAGGCAGATCGTCAGTCAGTCACGATGCGGCGGCTCGAGTCGACCCGGGAATTCAGTGTGATTAAGAATTCGACCCA
CCCCATTCATTTTACCAAAATGATCCTATAACCTTGATTCAGTCTAAAACCTAAGAATCGTTTGGGTTTGCCTTTATACTTTCTAAA
GCATGGTTAAAATCCGCAACGCTTTTAACTTCAACCTCTCAATTTTTGTGATAATGTTACCTTGCCATAATCCGGCTTGCTCTGCTG
GGGAATTTTCACTTGAAGACTAAAACCCCTTGAACATCATCGCTCAAACGCATAGACCTTTGGTTTCTTGAGTTAAATCTTC
TACTTGAAGCCCGTTCAATTTGGCCTTGGCGCCGTTTTGAGCAGAAATGGTTCTTTTTGTTAGGGTTTTCTTTTCAGCTAGAGTG
AGGGTGAAGCGCGTTCTTTTTGTCTCTAATGACTTTTAAAGTTACTCTTTGATTGGGTAGCATGGAGCCGATTAGATTTCTTAAT
CATTGCTGTTTTAACCTTTTTCCCATTTGACTTCGGTGATCAAATCCACACCAAAATCCCTGCTTTTTAGCCGGAGAGTCTTTTT
TACGCTAATGACTACCGCCCTTCTTTGTTGTCTATAAGAATTTTGCAATCGCCACTCAAATCTTGCAAGCCCACGCCAAGTAACCT
CT

Anhang B: Komplette Ergebnisse des retrospektiven virtuellen Screenings

Die folgenden Tabellen enthalten die kompletten Ergebnisse des retrospektiven virtuellen Screenings auf verschiedenen Wirkstoffzielen mit den COBRA-, MUV- und UGI-Datenbanken. Die Werte sind über die zehn Durchgänge der Kreuzvalidierung gemittelt.

Tabelle A1. Ergebnisse des retrospektiven virtuellen Screenings der COBRA Datenbank mit der euklidischen Metrik als Unähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB ID	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung			Blockskalierung			Skalierung auf Eins					
				ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)
ACE	1o86	1	1,5	0,34	(0,01)	0,00	(0,00)	0,50	(0,01)	0,01	(0,00)	0,24	(0,01)	0,00	(0,00)
	1o86	1	1,9	0,39	(0,01)	0,00	(0,00)	0,47	(0,01)	0,00	(0,00)	0,28	(0,01)	0,00	(0,00)
	1o86	1	4	0,35	(0,01)	0,00	(0,00)	0,47	(0,01)	0,01	(0,00)	0,26	(0,01)	0,00	(0,00)
COX-2	3pgh	1	1,5	0,46	(0,01)	0,03	(0,01)	0,40	(0,01)	0,01	(0,00)	0,72	(0,01)	0,16	(0,01)
	3pgh	1	1,9	0,48	(0,01)	0,06	(0,01)	0,37	(0,01)	0,00	(0,00)	0,54	(0,01)	0,02	(0,00)
	3pgh	1	4	0,44	(0,01)	0,06	(0,01)	0,44	(0,01)	0,02	(0,00)	0,50	(0,01)	0,02	(0,00)
	3pgh	1 (Teil) ³⁾	1,5	0,53	(0,02)	0,13	(0,01)	0,37	(0,01)	0,04	(0,01)	0,75	(0,01)	0,28	(0,02)
	3pgh	1 (Teil) ³⁾	1,9	0,57	(0,01)	0,16	(0,01)	0,47	(0,01)	0,15	(0,01)	0,80	(0,01)	0,35	(0,01)
	3pgh	1 (Teil) ³⁾	4	0,55	(0,01)	0,19	(0,01)	0,60	(0,02)	0,27	(0,02)	0,81	(0,01)	0,36	(0,01)
	5cox	3	1,5	0,39	(0,01)	0,01	(0,00)	0,31	(0,01)	0,01	(0,00)	0,47	(0,01)	0,01	(0,00)
	5cox	3	1,9	0,41	(0,01)	0,02	(0,00)	0,37	(0,01)	0,00	(0,00)	0,50	(0,01)	0,01	(0,00)
	5cox	3	4	0,37	(0,01)	0,02	(0,00)	0,45	(0,01)	0,03	(0,00)	0,44	(0,01)	0,01	(0,00)
DHFR	1kmv	1	1,5	0,63	(0,01)	0,11	(0,01)	0,61	(0,02)	0,05	(0,01)	0,65	(0,01)	0,13	(0,02)
	1kmv	1	1,9	0,60	(0,02)	0,19	(0,01)	0,64	(0,01)	0,06	(0,01)	0,67	(0,02)	0,28	(0,03)
	1kmv	1	4	0,59	(0,01)	0,11	(0,01)	0,71	(0,01)	0,07	(0,01)	0,65	(0,01)	0,23	(0,02)
fXa	2bok	1	1,5	0,70	(0,01)	0,07	(0,00)	0,62	(0,01)	0,07	(0,00)	0,35	(0,01)	0,00	(0,00)
	2bok	1	1,9	0,64	(0,01)	0,04	(0,00)	0,69	(0,01)	0,12	(0,01)	0,32	(0,01)	0,00	(0,00)
	2bok	1	4	0,64	(0,01)	0,05	(0,00)	0,62	(0,01)	0,08	(0,01)	0,31	(0,01)	0,00	(0,00)
	2bok	1, 6, 13, 17	1,5	0,78	(0,00)	0,15	(0,01)	0,69	(0,01)	0,09	(0,01)	0,82	(0,00)	0,17	(0,01)
	2bok	1, 6, 13, 17	1,9	0,73	(0,00)	0,09	(0,00)	0,68	(0,01)	0,07	(0,01)	0,78	(0,01)	0,21	(0,01)
	2bok	1, 6, 13, 17	4	0,74	(0,01)	0,10	(0,01)	0,57	(0,01)	0,03	(0,00)	0,82	(0,00)	0,21	(0,01)
PPAR γ	1zgy	1	1,5	0,57	(0,02)	0,06	(0,01)	0,53	(0,02)	0,03	(0,01)	0,54	(0,02)	0,06	(0,01)
	1zgy	1	1,9	0,56	(0,02)	0,04	(0,01)	0,55	(0,02)	0,06	(0,01)	0,53	(0,02)	0,07	(0,01)
	1zgy	1	4	0,52	(0,02)	0,04	(0,01)	0,54	(0,01)	0,03	(0,01)	0,51	(0,02)	0,07	(0,01)
Trypsin	1dpo	1	1,5	0,61	(0,02)	0,05	(0,02)	0,65	(0,02)	0,05	(0,01)	0,46	(0,04)	0,01	(0,00)
	1dpo	1	1,9	0,65	(0,03)	0,08	(0,01)	0,67	(0,02)	0,05	(0,01)	0,53	(0,03)	0,02	(0,00)
	1dpo	1	4	0,61	(0,04)	0,05	(0,01)	0,60	(0,02)	0,02	(0,00)	0,49	(0,02)	0,01	(0,00)
Tryptase	2fpz	2, 4, 17	1,5	0,71	(0,03)	0,17	(0,04)	0,65	(0,03)	0,09	(0,03)	0,74	(0,02)	0,15	(0,02)
	2fpz	2, 4, 17	1,9	0,72	(0,02)	0,22	(0,03)	0,71	(0,02)	0,26	(0,05)	0,74	(0,03)	0,18	(0,02)
	2fpz	2, 4, 17	4	0,68	(0,03)	0,21	(0,05)	0,70	(0,03)	0,13	(0,03)	0,67	(0,03)	0,13	(0,02)
UPA	2o8t	1	1,5	0,64	(0,01)	0,10	(0,01)	0,51	(0,02)	0,04	(0,02)	0,65	(0,03)	0,15	(0,02)
	2o8t	1	1,9	0,74	(0,02)	0,23	(0,03)	0,58	(0,02)	0,04	(0,01)	0,78	(0,02)	0,27	(0,01)
	2o8t	1	4	0,67	(0,02)	0,18	(0,02)	0,66	(0,01)	0,18	(0,02)	0,72	(0,02)	0,26	(0,02)

1) Die Nummerierung entspricht der Rheifolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

3) Bindetaschenvorhersage wurde manuell bearbeitet (siehe Text).

Tabelle A2. Ergebnisse des retrospektiven virtuellen Screenings der COBRA Datenbank mit der Manhattanmetrik als Unähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB ID	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung			Blockskalierung			Skalierung auf Eins					
				ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)
ACE	1o86	1	1,5	0,32	(0,01)	0,00	(0,00)	0,40	(0,01)	0,00	(0,00)	0,26	(0,01)	0,00	(0,00)
	1o86	1	1,9	0,42	(0,01)	0,00	(0,00)	0,42	(0,01)	0,00	(0,00)	0,35	(0,01)	0,00	(0,00)
	1o86	1	4	0,33	(0,01)	0,00	(0,00)	0,40	(0,02)	0,01	(0,00)	0,28	(0,01)	0,00	(0,00)
COX-2	3pgh	1	1,5	0,62	(0,01)	0,04	(0,00)	0,49	(0,01)	0,00	(0,00)	0,74	(0,01)	0,18	(0,01)
	3pgh	1	1,9	0,74	(0,01)	0,19	(0,01)	0,53	(0,00)	0,00	(0,00)	0,69	(0,01)	0,08	(0,00)
	3pgh	1	4	0,77	(0,00)	0,15	(0,01)	0,57	(0,01)	0,03	(0,00)	0,66	(0,01)	0,06	(0,00)
	3pgh	1 (Teil) ³⁾	1,5	0,69	(0,01)	0,11	(0,01)	0,42	(0,01)	0,04	(0,01)	0,76	(0,01)	0,21	(0,01)
	3pgh	1 (Teil) ³⁾	1,9	0,81	(0,00)	0,28	(0,01)	0,73	(0,01)	0,30	(0,02)	0,81	(0,01)	0,40	(0,02)
	3pgh	1 (Teil) ³⁾	4	0,84	(0,01)	0,29	(0,01)	0,82	(0,00)	0,36	(0,01)	0,85	(0,00)	0,44	(0,01)
	5cox	3	1,5	0,50	(0,01)	0,01	(0,00)	0,33	(0,01)	0,00	(0,00)	0,50	(0,01)	0,01	(0,00)
	5cox	3	1,9	0,63	(0,00)	0,06	(0,00)	0,53	(0,01)	0,02	(0,00)	0,62	(0,01)	0,04	(0,00)
	5cox	3	4	0,61	(0,01)	0,03	(0,00)	0,61	(0,01)	0,04	(0,01)	0,59	(0,01)	0,02	(0,00)
DHFR	1kmv	1	1,5	0,66	(0,01)	0,10	(0,01)	0,71	(0,01)	0,15	(0,01)	0,70	(0,01)	0,13	(0,01)
	1kmv	1	1,9	0,63	(0,02)	0,14	(0,01)	0,75	(0,01)	0,24	(0,01)	0,67	(0,02)	0,14	(0,01)
	1kmv	1	4	0,57	(0,01)	0,05	(0,01)	0,78	(0,01)	0,17	(0,01)	0,62	(0,01)	0,06	(0,01)
fXa	2bok	1	1,5	0,52	(0,01)	0,02	(0,00)	0,52	(0,00)	0,04	(0,00)	0,30	(0,00)	0,00	(0,00)
	2bok	1	1,9	0,41	(0,01)	0,00	(0,00)	0,64	(0,01)	0,05	(0,00)	0,29	(0,00)	0,00	(0,00)
	2bok	1	4	0,37	(0,00)	0,00	(0,00)	0,53	(0,00)	0,03	(0,00)	0,27	(0,01)	0,00	(0,00)
	2bok	1, 6, 13, 17	1,5	0,78	(0,00)	0,17	(0,01)	0,75	(0,01)	0,13	(0,00)	0,81	(0,01)	0,19	(0,01)
	2bok	1, 6, 13, 17	1,9	0,67	(0,01)	0,12	(0,01)	0,76	(0,00)	0,10	(0,00)	0,70	(0,01)	0,24	(0,01)
	2bok	1, 6, 13, 17	4	0,71	(0,00)	0,11	(0,00)	0,70	(0,01)	0,06	(0,00)	0,78	(0,00)	0,18	(0,01)
PPAR γ	1zgy	1	1,5	0,56	(0,01)	0,03	(0,01)	0,52	(0,02)	0,03	(0,01)	0,47	(0,01)	0,02	(0,00)
	1zgy	1	1,9	0,61	(0,01)	0,05	(0,01)	0,58	(0,02)	0,05	(0,01)	0,54	(0,01)	0,07	(0,02)
	1zgy	1	4	0,58	(0,02)	0,04	(0,01)	0,52	(0,02)	0,04	(0,01)	0,49	(0,02)	0,05	(0,01)
Trypsin	1dpo	1	1,5	0,44	(0,01)	0,01	(0,00)	0,55	(0,01)	0,03	(0,01)	0,34	(0,02)	0,00	(0,00)
	1dpo	1	1,9	0,48	(0,02)	0,00	(0,00)	0,53	(0,02)	0,01	(0,00)	0,39	(0,02)	0,00	(0,00)
	1dpo	1	4	0,32	(0,02)	0,00	(0,00)	0,47	(0,01)	0,00	(0,00)	0,29	(0,02)	0,00	(0,00)
Tryptase	2fpz	1	1,5	0,72	(0,03)	0,16	(0,03)	0,73	(0,01)	0,13	(0,03)	0,72	(0,01)	0,16	(0,04)
	2fpz	1	1,9	0,73	(0,04)	0,20	(0,03)	0,72	(0,02)	0,23	(0,03)	0,70	(0,02)	0,10	(0,02)
	2fpz	1	4	0,60	(0,03)	0,14	(0,02)	0,71	(0,01)	0,10	(0,02)	0,56	(0,03)	0,07	(0,02)
UPA	2o8t	2, 4, 17	1,5	0,66	(0,02)	0,12	(0,02)	0,58	(0,02)	0,09	(0,01)	0,64	(0,03)	0,16	(0,02)
	2o8t	2, 4, 17	1,9	0,75	(0,02)	0,20	(0,02)	0,67	(0,02)	0,10	(0,01)	0,75	(0,02)	0,25	(0,02)
	2o8t	2, 4, 17	4	0,66	(0,03)	0,28	(0,03)	0,73	(0,02)	0,22	(0,02)	0,66	(0,02)	0,29	(0,02)

1) Die Nummerierung entspricht der Rheifolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

3) Bindetaschenvorhersage wurde manuell bearbeitet, (siehe Text).

Tabelle A3. Ergebnisse des retrospektiven virtuellen Screenings der COBRA Datenbank mit dem Carbóindex als Ähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB ID	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung				Skalierung auf Eins					
				ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)
ACE	1o86	1	1,5	0,41	(0,01)	0,01	(0,00)	0,43	(0,01)	0,01	(0,01)	0,41	(0,01)	0,01	(0,00)
	1o86	1	1,9	0,44	(0,00)	0,00	(0,00)	0,32	(0,01)	0,00	(0,00)	0,44	(0,01)	0,00	(0,00)
	1o86	1	4	0,42	(0,01)	0,00	(0,00)	0,35	(0,01)	0,00	(0,01)	0,42	(0,01)	0,00	(0,00)
COX-2	3pgh	1	1,5	0,46	(0,01)	0,02	(0,00)	0,51	(0,01)	0,03	(0,00)	0,46	(0,01)	0,02	(0,00)
	3pgh	1	1,9	0,42	(0,01)	0,01	(0,00)	0,40	(0,01)	0,01	(0,00)	0,42	(0,01)	0,01	(0,00)
	3pgh	1	4	0,41	(0,01)	0,01	(0,00)	0,38	(0,00)	0,01	(0,00)	0,41	(0,01)	0,01	(0,00)
	3pgh	1 (Teil) ³⁾	1,5	0,48	(0,01)	0,21	(0,00)	0,37	(0,01)	0,05	(0,00)	0,49	(0,01)	0,21	(0,01)
	3pgh	1 (Teil) ³⁾	1,9	0,51	(0,01)	0,21	(0,01)	0,57	(0,01)	0,17	(0,01)	0,51	(0,01)	0,20	(0,01)
	3pgh	1 (Teil) ³⁾	4	0,52	(0,01)	0,21	(0,01)	0,68	(0,01)	0,27	(0,01)	0,51	(0,01)	0,21	(0,00)
	5cox	3	1,5	0,22	(0,00)	0,00	(0,00)	0,26	(0,01)	0,01	(0,00)	0,22	(0,00)	0,00	(0,00)
	5cox	3	1,9	0,25	(0,01)	0,00	(0,00)	0,44	(0,01)	0,01	(0,00)	0,25	(0,00)	0,00	(0,00)
	5cox	3	4	0,24	(0,00)	0,01	(0,00)	0,48	(0,01)	0,04	(0,01)	0,24	(0,00)	0,00	(0,00)
DHFR	1kmv	1	1,5	0,63	(0,01)	0,14	(0,01)	0,57	(0,01)	0,05	(0,01)	0,63	(0,01)	0,14	(0,01)
	1kmv	1	1,9	0,62	(0,01)	0,30	(0,01)	0,64	(0,01)	0,07	(0,00)	0,63	(0,01)	0,30	(0,01)
	1kmv	1	4	0,62	(0,01)	0,26	(0,01)	0,67	(0,01)	0,08	(0,01)	0,62	(0,01)	0,27	(0,01)
fXa	2bok	1	1,5	0,45	(0,00)	0,02	(0,00)	0,49	(0,00)	0,05	(0,00)	0,44	(0,00)	0,03	(0,00)
	2bok	1	1,9	0,65	(0,00)	0,06	(0,00)	0,59	(0,01)	0,03	(0,00)	0,65	(0,00)	0,06	(0,00)
	2bok	1	4	0,66	(0,00)	0,06	(0,00)	0,57	(0,01)	0,03	(0,00)	0,66	(0,01)	0,07	(0,00)
	2bok	1, 6, 13, 17	1,5	0,81	(0,00)	0,16	(0,00)	0,63	(0,00)	0,04	(0,00)	0,81	(0,00)	0,16	(0,01)
	2bok	1, 6, 13, 17	1,9	0,81	(0,00)	0,21	(0,01)	0,63	(0,00)	0,03	(0,00)	0,81	(0,00)	0,21	(0,01)
	2bok	1, 6, 13, 17	4	0,82	(0,00)	0,20	(0,01)	0,49	(0,00)	0,02	(0,00)	0,82	(0,00)	0,20	(0,01)
PPAR γ	1zgy	1	1,5	0,53	(0,01)	0,05	(0,00)	0,46	(0,01)	0,03	(0,00)	0,54	(0,01)	0,05	(0,01)
	1zgy	1	1,9	0,53	(0,01)	0,05	(0,01)	0,51	(0,01)	0,03	(0,00)	0,53	(0,01)	0,06	(0,01)
	1zgy	1	4	0,51	(0,02)	0,06	(0,02)	0,44	(0,02)	0,02	(0,00)	0,51	(0,01)	0,06	(0,01)
Trypsin	1dpo	1	1,5	0,68	(0,01)	0,04	(0,01)	0,65	(0,01)	0,04	(0,01)	0,68	(0,01)	0,04	(0,01)
	1dpo	1	1,9	0,70	(0,02)	0,05	(0,01)	0,67	(0,01)	0,04	(0,01)	0,69	(0,01)	0,05	(0,01)
	1dpo	1	4	0,68	(0,01)	0,05	(0,01)	0,64	(0,02)	0,02	(0,00)	0,68	(0,01)	0,04	(0,01)
Tryptase	2fpz	1	1,5	0,78	(0,01)	0,21	(0,02)	0,65	(0,01)	0,09	(0,01)	0,79	(0,02)	0,20	(0,04)
	2fpz	1	1,9	0,81	(0,02)	0,28	(0,04)	0,71	(0,01)	0,25	(0,02)	0,80	(0,01)	0,28	(0,02)
	2fpz	1	4	0,78	(0,01)	0,24	(0,02)	0,72	(0,02)	0,13	(0,02)	0,78	(0,02)	0,24	(0,04)
UPA	2o8t	2, 4, 17	1,5	0,68	(0,02)	0,13	(0,01)	0,47	(0,01)	0,04	(0,01)	0,68	(0,01)	0,14	(0,01)
	2o8t	2, 4, 17	1,9	0,74	(0,02)	0,25	(0,01)	0,63	(0,01)	0,07	(0,01)	0,74	(0,01)	0,27	(0,01)
	2o8t	2, 4, 17	4	0,71	(0,02)	0,23	(0,01)	0,72	(0,01)	0,20	(0,02)	0,70	(0,02)	0,23	(0,02)

1) Die Nummerierung entspricht der Rheifolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

3) Bindetaschenvorhersage wurde manuell bearbeitet (siehe Text).

Tabelle A4. Ergebnisse des retrospektiven virtuellen Screenings der UGI Datenbank mit der euklidischen Metrik als Unähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB ID	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung		Skalierung auf Eins	
				ROCAUC (σ)	BEDROC (σ)	ROCAUC (σ)	BEDROC (σ)	ROCAUC (σ)	BEDROC (σ)
fXa	2bok	1, 6, 13, 17	1,5	0,60 (0,00)	0,17 (0,00)	0,53 (0,00)	0,16 (0,00)	0,60 (0,00)	0,17 (0,00)
	2bok	1, 6, 13, 18	1,9	0,59 (0,00)	0,16 (0,00)	0,53 (0,00)	0,13 (0,00)	0,61 (0,00)	0,15 (0,00)
	2bok	1, 6, 13, 19	4	0,55 (0,00)	0,14 (0,00)	0,58 (0,00)	0,13 (0,00)	0,58 (0,00)	0,13 (0,00)
UPA	2o8t	1	1,5	0,64 (0,00)	0,20 (0,00)	0,48 (0,00)	0,12 (0,00)	0,64 (0,00)	0,28 (0,00)
	2o8t	1	1,9	0,65 (0,00)	0,21 (0,00)	0,53 (0,00)	0,14 (0,00)	0,64 (0,00)	0,29 (0,00)
	2o8t	1	4	0,64 (0,00)	0,20 (0,00)	0,55 (0,00)	0,11 (0,00)	0,63 (0,00)	0,29 (0,00)
Trypsin	1dpo	1	1,5	0,55 (0,00)	0,06 (0,00)	0,52 (0,00)	0,10 (0,00)	0,52 (0,01)	0,09 (0,01)
	1dpo	1	1,9	0,54 (0,00)	0,05 (0,00)	0,49 (0,01)	0,06 (0,00)	0,50 (0,01)	0,09 (0,01)
	1dpo	1	4	0,48 (0,00)	0,04 (0,00)	0,54 (0,00)	0,06 (0,00)	0,44 (0,01)	0,08 (0,01)
Tryptase	2fpz	2, 4, 17	1,5	0,67 (0,00)	0,59 (0,00)	0,54 (0,00)	0,34 (0,01)	0,70 (0,00)	0,65 (0,00)
	2fpz	2, 4, 18	1,9	0,67 (0,00)	0,59 (0,00)	0,54 (0,00)	0,39 (0,00)	0,70 (0,00)	0,62 (0,00)
	2fpz	2, 4, 19	4	0,66 (0,00)	0,57 (0,00)	0,57 (0,00)	0,33 (0,00)	0,70 (0,00)	0,58 (0,00)

1) Die Nummerierung entspricht der Rheinfolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

Tabelle A5. Ergebnisse des retrospektiven virtuellen Screenings der UGI Datenbank mit der Manhattanmetrik als Unähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB ID	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung		Skalierung auf Eins	
				ROCAUC (σ)	BEDROC (σ)	ROCAUC (σ)	BEDROC (σ)	ROCAUC (σ)	BEDROC (σ)
fXa	2bok	1, 6, 13, 17	1,5	0,60 (0,00)	0,18 (0,00)	0,56 (0,00)	0,16 (0,00)	0,60 (0,00)	0,17 (0,00)
	2bok	1, 6, 13, 18	1,9	0,59 (0,00)	0,16 (0,00)	0,57 (0,00)	0,13 (0,00)	0,62 (0,00)	0,16 (0,00)
	2bok	1, 6, 13, 19	4	0,54 (0,00)	0,14 (0,00)	0,59 (0,00)	0,11 (0,00)	0,58 (0,00)	0,12 (0,00)
UPA	2o8t	1	1,5	0,60 (0,00)	0,09 (0,01)	0,58 (0,00)	0,13 (0,00)	0,56 (0,00)	0,09 (0,00)
	2o8t	1	1,9	0,57 (0,00)	0,04 (0,00)	0,49 (0,00)	0,06 (0,00)	0,53 (0,00)	0,07 (0,01)
	2o8t	1	4	0,46 (0,01)	0,04 (0,00)	0,49 (0,00)	0,06 (0,00)	0,43 (0,00)	0,06 (0,00)
Trypsin	1dpo	1	1,5	0,67 (0,00)	0,58 (0,00)	0,62 (0,00)	0,44 (0,00)	0,69 (0,00)	0,65 (0,00)
	1dpo	1	1,9	0,67 (0,00)	0,60 (0,00)	0,60 (0,00)	0,42 (0,00)	0,67 (0,00)	0,67 (0,00)
	1dpo	1	4	0,66 (0,00)	0,52 (0,01)	0,60 (0,00)	0,33 (0,00)	0,65 (0,00)	0,59 (0,00)
Tryptase	2fpz	2, 4, 17	1,5	0,61 (0,00)	0,19 (0,00)	0,52 (0,00)	0,14 (0,00)	0,61 (0,00)	0,25 (0,00)
	2fpz	2, 4, 18	1,9	0,61 (0,00)	0,18 (0,00)	0,59 (0,00)	0,16 (0,00)	0,61 (0,00)	0,26 (0,00)
	2fpz	2, 4, 19	4	0,58 (0,00)	0,16 (0,00)	0,57 (0,00)	0,11 (0,00)	0,58 (0,00)	0,24 (0,00)

1) Die Nummerierung entspricht der Rheinfolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

Tabelle A6. Ergebnisse des retrospektiven virtuellen Screenings der UGI Datenbank mit dem Carbóindex als Ähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB ID	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung		Skalierung auf Eins	
				ROCAUC (σ)	BEDROC (σ)	ROCAUC (σ)	BEDROC (σ)	ROCAUC (σ)	BEDROC (σ)
fXa	2bok	1, 6, 13, 17	1,5	0,59 (0,00)	0,17 (0,00)	0,53 (0,00)	0,12 (0,00)	0,59 (0,00)	0,17 (0,00)
	2bok	1, 6, 13, 18	1,9	0,62 (0,00)	0,16 (0,00)	0,54 (0,00)	0,13 (0,00)	0,61 (0,00)	0,16 (0,00)
	2bok	1, 6, 13, 19	4	0,59 (0,00)	0,13 (0,00)	0,57 (0,00)	0,11 (0,00)	0,59 (0,00)	0,13 (0,00)
UPA	2o8t	1	1,5	0,48 (0,00)	0,09 (0,00)	0,50 (0,00)	0,11 (0,00)	0,48 (0,00)	0,08 (0,00)
	2o8t	1	1,9	0,52 (0,00)	0,09 (0,00)	0,50 (0,00)	0,07 (0,00)	0,53 (0,00)	0,09 (0,00)
	2o8t	1	4	0,50 (0,00)	0,09 (0,00)	0,51 (0,00)	0,06 (0,00)	0,50 (0,00)	0,09 (0,00)
Trypsin	1dpo	1	1,5	0,70 (0,00)	0,58 (0,00)	0,55 (0,00)	0,31 (0,00)	0,70 (0,00)	0,58 (0,00)
	1dpo	1	1,9	0,71 (0,00)	0,56 (0,00)	0,57 (0,00)	0,37 (0,00)	0,71 (0,00)	0,56 (0,00)
	1dpo	1	4	0,71 (0,00)	0,53 (0,00)	0,59 (0,00)	0,37 (0,00)	0,71 (0,00)	0,53 (0,00)
Tryptase	2fpz	2, 4, 17	1,5	0,68 (0,00)	0,22 (0,00)	0,47 (0,00)	0,12 (0,00)	0,68 (0,00)	0,22 (0,00)
	2fpz	2, 4, 18	1,9	0,70 (0,00)	0,23 (0,00)	0,56 (0,00)	0,13 (0,00)	0,70 (0,00)	0,23 (0,00)
	2fpz	2, 4, 19	4	0,68 (0,00)	0,23 (0,00)	0,56 (0,00)	0,10 (0,00)	0,68 (0,00)	0,23 (0,00)

1) Die Nummerierung entspricht der Rheinfolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

Tabelle A7. Ergebnisse des retrospektiven virtuellen Screenings der MUV Datenbank mit der euklidischen Metrik als Unähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB Nr.	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung				Skalierung auf Eins					
				ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)
CathepsinG	1cgh	1	1,5	0,69	(0,03)	0,19	(0,03)	0,61	(0,02)	0,07	(0,01)	0,46	(0,03)	0,10	(0,02)
	1cgh	1	1,9	0,60	(0,02)	0,13	(0,02)	0,39	(0,03)	0,01	(0,01)	0,46	(0,03)	0,06	(0,02)
	1cgh	1	4	0,70	(0,04)	0,15	(0,02)	0,49	(0,02)	0,02	(0,01)	0,51	(0,03)	0,08	(0,02)
Eph	3ckh	2	1,5	0,57	(0,02)	0,05	(0,02)	0,57	(0,03)	0,03	(0,01)	0,52	(0,02)	0,04	(0,01)
	3ckh	2	1,9	0,54	(0,03)	0,07	(0,02)	0,59	(0,02)	0,03	(0,01)	0,53	(0,03)	0,05	(0,01)
	3ckh	2	4	0,56	(0,02)	0,05	(0,01)	0,55	(0,03)	0,06	(0,01)	0,54	(0,02)	0,07	(0,01)
ER-a	1xpc	1	1,5	0,72	(0,02)	0,28	(0,03)	0,54	(0,01)	0,03	(0,01)	0,46	(0,03)	0,07	(0,02)
	1xpc	1	1,9	0,70	(0,02)	0,33	(0,02)	0,68	(0,01)	0,23	(0,03)	0,60	(0,04)	0,18	(0,03)
	1xpc	1	4	0,75	(0,02)	0,35	(0,04)	0,66	(0,02)	0,06	(0,01)	0,63	(0,01)	0,18	(0,01)
ER-b	1qkm	1	1,5	0,58	(0,02)	0,10	(0,02)	0,53	(0,04)	0,07	(0,02)	0,61	(0,01)	0,10	(0,03)
	1qkm	1	1,9	0,56	(0,02)	0,07	(0,02)	0,55	(0,03)	0,09	(0,02)	0,63	(0,03)	0,10	(0,02)
	1qkm	1	4	0,58	(0,03)	0,08	(0,01)	0,50	(0,03)	0,06	(0,02)	0,63	(0,02)	0,10	(0,02)
FAK	1mp8	1	1,5	0,65	(0,03)	0,11	(0,02)	0,59	(0,04)	0,08	(0,01)	0,62	(0,04)	0,13	(0,02)
	1mp8	1	1,9	0,61	(0,04)	0,15	(0,02)	0,58	(0,04)	0,15	(0,03)	0,60	(0,02)	0,08	(0,01)
	1mp8	1	4	0,57	(0,04)	0,11	(0,02)	0,70	(0,02)	0,08	(0,01)	0,60	(0,04)	0,07	(0,02)
FXIa	1zsj	1	1,5	0,52	(0,04)	0,03	(0,01)	0,30	(0,03)	0,00	(0,00)	0,48	(0,02)	0,00	(0,00)
	1zsj	1	1,9	0,47	(0,02)	0,01	(0,01)	0,32	(0,01)	0,00	(0,00)	0,33	(0,02)	0,00	(0,00)
	1zsj	1	4	0,45	(0,02)	0,00	(0,00)	0,34	(0,02)	0,00	(0,00)	0,31	(0,02)	0,00	(0,00)
HIV-RT	2zd1	1	1,5	0,59	(0,01)	0,11	(0,02)	0,51	(0,02)	0,02	(0,00)	0,36	(0,02)	0,05	(0,01)
	2zd1	1	1,9	0,54	(0,02)	0,08	(0,01)	0,63	(0,01)	0,09	(0,01)	0,32	(0,02)	0,05	(0,01)
	2zd1	1	4	0,63	(0,02)	0,09	(0,01)	0,58	(0,02)	0,12	(0,02)	0,36	(0,02)	0,05	(0,02)
Hsp90	1uyl	1	1,5	0,74	(0,03)	0,12	(0,02)	0,53	(0,02)	0,01	(0,00)	0,65	(0,03)	0,21	(0,03)
	1uyl	1	1,9	0,66	(0,03)	0,13	(0,02)	0,61	(0,01)	0,03	(0,01)	0,65	(0,04)	0,16	(0,03)
	1uyl	1	4	0,68	(0,02)	0,13	(0,01)	0,64	(0,03)	0,24	(0,03)	0,68	(0,03)	0,18	(0,02)
PKA	2uzt	1	1,5	0,57	(0,03)	0,08	(0,02)	0,34	(0,03)	0,03	(0,01)	0,62	(0,02)	0,03	(0,01)
	2uzt	1	1,9	0,51	(0,03)	0,08	(0,01)	0,39	(0,04)	0,02	(0,01)	0,54	(0,03)	0,03	(0,01)
	2uzt	1	4	0,49	(0,03)	0,06	(0,02)	0,52	(0,03)	0,07	(0,01)	0,50	(0,02)	0,03	(0,01)
Rho-kinase2	2f2u	1	1,5	0,49	(0,02)	0,09	(0,02)	0,35	(0,03)	0,03	(0,01)	0,51	(0,02)	0,03	(0,01)
	2f2u	1	1,9	0,42	(0,04)	0,04	(0,02)	0,39	(0,02)	0,02	(0,01)	0,52	(0,02)	0,01	(0,01)
	2f2u	1	4	0,48	(0,03)	0,06	(0,02)	0,54	(0,02)	0,01	(0,01)	0,59	(0,01)	0,04	(0,01)

1) Die Nummerierung entspricht der Reihenfolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

Tabelle A8. Ergebnisse des retrospektiven virtuellen Screenings der MUV Datenbank mit der Manhattanmetrik als Unähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB Nr.	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung				Skalierung auf Eins					
				ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)
CathepsinG	1cgh	1	1,5	0,62	(0,03)	0,16	(0,03)	0,69	(0,02)	0,10	(0,03)	0,49	(0,04)	0,14	(0,03)
	1cgh	1	1,9	0,52	(0,04)	0,16	(0,03)	0,62	(0,03)	0,08	(0,01)	0,46	(0,03)	0,10	(0,03)
	1cgh	1	4	0,67	(0,02)	0,11	(0,02)	0,41	(0,02)	0,01	(0,01)	0,47	(0,04)	0,06	(0,02)
Eph	3ckh	2	1,5	0,53	(0,03)	0,06	(0,02)	0,49	(0,01)	0,02	(0,01)	0,50	(0,01)	0,05	(0,01)
	3ckh	2	1,9	0,52	(0,03)	0,08	(0,02)	0,57	(0,02)	0,07	(0,01)	0,52	(0,03)	0,02	(0,01)
	3ckh	2	4	0,52	(0,05)	0,08	(0,02)	0,54	(0,02)	0,06	(0,01)	0,53	(0,02)	0,05	(0,02)
ER-a	1xpc	1	1,5	0,70	(0,02)	0,12	(0,01)	0,55	(0,02)	0,05	(0,01)	0,45	(0,02)	0,04	(0,01)
	1xpc	1	1,9	0,65	(0,03)	0,31	(0,05)	0,62	(0,01)	0,10	(0,01)	0,54	(0,03)	0,08	(0,01)
	1xpc	1	4	0,76	(0,02)	0,35	(0,03)	0,61	(0,02)	0,05	(0,01)	0,61	(0,02)	0,13	(0,01)
ER-b	1qkm	1	1,5	0,58	(0,02)	0,07	(0,02)	0,49	(0,03)	0,04	(0,02)	0,59	(0,03)	0,07	(0,01)
	1qkm	1	1,9	0,59	(0,03)	0,04	(0,02)	0,53	(0,02)	0,04	(0,01)	0,58	(0,02)	0,08	(0,02)
	1qkm	1	4	0,63	(0,02)	0,08	(0,01)	0,50	(0,02)	0,06	(0,02)	0,62	(0,01)	0,08	(0,03)
FAK	1mp8	1	1,5	0,67	(0,04)	0,09	(0,02)	0,58	(0,03)	0,08	(0,02)	0,63	(0,03)	0,11	(0,02)
	1mp8	1	1,9	0,63	(0,01)	0,08	(0,02)	0,56	(0,03)	0,05	(0,02)	0,56	(0,04)	0,08	(0,02)
	1mp8	1	4	0,62	(0,02)	0,08	(0,02)	0,68	(0,03)	0,08	(0,02)	0,57	(0,04)	0,06	(0,01)
FXIa	1zsj	1	1,5	0,60	(0,03)	0,06	(0,01)	0,33	(0,03)	0,02	(0,01)	0,60	(0,03)	0,04	(0,02)
	1zsj	1	1,9	0,48	(0,02)	0,01	(0,01)	0,26	(0,02)	0,01	(0,01)	0,41	(0,01)	0,00	(0,00)
	1zsj	1	4	0,36	(0,03)	0,00	(0,00)	0,26	(0,02)	0,00	(0,00)	0,32	(0,02)	0,00	(0,00)
HIV-RT	2zd1	1	1,5	0,54	(0,02)	0,08	(0,01)	0,46	(0,01)	0,02	(0,01)	0,36	(0,02)	0,04	(0,02)
	2zd1	1	1,9	0,44	(0,02)	0,08	(0,02)	0,63	(0,01)	0,09	(0,02)	0,32	(0,03)	0,05	(0,01)
	2zd1	1	4	0,63	(0,02)	0,10	(0,01)	0,61	(0,02)	0,13	(0,02)	0,41	(0,02)	0,06	(0,02)
Hsp90	1uyl	1	1,5	0,69	(0,03)	0,07	(0,01)	0,56	(0,02)	0,01	(0,00)	0,62	(0,03)	0,13	(0,02)
	1uyl	1	1,9	0,60	(0,04)	0,14	(0,02)	0,63	(0,03)	0,11	(0,01)	0,59	(0,03)	0,11	(0,02)
	1uyl	1	4	0,66	(0,03)	0,13	(0,02)	0,69	(0,02)	0,23	(0,02)	0,65	(0,03)	0,19	(0,03)
PKA	2uzt	1	1,5	0,68	(0,02)	0,08	(0,02)	0,40	(0,02)	0,08	(0,02)	0,65	(0,03)	0,05	(0,01)
	2uzt	1	1,9	0,58	(0,03)	0,07	(0,02)	0,39	(0,03)	0,05	(0,01)	0,58	(0,02)	0,05	(0,01)
	2uzt	1	4	0,54	(0,02)	0,04	(0,01)	0,56	(0,05)	0,12	(0,03)	0,50	(0,02)	0,02	(0,01)
Rho-kinase2	2f2u	1	1,5	0,53	(0,02)	0,06	(0,02)	0,39	(0,03)	0,04	(0,01)	0,57	(0,02)	0,03	(0,00)
	2f2u	1	1,9	0,42	(0,03)	0,02	(0,01)	0,44	(0,01)	0,04	(0,01)	0,50	(0,01)	0,02	(0,01)
	2f2u	1	4	0,57	(0,03)	0,03	(0,01)	0,63	(0,02)	0,04	(0,01)	0,61	(0,03)	0,04	(0,01)

1) Die Nummerierung entspricht der Reihenfolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

Tabelle A9. Ergebnisse des retrospektiven virtuellen Screenings der MUV Datenbank mit dem Carbóindex als Ähnlichkeitsmaß. Der höchste ROCAUC Wert für jedes Enzym ist fett hervorgehoben.

Enzym	PDB Nr.	Taschen(n) ¹⁾	Clusterradius [Å] ²⁾	keine Skalierung		Blockskalierung				Skalierung auf Eins					
				ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)	ROCAUC	(σ)	BEDROC	(σ)
CathepsinG	1cgh	1	1,5	0,61	(0,02)	0,03	(0,00)	0,58	(0,03)	0,06	(0,02)	0,60	(0,04)	0,03	(0,01)
	1cgh	1	1,9	0,52	(0,03)	0,02	(0,01)	0,48	(0,02)	0,02	(0,01)	0,51	(0,02)	0,02	(0,01)
	1cgh	1	4	0,48	(0,03)	0,02	(0,01)	0,48	(0,02)	0,03	(0,01)	0,50	(0,02)	0,02	(0,00)
Eph	3ckh	2	1,5	0,48	(0,03)	0,03	(0,02)	0,49	(0,04)	0,03	(0,02)	0,49	(0,03)	0,04	(0,02)
	3ckh	2	1,9	0,49	(0,04)	0,03	(0,02)	0,49	(0,05)	0,04	(0,01)	0,50	(0,03)	0,03	(0,02)
	3ckh	2	4	0,50	(0,04)	0,03	(0,02)	0,51	(0,04)	0,07	(0,03)	0,48	(0,04)	0,04	(0,02)
ER-a	1xpc	1	1,5	0,57	(0,02)	0,08	(0,01)	0,51	(0,02)	0,03	(0,00)	0,56	(0,01)	0,08	(0,00)
	1xpc	1	1,9	0,55	(0,03)	0,07	(0,01)	0,59	(0,03)	0,10	(0,02)	0,54	(0,02)	0,07	(0,01)
	1xpc	1	4	0,59	(0,02)	0,09	(0,01)	0,60	(0,03)	0,08	(0,01)	0,60	(0,03)	0,09	(0,01)
ER-b	1qkm	1	1,5	0,50	(0,03)	0,11	(0,01)	0,51	(0,03)	0,06	(0,01)	0,50	(0,03)	0,11	(0,02)
	1qkm	1	1,9	0,51	(0,05)	0,07	(0,02)	0,52	(0,03)	0,07	(0,02)	0,51	(0,03)	0,07	(0,01)
	1qkm	1	4	0,50	(0,03)	0,06	(0,01)	0,48	(0,02)	0,05	(0,01)	0,49	(0,03)	0,06	(0,02)
FAK	1mp8	1	1,5	0,60	(0,02)	0,08	(0,02)	0,60	(0,04)	0,06	(0,01)	0,59	(0,02)	0,07	(0,02)
	1mp8	1	1,9	0,60	(0,02)	0,09	(0,02)	0,60	(0,03)	0,08	(0,01)	0,61	(0,03)	0,10	(0,01)
	1mp8	1	4	0,61	(0,04)	0,09	(0,02)	0,64	(0,05)	0,06	(0,02)	0,61	(0,03)	0,10	(0,02)
FXIa	1zsj	1	1,5	0,38	(0,02)	0,00	(0,00)	0,38	(0,02)	0,00	(0,00)	0,37	(0,02)	0,00	(0,00)
	1zsj	1	1,9	0,37	(0,02)	0,00	(0,00)	0,37	(0,02)	0,00	(0,00)	0,37	(0,01)	0,00	(0,00)
	1zsj	1	4	0,36	(0,02)	0,00	(0,00)	0,39	(0,02)	0,00	(0,00)	0,37	(0,03)	0,00	(0,00)
HIV-RT	2zd1	1	1,5	0,55	(0,04)	0,07	(0,02)	0,47	(0,02)	0,03	(0,01)	0,56	(0,02)	0,07	(0,01)
	2zd1	1	1,9	0,63	(0,02)	0,10	(0,02)	0,57	(0,03)	0,10	(0,01)	0,64	(0,02)	0,09	(0,02)
	2zd1	1	4	0,60	(0,03)	0,11	(0,01)	0,51	(0,03)	0,11	(0,02)	0,60	(0,02)	0,10	(0,02)
Hsp90	1uyl	1	1,5	0,66	(0,03)	0,14	(0,03)	0,56	(0,02)	0,01	(0,00)	0,63	(0,02)	0,12	(0,01)
	1uyl	1	1,9	0,65	(0,04)	0,12	(0,02)	0,62	(0,03)	0,08	(0,02)	0,66	(0,03)	0,13	(0,02)
	1uyl	1	4	0,67	(0,04)	0,17	(0,03)	0,62	(0,05)	0,19	(0,03)	0,67	(0,03)	0,15	(0,03)
PKA	2uzt	1	1,5	0,49	(0,04)	0,05	(0,03)	0,42	(0,04)	0,04	(0,02)	0,51	(0,02)	0,05	(0,01)
	2uzt	1	1,9	0,48	(0,02)	0,06	(0,02)	0,39	(0,03)	0,02	(0,01)	0,49	(0,02)	0,06	(0,02)
	2uzt	1	4	0,56	(0,04)	0,05	(0,03)	0,51	(0,02)	0,07	(0,01)	0,55	(0,02)	0,06	(0,03)
Rho-kinase2	2f2u	1	1,5	0,49	(0,02)	0,05	(0,02)	0,52	(0,03)	0,07	(0,02)	0,49	(0,04)	0,06	(0,02)
	2f2u	1	1,9	0,53	(0,04)	0,03	(0,01)	0,48	(0,03)	0,02	(0,01)	0,52	(0,04)	0,03	(0,02)
	2f2u	1	4	0,55	(0,03)	0,06	(0,01)	0,48	(0,04)	0,01	(0,01)	0,55	(0,03)	0,05	(0,02)

1) Die Nummerierung entspricht der Reihenfolge der Ausgabe von PocketPicker und damit dem Größenrang der jeweiligen Tasche.

2) LIQUID Clusterradius für lipophile Interaktionen.

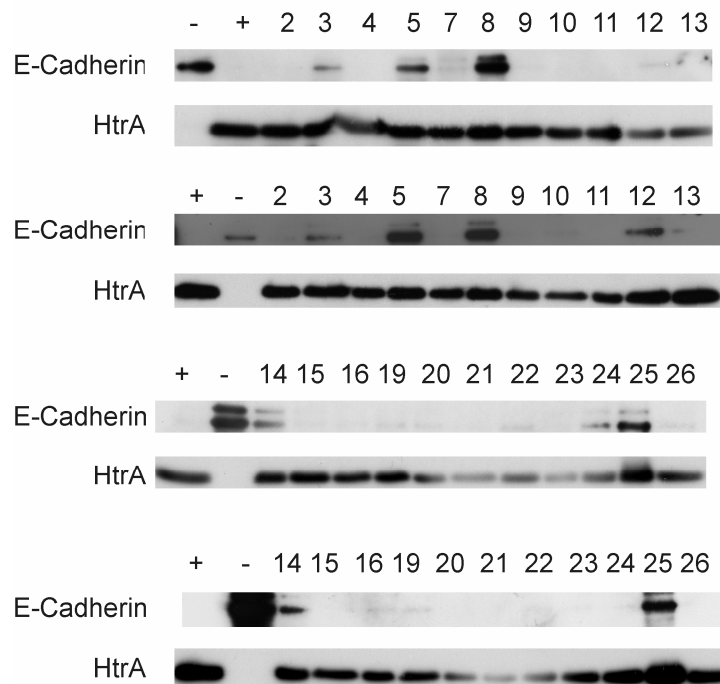
Anhang C: Markierte Western-Blots der *in-vitro* Kontrollexperimente

Abbildung A1. Mit E-Cadherin bzw. HtrA markierte Western-Blots des Screenings nach Inhibitoren für die Protease HtrA. Die Spur - ist eine Negativkontrolle ohne HtrA, die Spur + die Positivkontrolle ohne Inhibitor. Die anderen Spuren sind entsprechend der benutzten Substanz nummeriert (vgl. Tabelle 11). Dargestellt ist das Screening (oben) mit Kontrolle (unten). Alle Substanzen wurden mit einer Konzentration von 100 μ M eingesetzt.

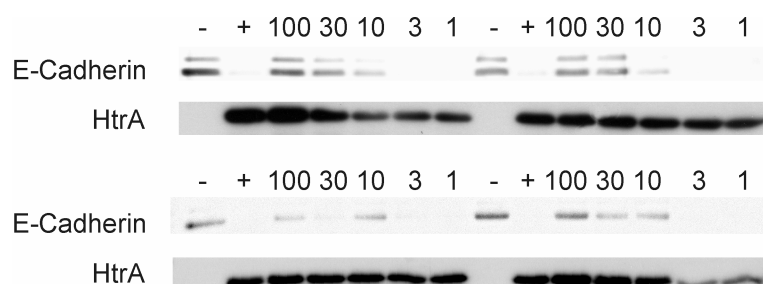


Abbildung A2. Mit E-Cadherin bzw. HtrA markierte Western-Blots für die Bestimmung eines IC_{50} der Substanz **8** (vgl. Tabelle 11). Die Substanz wurde in verschiedenen Konzentrationen eingesetzt. Der Versuch wurde vier Mal durchgeführt. Die Spur - ist eine Negativkontrolle ohne HtrA, die Spur + die Positivkontrolle ohne Inhibitor. Die Beschriftungen der anderen Spuren sind die Konzentrationen von Substanz **8** in μ M.

Anhang D: Parametereinstellungen von GOLD

```
autoscale = 1
popsiz = auto
select_pressure = auto
n_islands = auto
maxops = auto
niche_siz = auto
pt_crosswt = auto
allele_mutatewt = auto
migratewt = auto
radius = 10
origin = 0 0 0
do_cavity = 1
floodfill_atom_no = 0
param_file = DEFAULT
set_ligand_atom_types = 1
set_protein_atom_types = 0
tordist_file = DEFAULT
save_lone_pairs = 1
fit_points_file = fit_pts.mol2
read_fitpts = 0
internal_ligand_h_bonds = 0
n_ligand_bumps = 0
flip_free_corners = 0
flip_amide_bonds = 0
flip_planar_n = 1
flip_ring_NRR
flip_ring_NHR
flip_pyramidal_n = 0
rotate_carboxylic_oh = flip
use_tordist = 1
postprocess_bonds = 1
rotatable_bond_override_file = DEFAULT
early_termination = 1
n_top_solutions = 3
rms_tolerance = 1.5
force_constraints = 0
covalent = 0
initial_virtual_pt_match_max = 3
relative_ligand_energy = 0
start_vdw_linear_cutoff = 6
score_param_file = DEFAULT
```

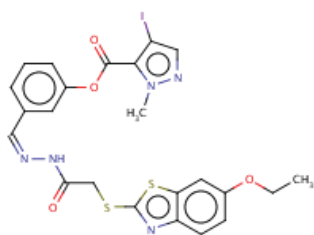
Die folgenden Aminosäuren wurden als Definition der Bindetasche benutzt:

GLU207, LEU99, PHE182, SER241, ASN208, SER144, LY244, VAL117, ILE201, HIS116, ASP147, TYR206, SER221, ILE253, MET257, LE239, SER146, PRO218, PHE209, GLU145, GLY219, SER205, THR237, ALA238, GLY202, ASN217.

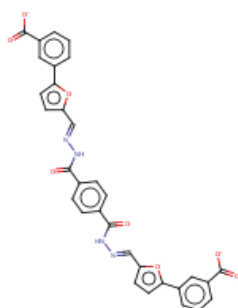
Anhang E: Ergebnislisten der retrospektiven virtuellen Screenings

Im Folgenden finden sich die jeweils 100 ersten Einträge der sortierten Datenbank nach dem virtuellen Screening mit den strukturabgeleiteten Pharmakophormodellen 1, 2 und 3 (vgl. Abschnitt 6.8.1). Gezeigt sind jeweils die Strukturformel, die Listenposition und der Datenbankbezeichner. Substanzen, deren Bezeichner mit „ASN“ oder „BAS“ beginnen, stammen von der Firma Asinex, alle anderen von der Firma Specs.

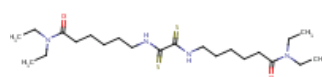
Modell 1



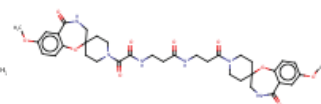
1: AK-968/40385474



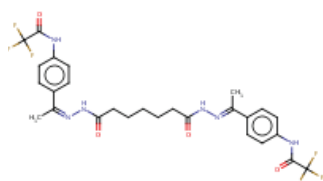
2: BAS 00314028



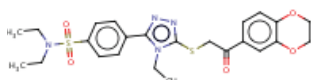
3: BAS 00005210



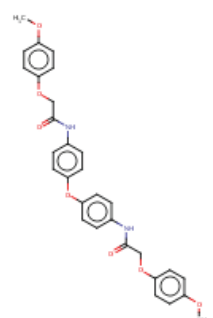
4: ASN 17455406



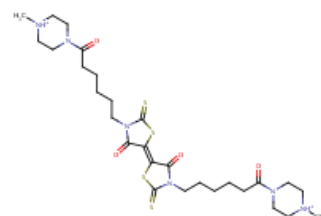
5: AK-968/40391425



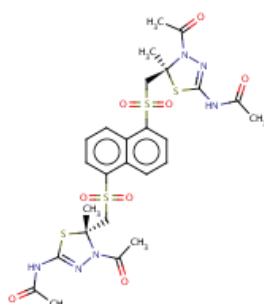
6: ASN 04363244



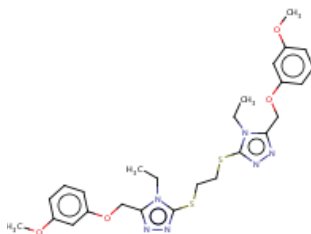
7: AN-329/41692968



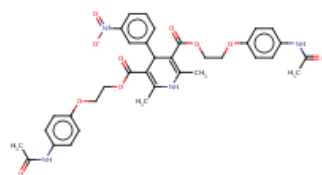
8: AG-690/34470005



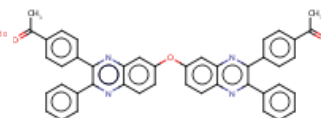
9: AG-690/34443019



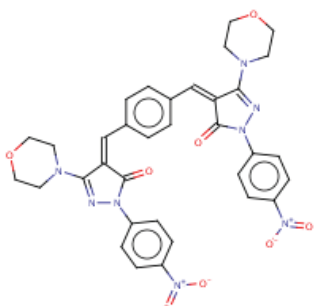
10: ASN 04456672



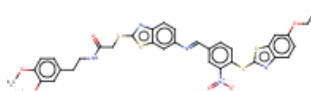
11: AQ-405/42300192



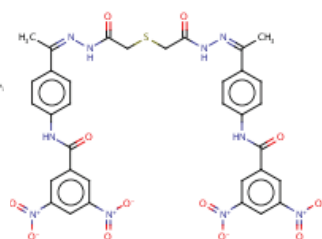
12: AG-690/36760005



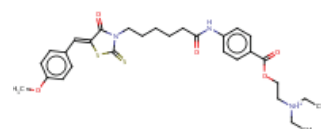
13: BAS 00412574



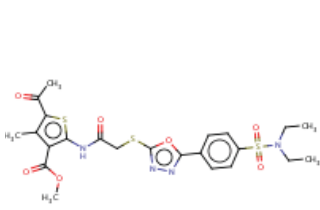
14: AG-205/11552188



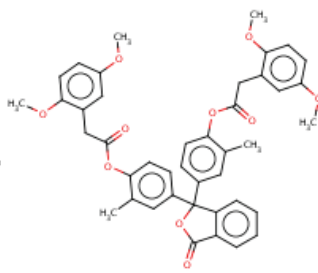
15: AG-205/32455029



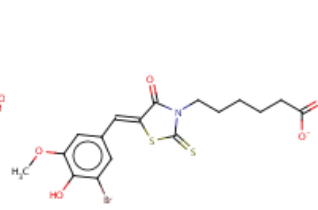
16: BAS 00582371



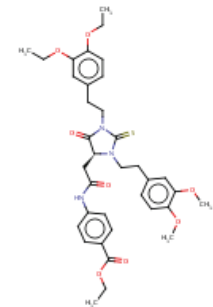
17: ASN 04363147



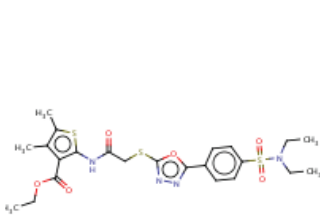
18: AN-652/13689016



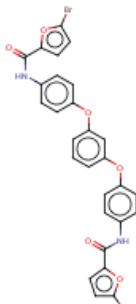
19: BAS 02167321



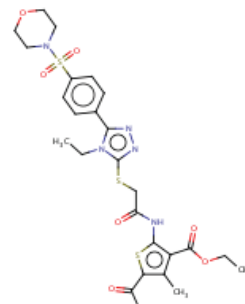
20: AK-968/41171673



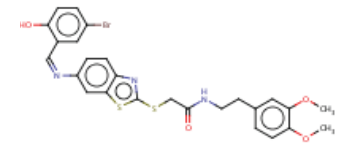
21: ASN 04363145



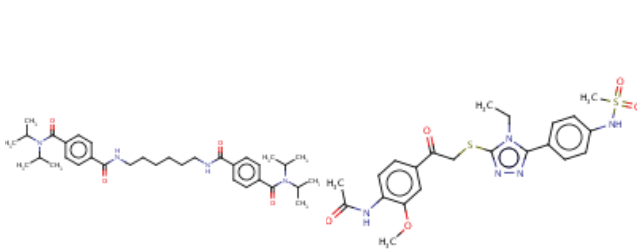
22: AG-690/10117038



23: ASN 02992586

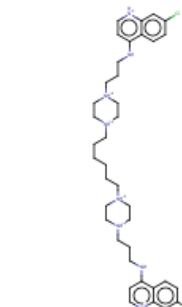


24: AG-205/11218060

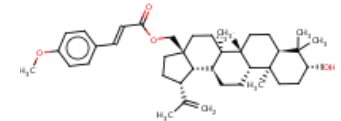


25: AK-968/15363102

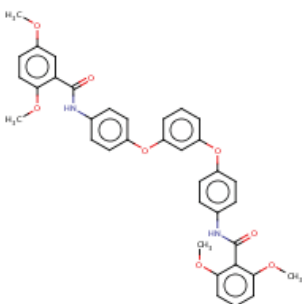
26: ASN 04374806



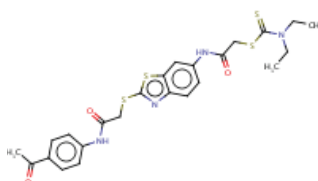
27: AN-740/37278044



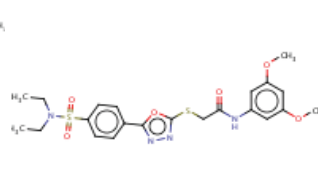
28: BAS 01279922



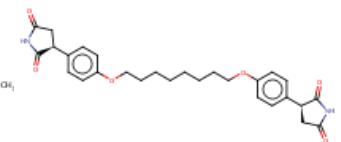
29: AK-918/41204444



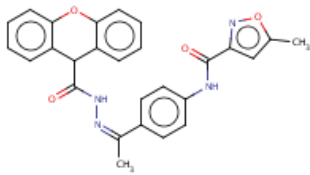
30: BAS 01853743



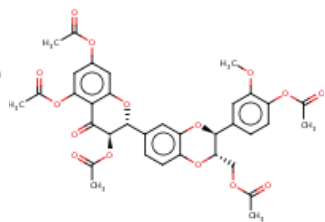
31: ASN 04363098



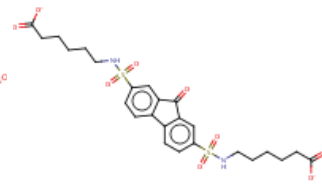
32: BAS 00404260



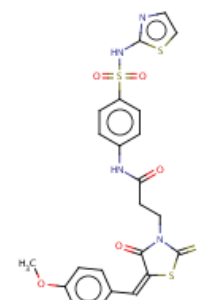
33: AK-968/41025690



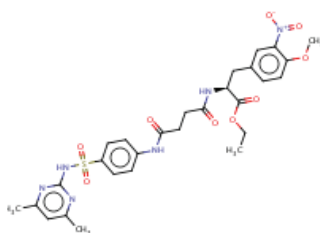
34: BAS 00704777



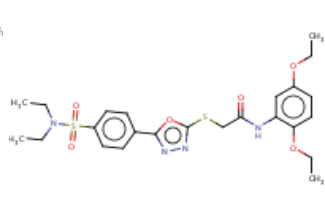
35: BAS 00668847



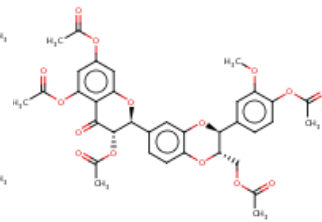
36: AF-399/32023047



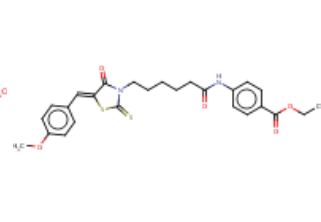
37: BAS 00549582



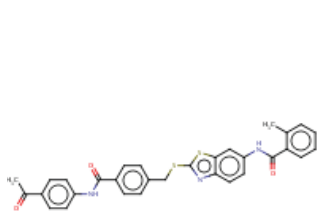
38: ASN 04363067



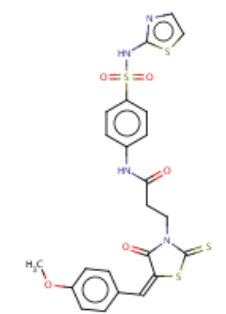
39: AE-848/32723040



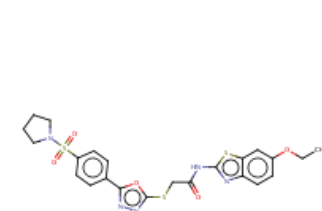
40: BAS 00502851



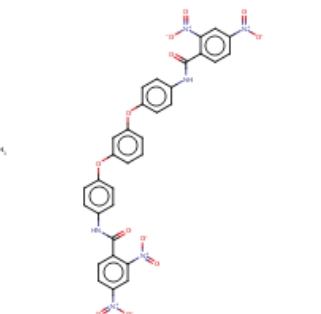
41: AQ-088/42014307



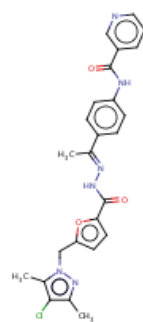
42: BAS 00555414



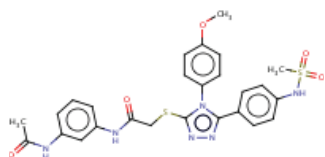
43: ASN 03798424



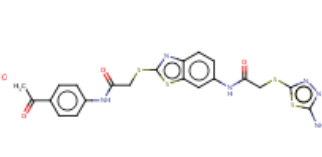
44: BAS 00243183



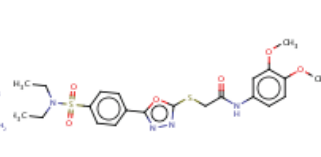
45: AK-968/40707810



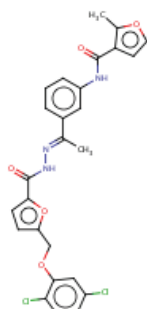
46: ASN 04445296



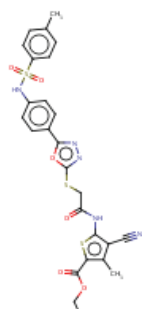
47: BAS 01853759



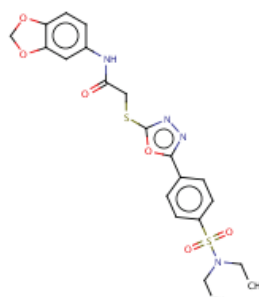
48: ASN 04363097



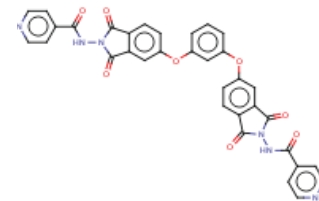
49: AK-968/40641861



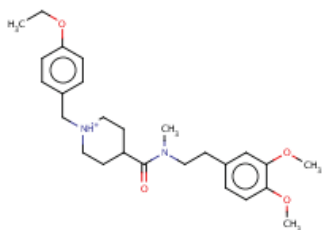
50: ASN 03793304



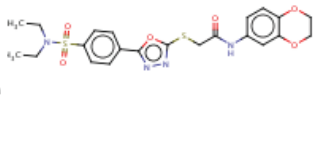
51: ASN 04363152



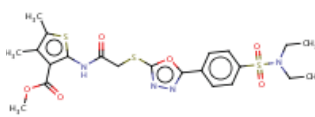
52: BAS 00287505



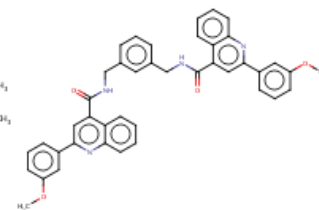
53: BAS 03838056



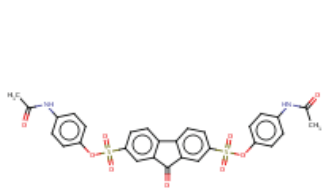
54: ASN 04363151



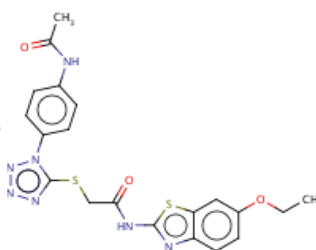
55: ASN 04363144



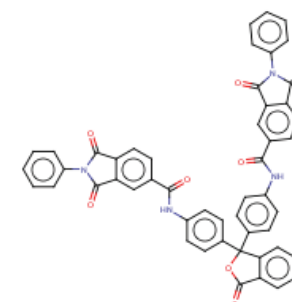
56: AK-968/12974774



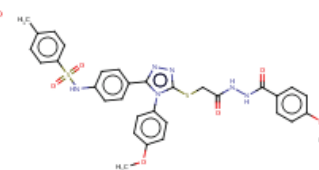
57: BAS 00621520



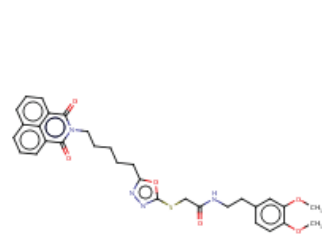
58: ASN 05116187



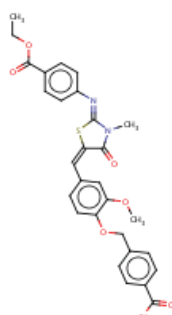
59: AI-020/33342010



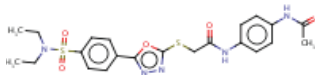
60: ASN 03977333



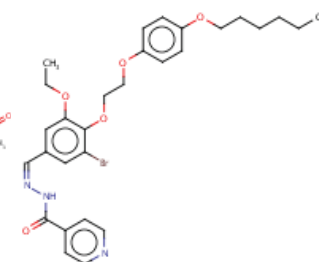
61: AF-399/41219715



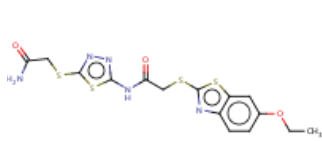
62: AM-879/40777304



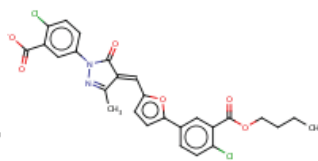
63: ASN 04363106



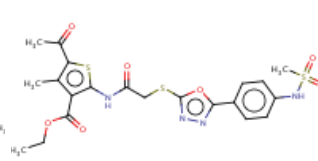
64: AM-900/15248041



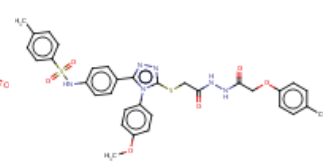
65: BAS 12711821



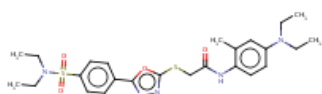
66: BAS 00761488



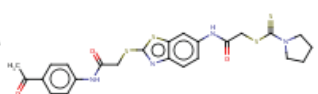
67: ASN 05343043



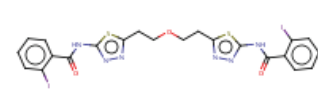
68: ASN 03977351



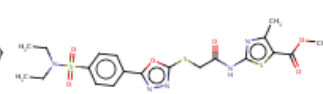
69: ASN 04363081



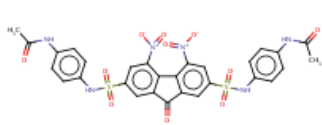
70: BAS 01853746



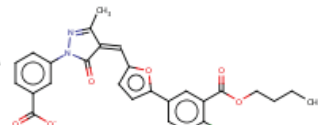
71: AG-690/36272042



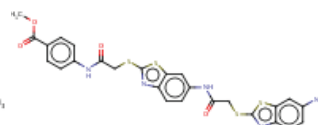
72: ASN 04363224



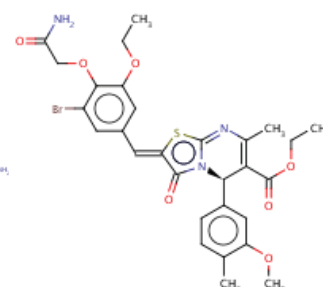
73: BAS 00523519



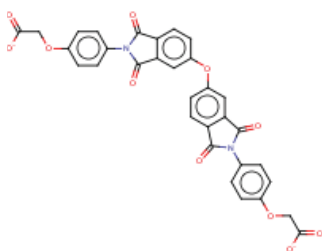
74: AG-690/37152140



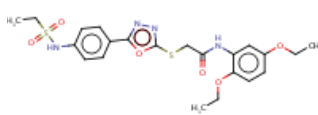
75: AG-205/13056178



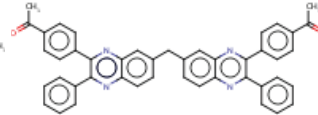
76: AO-081/40694606



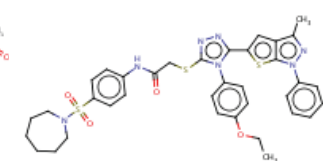
77: BAS 00505345



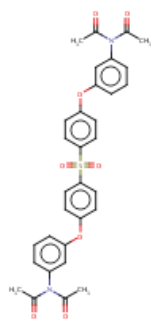
78: ASN 05342906



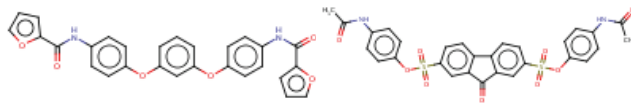
79: AG-690/10517059



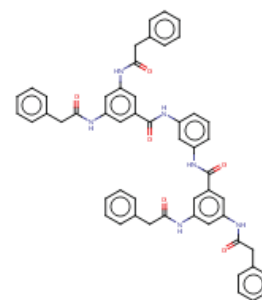
80: ASN 04036050



81: BAS 00280828

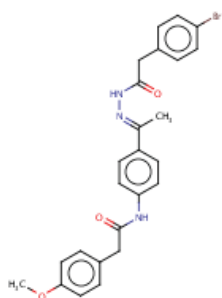


82: BAS 00119630

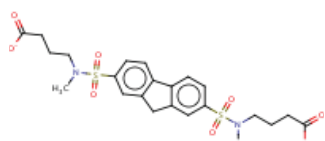


83: AG-690/37071045

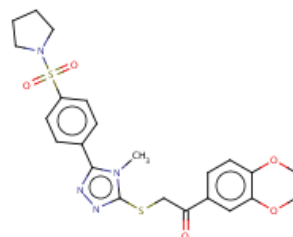
84: AE-848/08584029



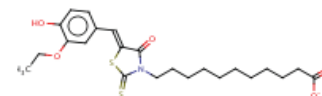
85: AK-968/40641738



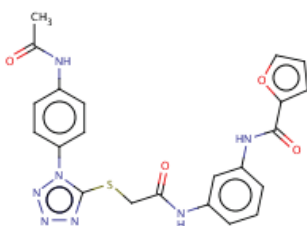
86: BAS 00668848



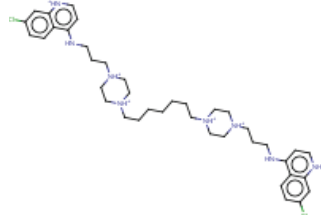
87: ASN 04448400



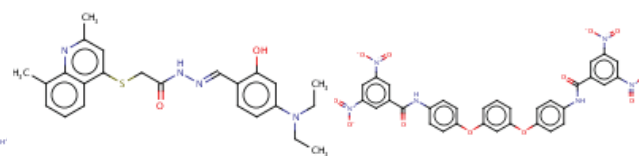
88: BAS 00534130



89: ASN 05836332

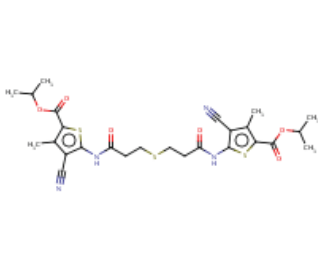


90: AO-295/15484020

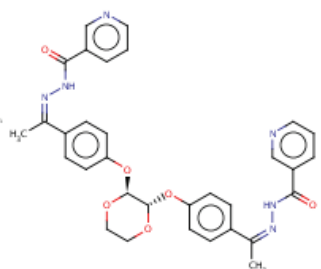


91: AN-698/40745340

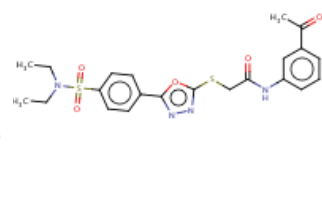
92: BAS 00465079



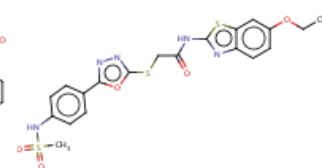
93: AK-968/12686027



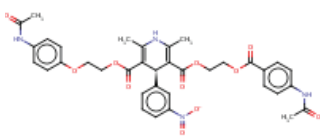
94: BAS 00523834



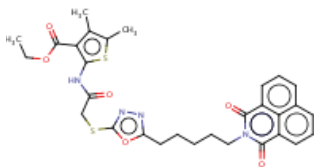
95: ASN 04363104



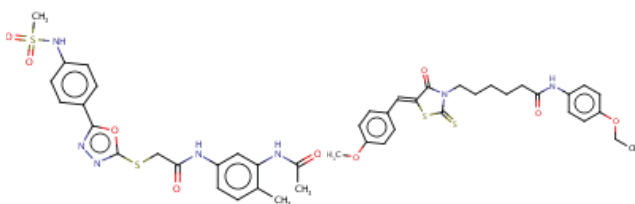
96: ASN 05343143



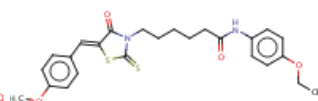
97: AQ-405/42300218



98: AF-399/41346082

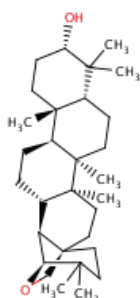


99: ASN 05343193

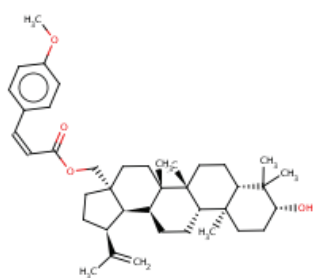


100: BAS 00555450

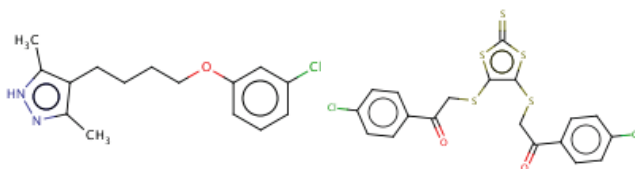
Modell 2



1: BAS 01279935

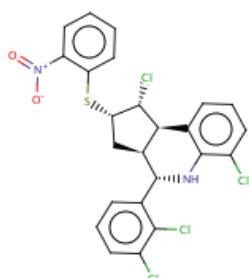


2: BAS 01279922

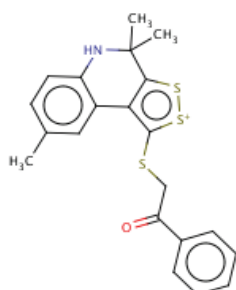


3: AG-664/14117574

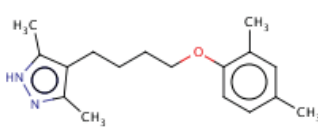
4: BAS 00733101



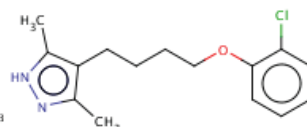
5: AG-690/12243088



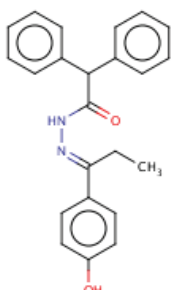
6: BAS 00126935



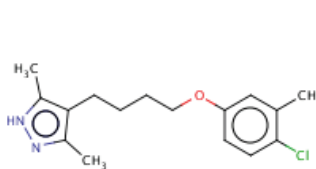
7: AG-664/14117583



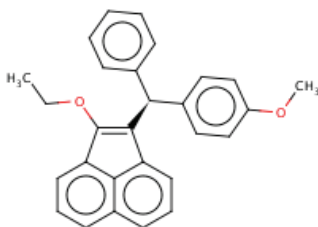
8: AG-664/14117571



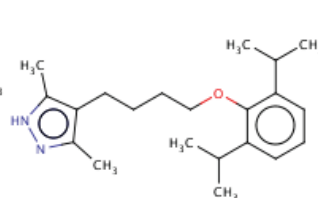
9: AN-329/40529483



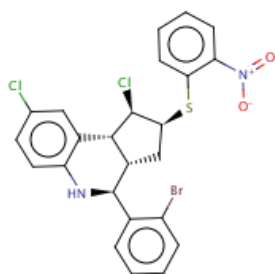
10: AG-664/14117324



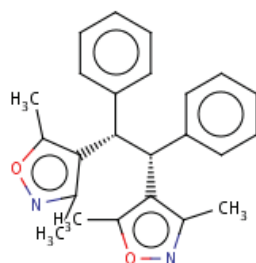
11: AQ-344/43100147



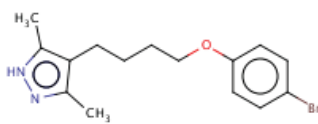
12: AJ-292/41721705



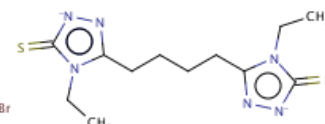
13: AG-690/13153022



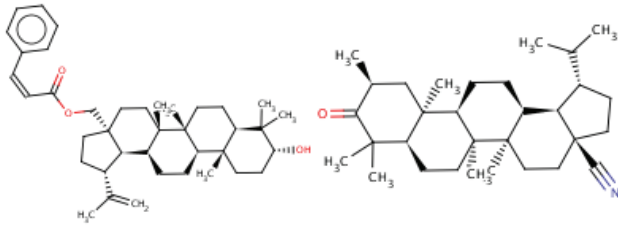
14: BAS 00122318



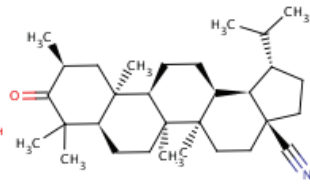
15: AG-664/14117580



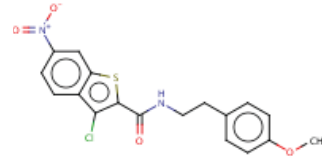
16: AF-399/41945649



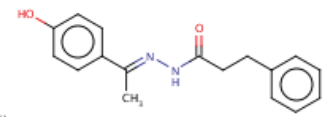
17: BAS 01279920



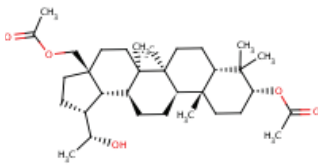
18: AE-641/00132007



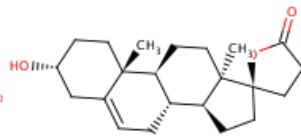
19: AK-968/41026031



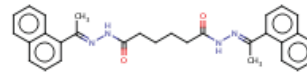
20: BAS 00295193



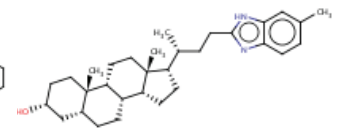
21: AE-641/00404028



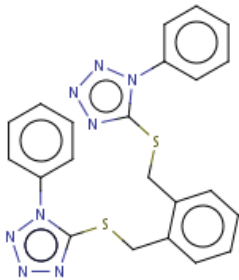
22: BAS 01417440



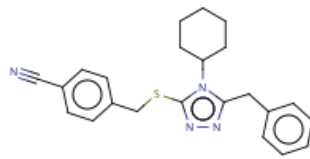
23: AK-968/40388888



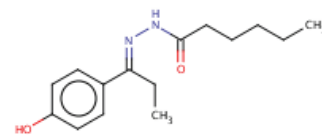
24: AE-641/15486018



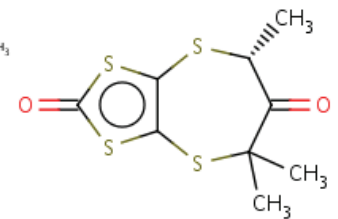
25: AI-204/42873860



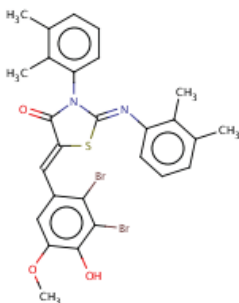
26: ASN 02832908



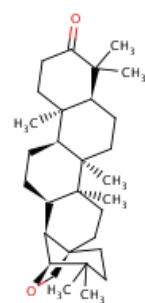
27: AK-968/40386626



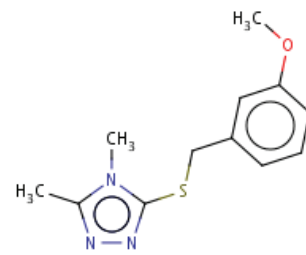
28: BAS 00733195



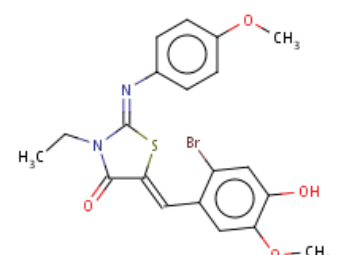
29: AN-655/13672005



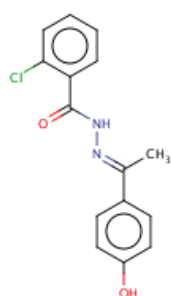
30: BAS 01279949



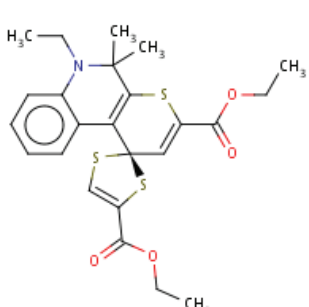
31: ASN 07331285



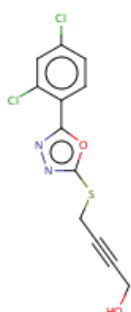
32: AO-081/15571755



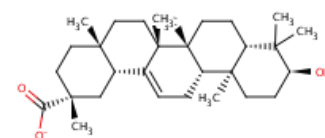
33: AN-329/10406012



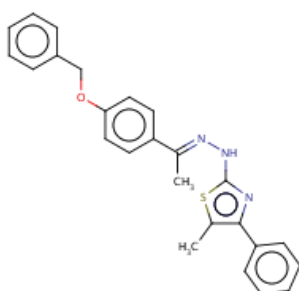
34: AG-690/12249217



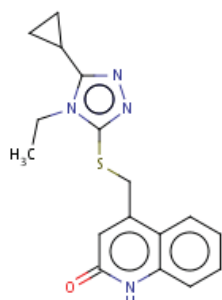
35: AP-853/43261177



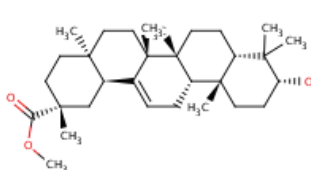
36: BAS 01279930



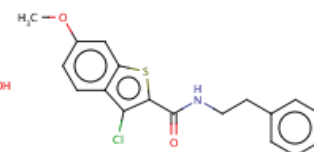
37: AH-487/40937099



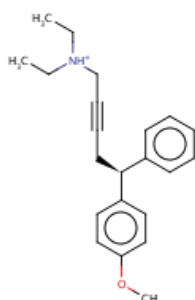
38: ASN 05988741



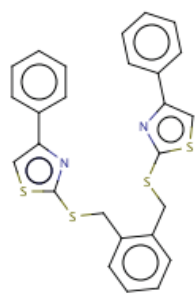
39: BAS 01279931



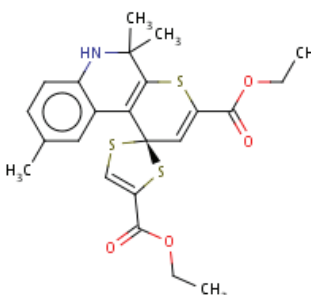
40: BAS 06656379



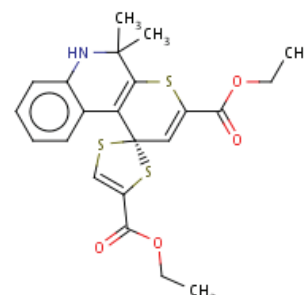
41: BAS 03200291



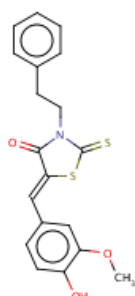
42: AG-401/43287187



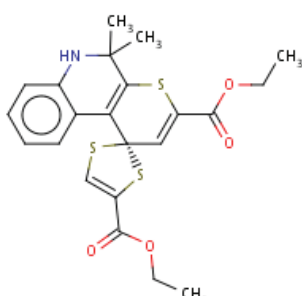
43: AG-690/12247019



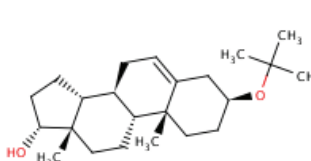
44: AG-690/12249214



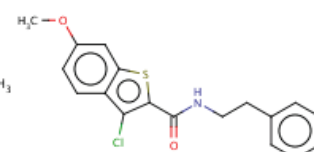
45: BAS 01151335



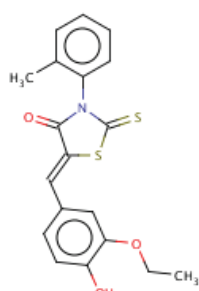
46: BAS 00861753



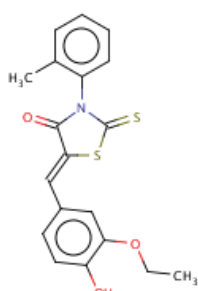
47: AE-562/12222512



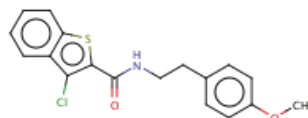
48: AN-329/41830764



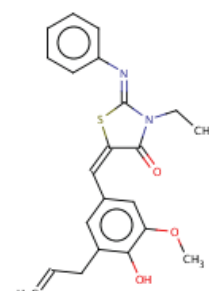
49: BAS 00682842



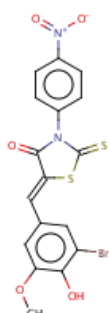
50: AG-690/11665374



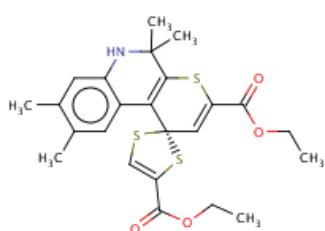
51: BAS 04085580



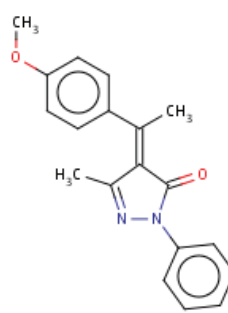
52: AO-081/15569120



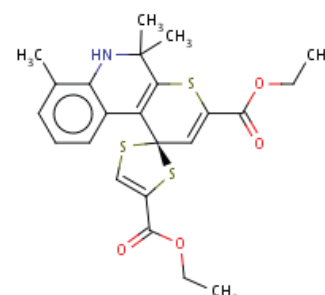
53: BAS 00744445



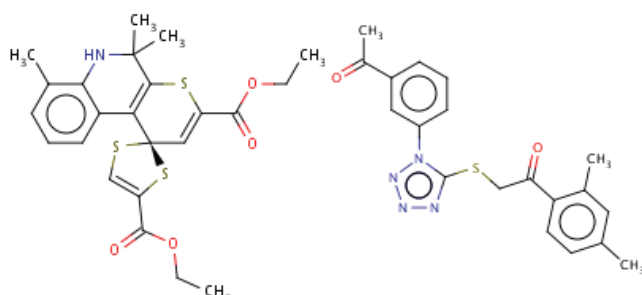
54: AG-690/12247020



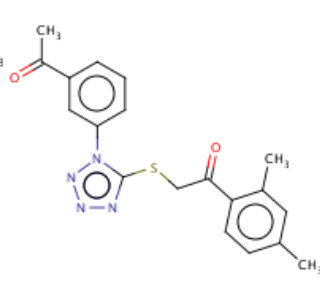
55: AE-848/12124440



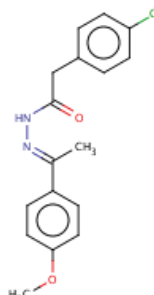
56: BAS 01173529



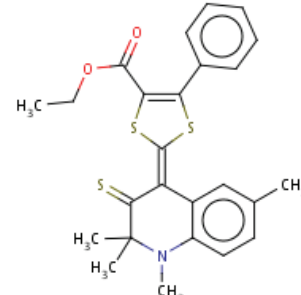
57: AJ-292/13095127



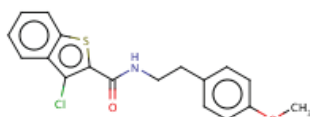
58: ASN 05116431



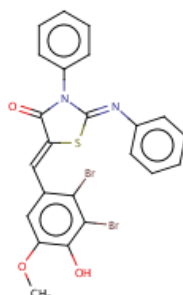
59: BAS 05486425



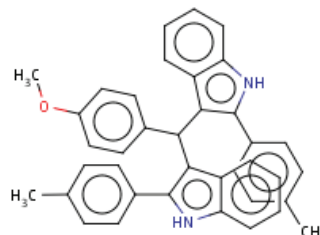
60: BAS 01045201



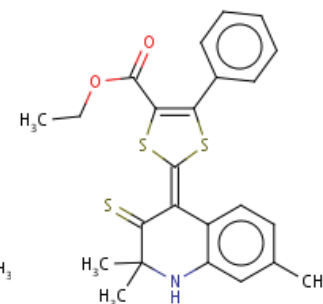
61: AG-690/36005033



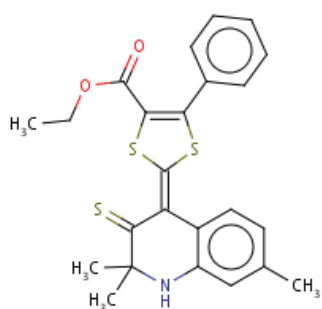
62: AK-968/12342072



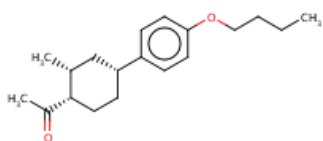
63: AG-664/14116120



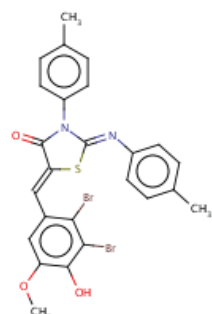
64: BAS 00458916



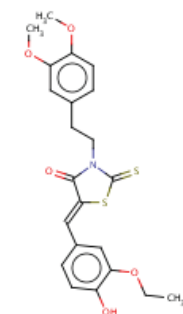
65: AG-690/11096063



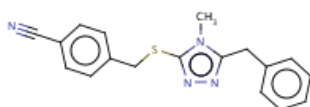
66: BAS 00396059



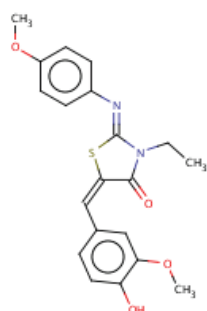
67: AN-655/13401012



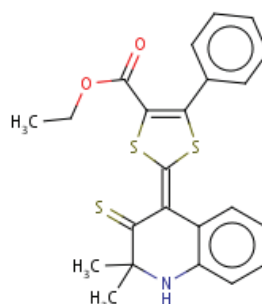
68: BAS 01151846



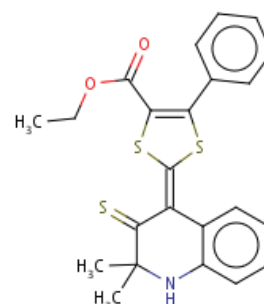
69: AH-487/42368819



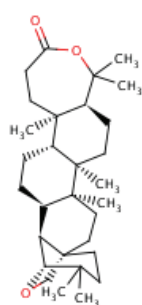
70: AO-081/15570058



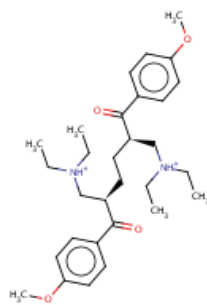
71: AG-690/36897030



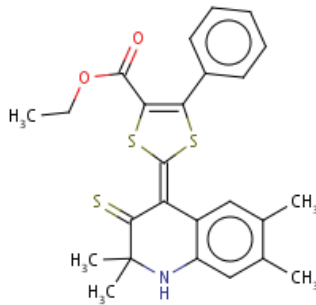
72: BAS 00458954



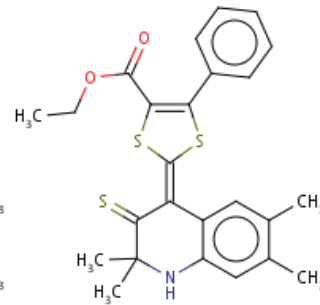
73: AE-641/00405027



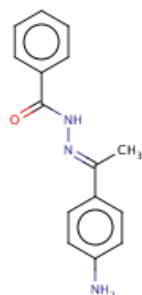
74: AG-690/09687011



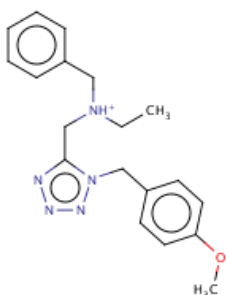
75: BAS 00584675



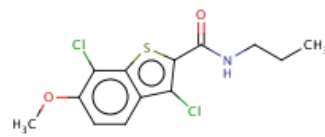
76: AG-690/37025122



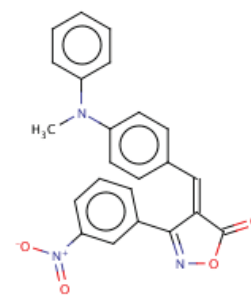
77: BAS 00125489



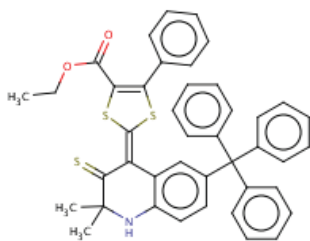
78: ASN 05071993



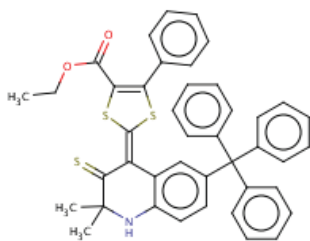
79: BAS 04296255



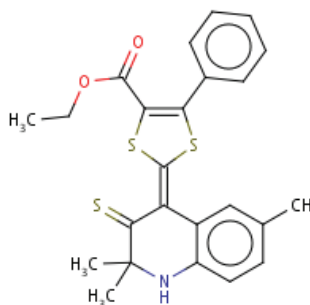
80: AG-205/33684012



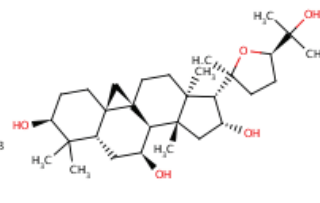
81: AK-968/37053118



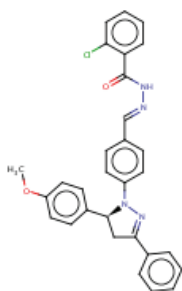
82: BAS 00619883



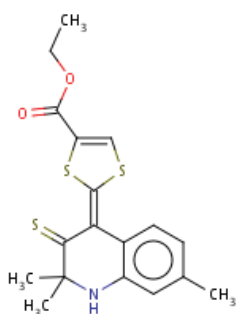
83: BAS 00458913



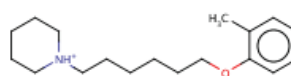
84: BAS 01832189



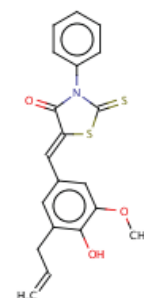
85: BAS 00603590



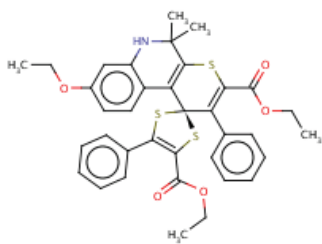
86: BAS 02232554



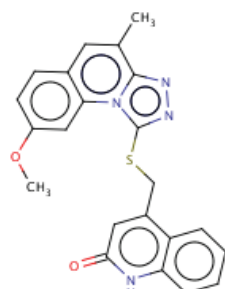
87: AJ-292/41721746



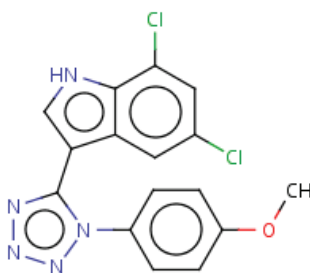
88: BAS 01108427



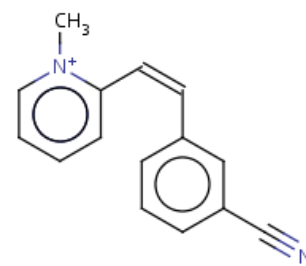
89: AK-968/13030073



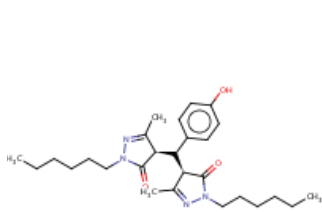
90: ASN 04371504



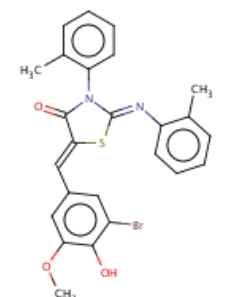
91: AG-401/43287212



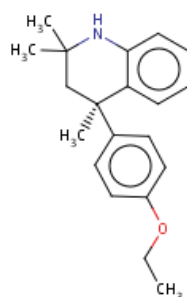
92: ASN 10791121



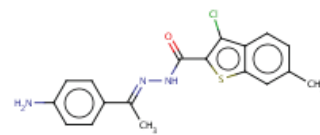
93: BAS 01249429



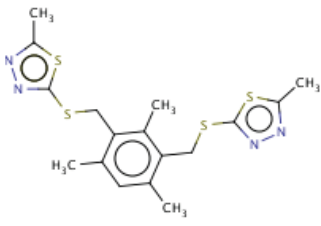
94: AK-968/11840102



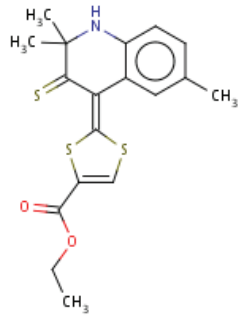
95: AK-968/12096335



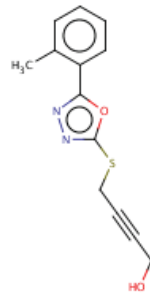
96: AK-968/15254114



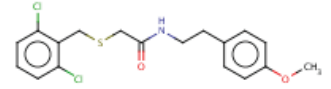
97: AG-205/13547011



98: BAS 01094998

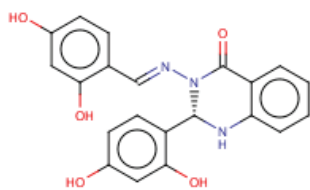


99: AP-853/43368071

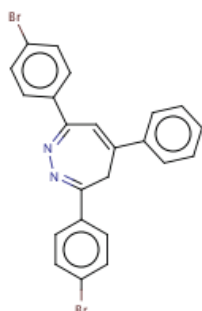


100: AO-081/42097440

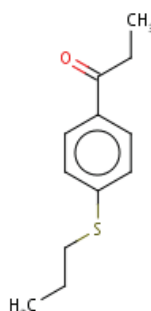
Modell 3



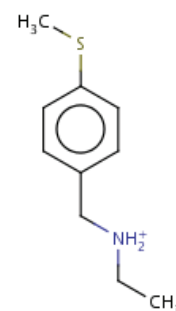
1: BAS 01074227



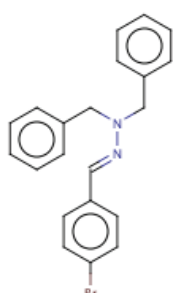
2: AG-401/37131020



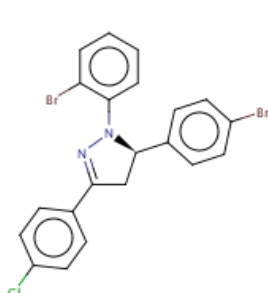
3: AN-651/43278366



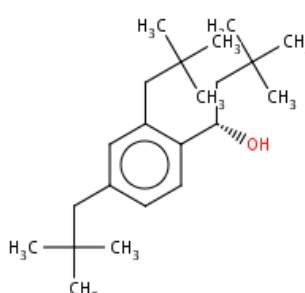
4: AN-465/42886904



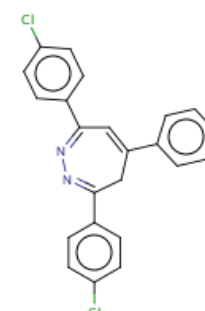
5: BAS 00113484



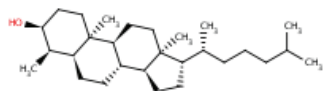
6: AG-205/07907014



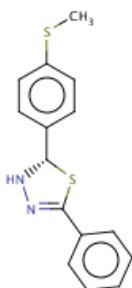
7: AE-562/12222049



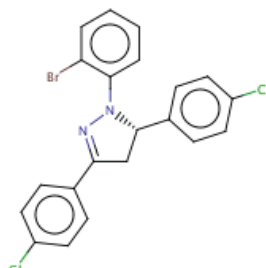
8: AG-401/13978013



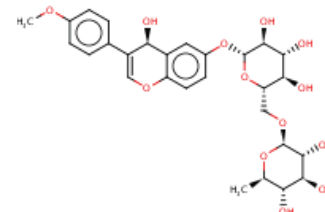
9: AG-803/12478032



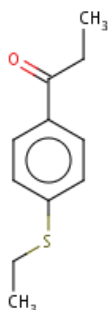
10: AG-670/41011532



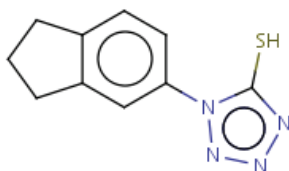
11: AG-205/07762006



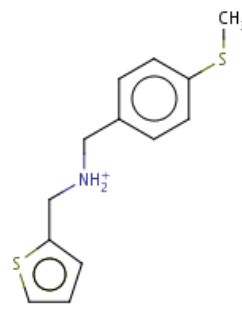
12: BAS 00225279



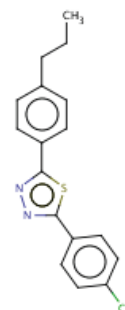
13: AN-651/43278365



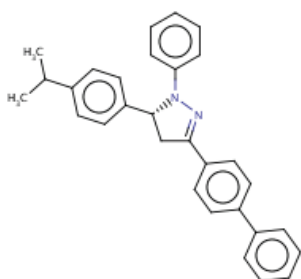
14: ASN 17326423



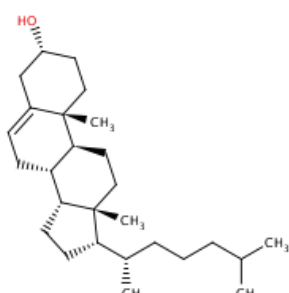
15: AN-465/42886511



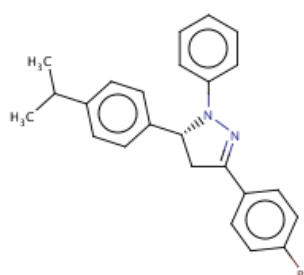
16: BAS 00260605



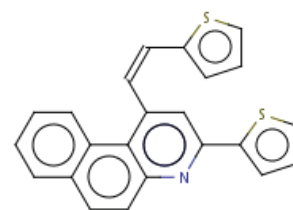
17: AG-690/36789008



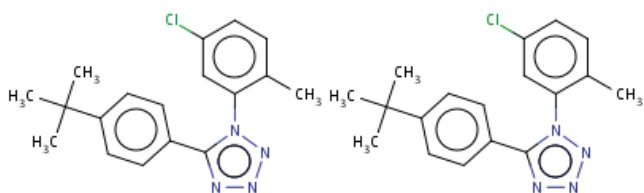
18: BAS 03662516



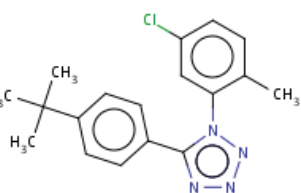
19: AG-690/10630045



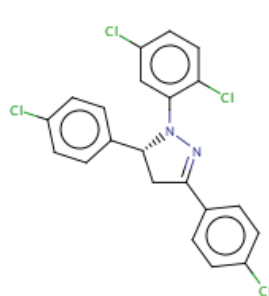
20: AG-205/32408006



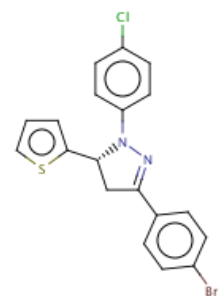
21: AG-690/40750186



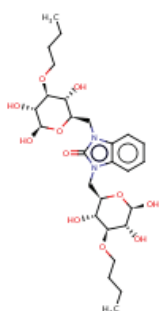
22: BAS 01842738



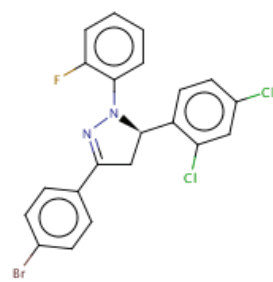
23: AG-205/07762008



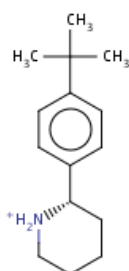
24: AG-690/33369030



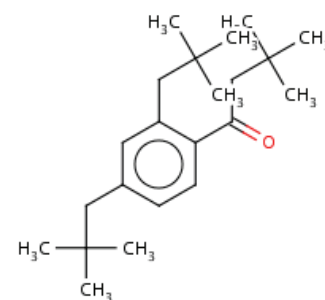
25: AO-763/14815008



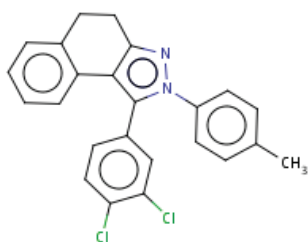
26: AG-205/33132053



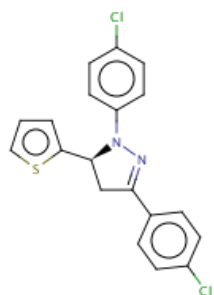
27: AP-836/41220127



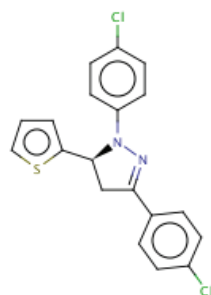
28: AE-562/12222050



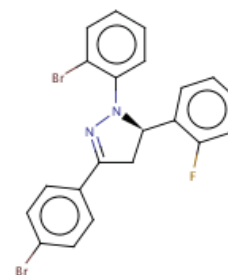
29: AG-401/43237640



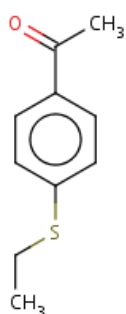
30: BAS 00148195



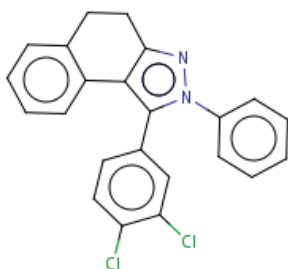
31: AG-690/33482011



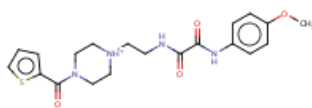
32: AG-205/07765020



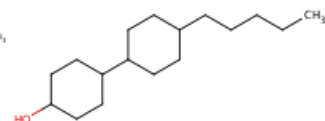
33: AN-651/43278362



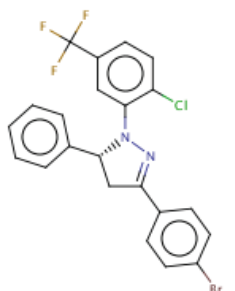
34: AG-401/43237642



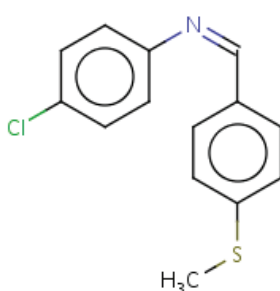
35: BAS 04327228



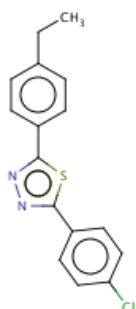
36: BAS 01123619



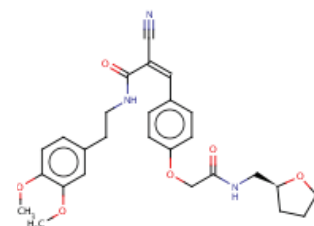
37: AG-690/34682010



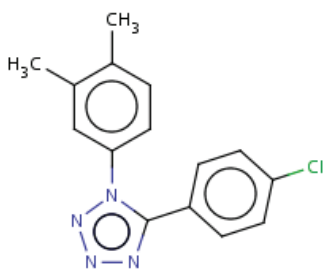
38: BAS 00485000



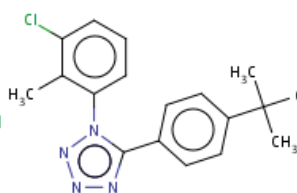
39: BAS 00333486



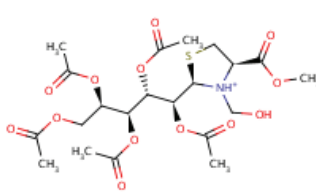
40: BAS 02301436



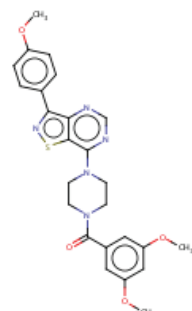
41: AL-291/41014938



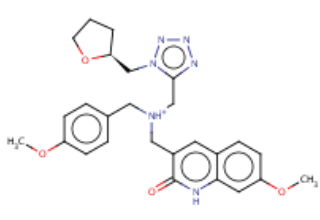
42: BAS 01842743



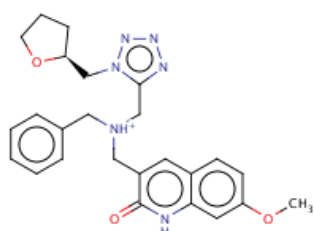
43: AJ-264/36455013



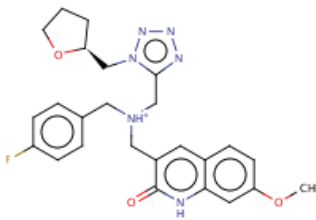
44: ASN 06353605



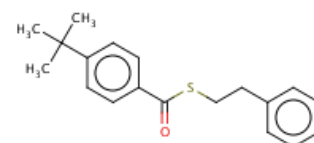
45: ASN 04370119



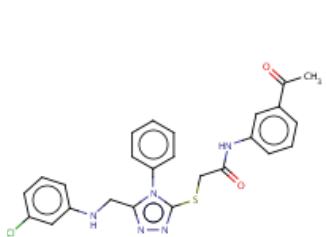
46: ASN 04370116



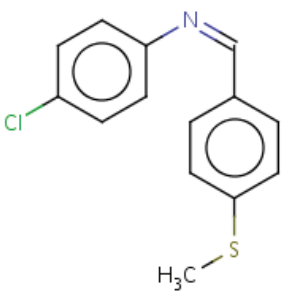
47: ASN 04370118



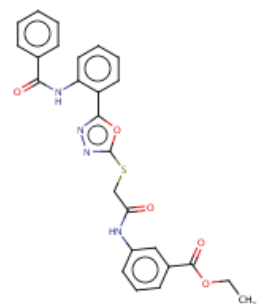
48: AQ-917/42754699



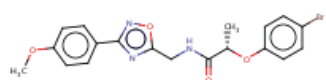
49: ASN 02070012



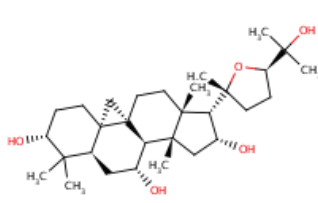
50: AG-690/12765745



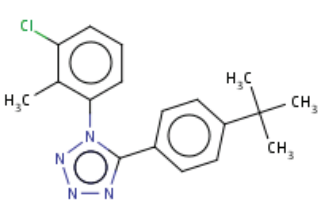
51: ASN 02069959



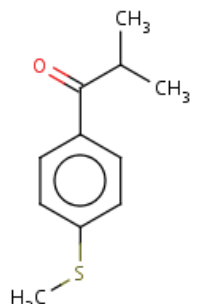
52: BAS 13152009



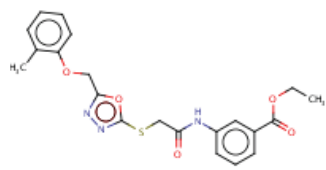
53: BAS 01832189



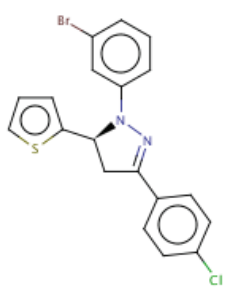
54: AG-690/40750188



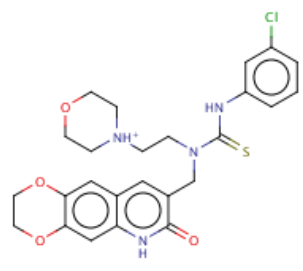
55: AQ-917/42753998



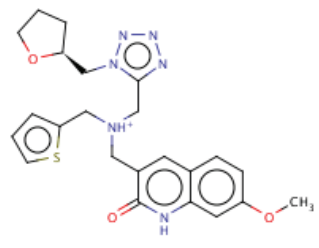
56: BAS 02325366



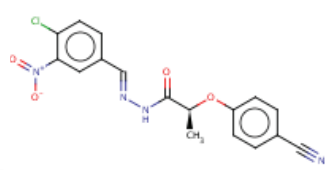
57: AG-690/33369027



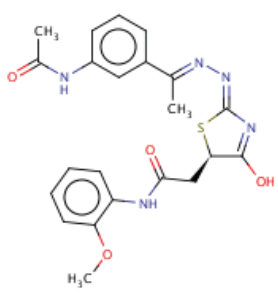
58: ASN 03776561



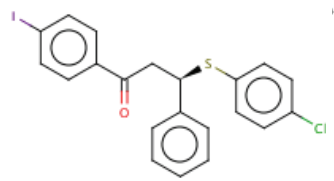
59: ASN 04370123



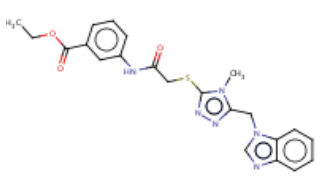
60: AK-968/40708438



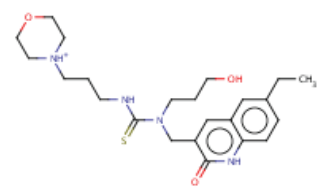
61: BAS 02175958



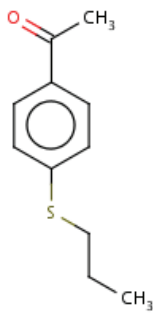
62: AG-401/11255001



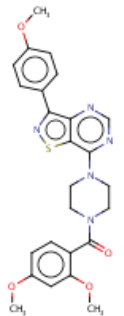
63: ASN 02993203



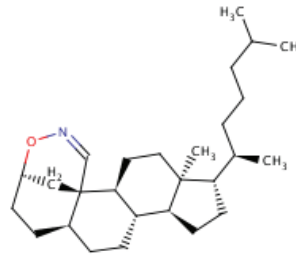
64: ASN 03206733



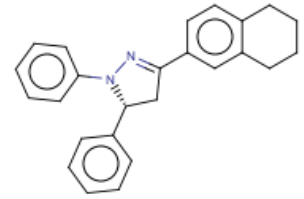
65: AN-651/43278363



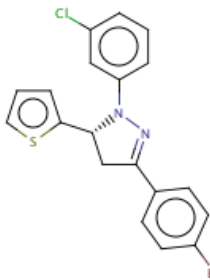
66: ASN 06353628



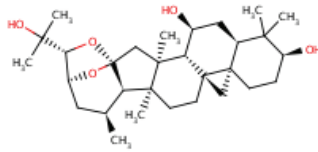
67: AN-835/13721003



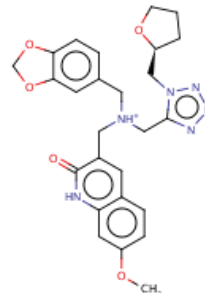
68: AL-182/11876024



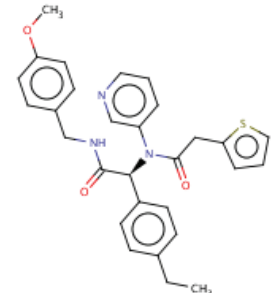
69: AG-690/33369044



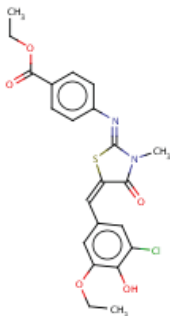
70: BAS 01832195



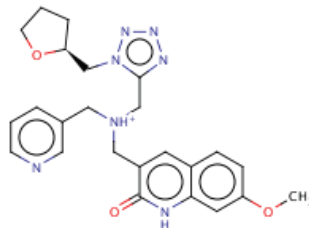
71: ASN 04370120



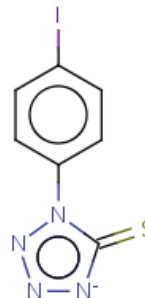
72: ASN 05580115



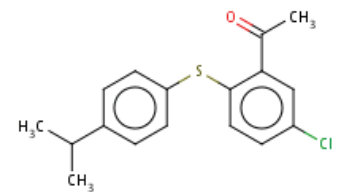
73: AN-989/40683839



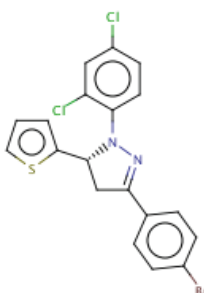
74: ASN 04370117



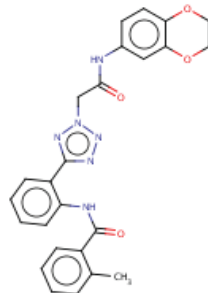
75: AE-848/32010051



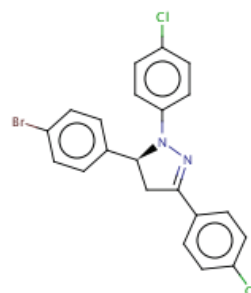
76: AE-641/00133005



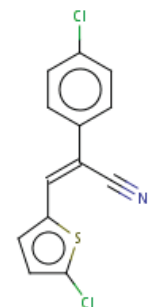
77: AG-690/33369012



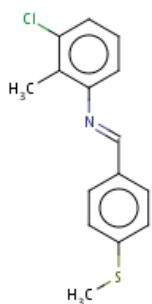
78: ASN 05017962



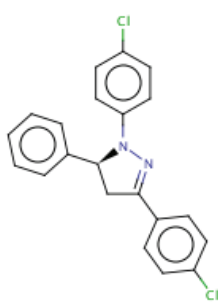
79: BAS 00125977



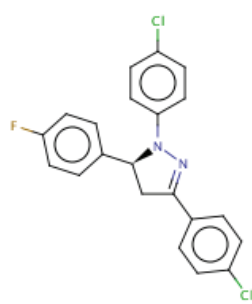
80: BAS 06613690



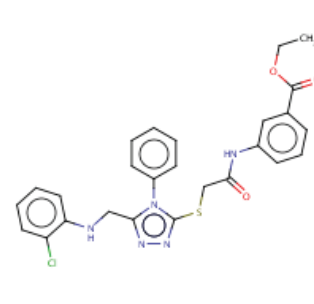
81: AN-329/11481574



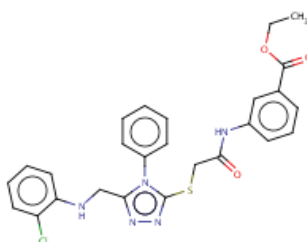
82: BAS 02866923



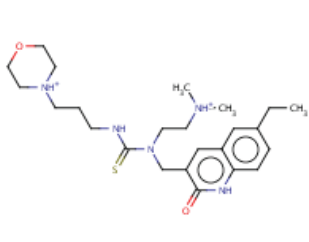
83: BAS 02800829



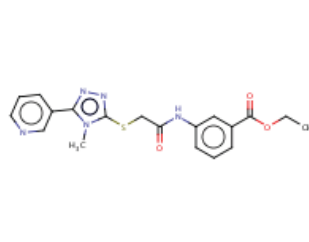
84: AG-690/40750240



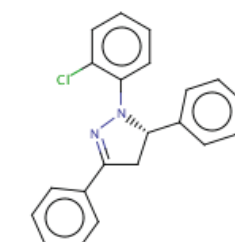
85: BAS 01917998



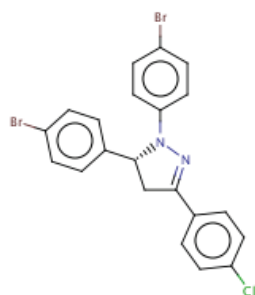
86: ASN 03367818



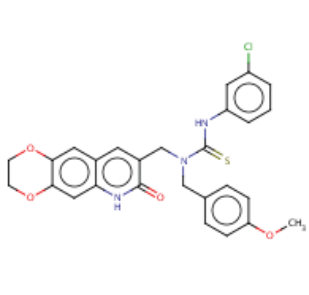
87: ASN 02070002



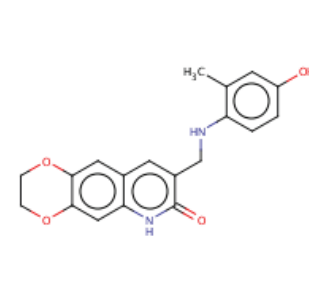
88: AG-690/33392024



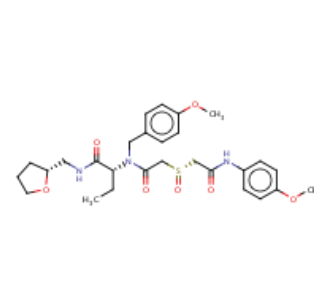
89: BAS 02767775



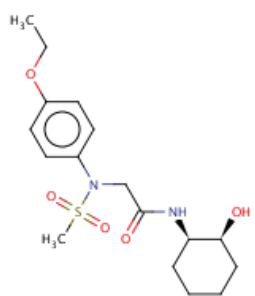
90: ASN 03776673



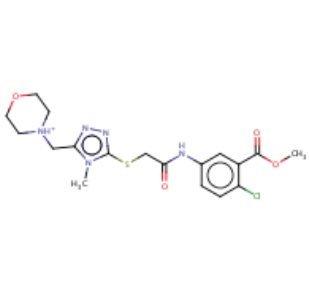
91: ASN 07404614



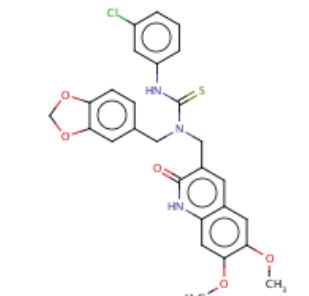
92: ASN 04172156



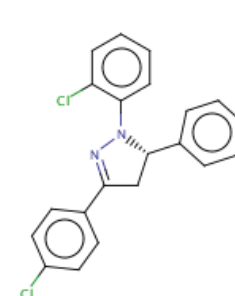
93: BAS 01979263



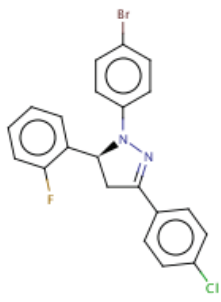
94: BAS 07228965



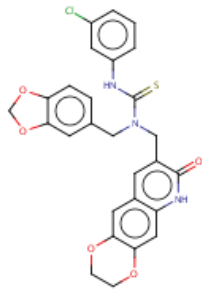
95: ASN 03778837



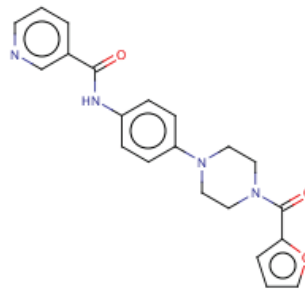
96: BAS 00126026



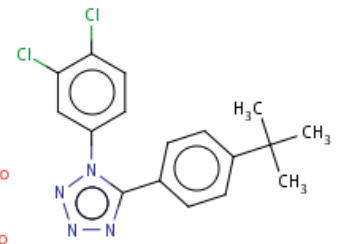
97: BAS 02800827



98: ASN 03776701



99: AP-970/43375665



100: BAS 01842726

Anhang F: Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA

Teile der vorliegenden Arbeit wurden in der Publikation „Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA“ (Löwer *et al.*, 2008) bereits veröffentlicht. Dieser Zeitschriftenartikel ist auf den nächsten Seiten eingefügt.

Zusammenfassung auf Deutsch

Exportierte Proteasen von *Helicobacter pylori* (*H. pylori*) sind potentiell an krankheits-zugehörigen Störungen beteiligt, die zu Entzündungen und Neoplasien in Magen führen. Durch eine vergleichende Sequenzanalyse des Proteoms von *H. pylori* wurden Gene gefunden, die Kandidaten für die Expression von sezernierten Proteasen sind. Es wurde eine kaseinolytische Aktivität bei einigen dieser Proteasen festgestellt, welche unabhängig vom Typ IV Sekretionssystem (T4SS) von *H. pylori* abgegeben werden. Das T4SS wird durch die *cag pathogenicity island* (*cagPAI*) kodiert. Unter diesen Proteasen ist die vorhergesagte Serinprotease HtrA (*hp1019*), welche bereits im bakteriellen Sekretom von *H. pylori* nachgewiesen wurde. Weiterhin konnte gezeigt werden, dass die Gene *hp1018* und *hp1019* höchstwahrscheinlich ein einziges Gen bilden, welches für ein exportiertes Protein kodiert. Die proteolytische Aktivität von HtrA wurde direkt *in-vitro* gezeigt, HtrA wurde in Zymogrammen durch Massenspektroskopie nachgewiesen. Die rekombinant exprimierte und gereinigte Protease HtrA zeigt eine ausgeprägte proteolytische Aktivität, welche durch eine Mutation der Aminosäure Serin 205 zu Alanin inaktiviert wird. Diese Daten zeigen, dass *H. pylori* HtrA als aktive Protease sezerniert. HtrA könnte daher ein neuartiges Ziel für therapeutische Strategien einer Infektionsbehandlung sein.

Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA

Martin Löwer¹, Christiane Weydig², Dirk Metzler³, Andreas Reuter⁴, Anna Starzinski-Powitz¹, Silja Wessler^{2,3}, Gisbert Schneider^{1,3*}

1 Goethe-University, Institute of Cell Biology and Neuroscience / CMP, Frankfurt am Main, Germany, **2** Junior Research Group, Paul-Ehrlich Institute, Langen, Germany, **3** Goethe-University, Institute of Computer Science, Frankfurt am Main, Germany, **4** Paul-Ehrlich Institute, Department of Allergology, Langen, Germany

Abstract

Exported proteases of *Helicobacter pylori* (*H. pylori*) are potentially involved in pathogen-associated disorders leading to gastric inflammation and neoplasia. By comprehensive sequence screening of the *H. pylori* proteome for predicted secreted proteases, we retrieved several candidate genes. We detected caseinolytic activities of several such proteases, which are released independently from the *H. pylori* type IV secretion system encoded by the *cag* pathogenicity island (*cagPAI*). Among these, we found the predicted serine protease HtrA (Hp1019), which was previously identified in the bacterial secretome of *H. pylori*. Importantly, we further found that the *H. pylori* genes *hp1018* and *hp1019* represent a single gene likely coding for an exported protein. Here, we directly verified proteolytic activity of HtrA *in vitro* and identified the HtrA protease in zymograms by mass spectrometry. Overexpressed and purified HtrA exhibited pronounced proteolytic activity, which is inactivated after mutation of Ser205 to alanine in the predicted active center of HtrA. These data demonstrate that *H. pylori* secretes HtrA as an active protease, which might represent a novel candidate target for therapeutic intervention strategies.

Citation: Löwer M, Weydig C, Metzler D, Reuter A, Starzinski-Powitz A, et al. (2008) Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA. PLoS ONE 3(10): e3510. doi:10.1371/journal.pone.0003510

Editor: Raphael H. Valdivia, Duke University Medical Center, United States of America

Received: June 25, 2008; **Accepted:** September 30, 2008; **Published:** October 23, 2008

Copyright: © 2008 Löwer et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften Frankfurt am Main (Germany), the Centre for Membrane Proteomics (CMP) Frankfurt am Main (Germany), the Jürgen-Manchot Stiftung, the Paul-Ehrlich Institut Langen (Germany), and the Deutsche Forschungsgemeinschaft (SFB-579, project A11).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gisbert.schneider@modlab.de

These authors contributed equally to this work.

Introduction

The mucosal epithelium in the human stomach forms the first barrier that prevents infiltration of pathogens into the host organism. The human pathogen *H. pylori* developed efficient strategies to colonize the gastric epithelium as a unique niche, where it induces the disruption of the epithelial layer contributing to inflammatory diseases (e.g. chronic gastritis, ulceration), mucosa-associated lymphoid tissue (MALT) lymphoma and gastric cancer in humans [1,2]. More virulent *H. pylori* strains express a combination of key disease-associated virulence factors allowing successful colonization in the stomach [3]. Among those, *H. pylori* harbors *cag* pathogenicity island (*cagPAI*), which encodes a type IV secretion system (T4SS) to inject the bacterial CagA (cytotoxin-associated gene A) oncoprotein into host cells [4]. *In vitro*, translocated CagA can strongly enhance the disruption of intercellular adhesions [4,5]. This process is believed to contribute to inflammation, carcinogenesis and invasive growth. Although the cellular aspects of CagA have been investigated intensively, the complex mechanisms of the actual interaction of *H. pylori* and the human epithelium are not fully understood yet.

Many pathogens developed elegant mechanisms for tissue destruction by secreting proteins with proteolytic activity.

Exported bacterial enzymes can directly activate host *pro*-matrix-metalloproteinases (*pro*-MMPs) representing a biochemical efficient way for matrix degradation. An example is set by the wide range of proteases of the thermolysin family secreted by *Pseudomonas aeruginosa* and *Vibrio cholera* that activate *pro*-MMP-1, -8, and -9 [6]. It has been further observed that serine proteases associated with lipopolysaccharides can induce MMP-9 activity in macrophages [7]. MMP-9 cleavage was also detected by a secreted zinc metalloproteinase (ZmpC) from *Streptococcus pneumoniae*, which indicates that ZmpC may play a role in pneumococcal virulence and pathogenicity in the lung [8].

Proteases might also play a role in *H. pylori* pathogenesis, and protease secretion has already been described for this organism [9]. *H. pylori* sheds an unknown protease that efficiently degrades PDGF (platelet derived growth factor) and TGF- β (transforming growth factor beta), which can be inhibited with sulglycotide [10]. Some features present in the primary sequence of *H. pylori* virulence factor vacuolating cytotoxin A (VacA) are reminiscent of serine proteases [11], although the predicted proteolytic activity of VacA has not been detected yet. In 1997, a *H. pylori* metalloproteinase with a native molecular size of approximately 200 kDa was discovered, which was secreted when *H. pylori* was grown in liquid culture [12]. The authors hypothesized that

surface expression of this metalloprotease activity may be involved in proteolysis of a variety of host proteins *in vivo* and thereby contribute to gastric pathology [12]. Importantly, *H. pylori* secretes a collagenase, encoded by *hp0169*, which might represent an essential virulence factor for *H. pylori* stomach colonization [13]. The predicted serine protease and chaperone HtrA (Hp1019) was previously identified as an extracellular protein of *H. pylori* [14], but its proteolytic role and substrates are still unknown.

As 658 of the 1,576 identified genes of the *H. pylori* genome [15] are annotated as “hypothetical” or as bearing a hypothetical function [16], we aimed at the identification of *H. pylori* genes possibly coding for secreted proteases by combining genomic data analysis with zymography. Indeed, we found that *H. pylori* secretes unknown proteins exhibiting caseinolytic activity. By calculating similarities to known proteases and using localization prediction methods, we inferred function and localization of these hypothetical *H. pylori* proteins. We also identified a sequencing error in the *hp1018* gene, which after correction encodes for a signal peptide for the putative serine protease HtrA (Hp1019). Eventually, we verified proteolytic activity of HtrA in biochemical approaches. The present study demonstrates the usefulness of sequence-based genome mining for potential drug targets representing one possible route for the prevention of matrix degradation of the mucosal epithelium by *H. pylori* and other pathogens.

Results and Discussion

H. pylori secretes caseinolytic proteases

Data are accumulating that bacteria secrete proteases with functional roles in microbial pathogenesis, but knowledge of *H. pylori*-secreted proteases and their functions is still limited. To analyze whether *H. pylori* actually secretes proteases, we performed casein zymography to monitor proteolytic activity in the supernatants of *H. pylori* lysates (Figure 1A, lane 1) and *H. pylori* culture medium (Figure 1A, lane 2). At least three casein-cleaving proteases were exported by *H. pylori* exhibiting apparent molecular weights of approximately 170 kDa, 140 kDa, and 50 kDa

(Figure 1A, lane 2). Interestingly, the protein band pattern present in the supernatant of the *H. pylori* medium obviously differs from the equivalent *H. pylori* lysate (Figure 1A lanes 1). The detected 170 kDa protease present in the supernatant of *H. pylori* (BHI Hp) consistently migrated slightly faster than in the *H. pylori* lysate, while the 140 kDa protein was only present in the supernatant, but absent in the lysate of *H. pylori* (Hp son). In contrast to the double band detected in the lysates, we observed only a single proteolytic activity in the supernatant (Figure 1A, lanes 1–2). These data indicate that the export of the proteases might occur *via* active signal peptide-dependent translocation, rather than being an artifact of bacterial autolysis in the *H. pylori* liquid culture.

Since *H. pylori* encodes a well-described T4SS and the T4SS-independently secreted pathogenic factor VacA with a hypothesized protease function [11], we also included supernatants of isogenic *H. pylori* mutants which are deficient of T4SS and CagA (Δ PAI, Figure 1B, lane 3), or VacA (Δ VacA, Figure 1B, lane 4), and compared them with the *H. pylori* wildtype strain (wt, Figure 1B, lane 2) and *H. pylori*-free culture medium (-, Figure 1B, lane 1). Compared to the wildtype strain, the Δ PAI mutant showed the same secretion pattern of proteins with caseinolytic activity in the extracellular space suggesting that their secretion might occur independently from the T4SS (Figure 1B, lane 3). Although initial publications indicated a predicted serine protease activity of the pathogenic factor VacA [11], we can exclude a caseinolytic effect of VacA since the isogenic *vacA*-deficient *H. pylori* mutant showed a similar pattern of proteases (Figure 1B, lane 4). Gelatin zymographies were also performed by us and clearly demonstrated the lack of gelatinolytic *H. pylori* proteases (not shown). A positive result here would have demonstrated a closer link to matrix degradation, as gelatin is a product of collagen, a major extracellular matrix protein.

So far, the identity of the detected *H. pylori* proteases was unknown. A previously described multi-metalloprotease-like complex secreted by *H. pylori* with a molecular weight of about 200 kDa [12] might be an explanation for the largest protein seen in the zymogram, since its size is four to six times greater than

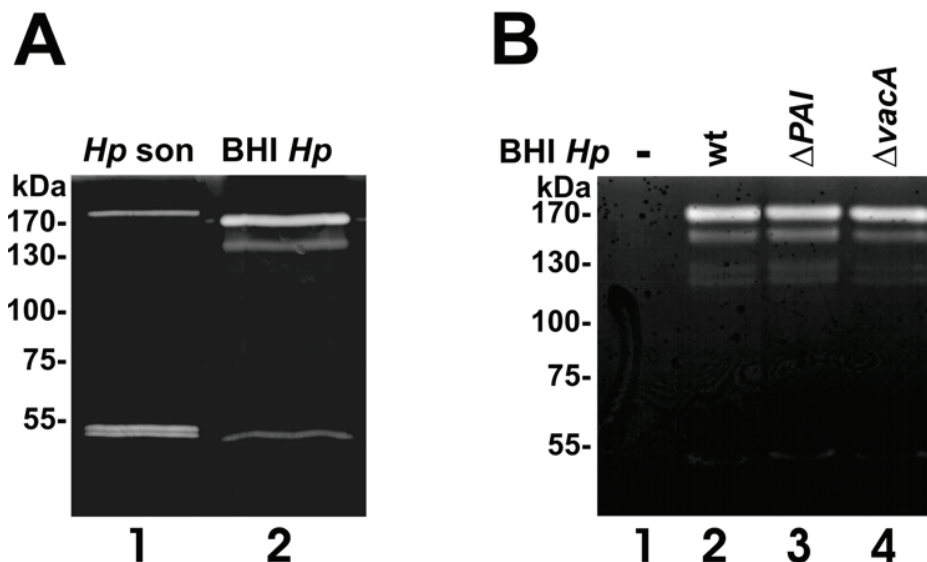


Figure 1. *H. pylori* secretes bacterial factors with caseinolytic activities. (A) The *H. pylori* strain 26695 was grown in protein-free BHI medium. After 48 hours, the bacteria were harvested and lysed by sonification (Hp son). 30 μ l aliquots of the supernatants (BHI Hp) and bacterial lysates were separated by casein zymography and analyzed proteolytic activities. (B) *H. pylori* strains wild type (P12, wt), Δ PAI, and Δ VacA were grown in protein-free liquid growth medium. 30 μ l of the aliquots of the medium were analyzed in casein zymograms for proteolytic activities. doi:10.1371/journal.pone.0003510.g001

comparable proteases of other Gram-negative bacteria [12]. Also, protease DegP of *Escherichia coli*, which is a homolog of Hp1019 from *H. pylori*, was shown to form hexamers when crystallized [17]. Therefore, as zymography was performed under non-reducing conditions, the upper band(s) might result from smaller proteins forming a macromolecular complex.

In silico genome screening for candidates of *H. pylori* secreted hypothetical proteases

Based on the finding that *H. pylori* actively secretes proteases, we then aimed to identify suitable candidates by *in silico* analysis. Thus, we compared the *H. pylori* proteome to a set of known proteases from various organisms using sequence alignment techniques. A reference set of known proteases containing 3,566 amino acid sequences was compiled from the UniProtKB/SwissProt database (version 6.7) [18], which served as queries for exhaustive pairwise alignment to genomic and protein sequence data of *H. pylori* strain 26695 with 1,576 annotated genes from the NCBI RefSeq database (accession number NC_000915) [16]. For the 1,576 putative *H. pylori* proteins, 75,524 local alignments were returned by the BLAST algorithm [19]. Alignments yielding an *E*-value ≤ 0.5 were selected and divided into four classes:

- Class A: alignments showing complete conservation of the active-site region,
- Class B: alignments showing partial conservation of the active-site region,
- Class C: alignments with proteases lacking an active-site annotation, and
- Class D: all other alignments.

The latter (class D) were not further examined. Information about the localization of the active sites was retrieved from the feature tables of the respective SwissProt entries [18].

Then, we predicted protein localization using prediction systems, which are publicly available on the World Wide Web: SignalP [20], SecretomeP [21], Phobius [22], CELLO [23], PA-SUB [24], and PSORTb [25]. All systems are capable or explicitly designed to analyze amino acid sequences from Gram-negative bacteria. Alignments were selected for further examination when the corresponding predictions for a protein sequence matched one or more of the following criteria:

- i) predicted extracellular localization (CELLO, PSORTb, PA-SUB),
- ii) predicted signal peptide (SignalP),
- iii) predicted signal peptide, but no transmembrane helices (Phobius), and a SecretomeP *score* ≥ 0.5 .

By filtering the alignments with respect to the active residues marked in the sequence of reference protease and the localization prediction, we obtained 47 class A, 39 class B and 32 class C proteins (*vide supra*) and their corresponding genes. The best-scoring alignments of those proteins to proteases of the reference set were manually inspected. Among those, nine genes have not been described to code for *H. pylori* proteases yet, but can be aligned with a statistically significant score to proteases of the reference protease sequences (Table 1). Interestingly, the putative translation products of genes *hp0289*, *hp0609*, and *hp0922* form a group of paralogs to VacA cytotoxin [26], which can be seen in a multiple sequence alignment (not shown). Structural similarities of VacA to extracellular IgA proteases of *Haemophilus influenzae* have been described previously [11]. Pairwise sequence identity of these

VacA paralogs to VacA ranges between 25% and 30% which also include the C-terminal autotransporter sequence [27]. As this sequence is sufficient to translocate the N-terminal part of the protein across the outer membrane – which is often followed by an autoproteolytic event to release the translocated part into the extracellular space [27] – it seems likely that some of these proteins possess a proteolytic function.

Finally, although we could not detect caseinolytic activity of VacA in our casein zymography study, we cannot exclude an effect of VacA and its paralogs on other substrates *per se*. However, the alignments do not reveal conserved active site residues in VacA paralogs. Still they might represent autoproteolytic autotransporter proteins without common protease motifs which have been reported already [27]. Notably, their precursor proteins have a molecular weight of 136 to 311 kDa (according to SwissProt entries O25063, O25330 and O25579) which is in accordance with the molecular weights we observed in the zymography after a possible cleavage of the N-terminal signal peptide and the autotransporter sequence.

H. pylori harbors five genes that are described in the literature and/or database annotations to code for potential extracellular proteases (Table 1). Processing protease YmxG (Hp0657) and protease pqqE (Hp1012) are predicted to possess a signal peptide (Table 1) and to be extracellular or outer membrane-bound. The protease coded by the gene *hp1350* could be extracellular, as SecretomeP and PA-Sub vote for this localization and the existence of a signal peptide is also predicted (Table 1). The product of *hp1019*, which is annotated as a serine protease in the respective GenBank file, seems to be a homologue to heat shock protein HtrA from *Escherichia coli*. Its active site is fully conserved, and the extracellular localization has been determined previously [14]. The gene product of *hp1584* is annotated as a sialoglycoprotease (gcp). Its amino acid sequence does not contain known export motifs, and the amino acid composition is predicted to be cytoplasmic. However, the PA-SUB and PSORTb predictors categorized the protein as extracellular (Table 1) based on the extracellular localization of the homologous o-sialoglycoprotein endopeptidase of *Mannheimia haemolytica* (SwissProt identifier GCP_PASHA), which also lacks an N-terminal targeting signal [28]. In fact, very recently Hp0657, Hp1012, Hp1019, and Hp1350 have been identified in the extracellular *H. pylori* proteome [29] indicating the high specificity of our bioinformatical prediction of hypothetical extracellular *H. pylori* proteases (Table 1).

Since we demonstrated that several caseinolytic proteases are secreted by *H. pylori* independently of functional T4SS, it is likely that other secretion systems exist. This is underlined by our observation that nine out of 14 genes either contain a signal peptide, which only explains a transportation to the periplasm, or receive a high SecP prediction score (Table 1). We stress that these predicted features are common for extracellular proteins but do not explain a possible transport pathway. Thus one can speculate that a secretory machinery not yet attributed to *H. pylori*, or entirely novel ones, might be involved which require export signals of an unknown nature. For example, *H. pylori* might involve a specific type I (ABC) or a type III transportation system.

H. pylori HtrA is an active protease

We were then interested in answering the question whether one of the predicted *H. pylori* proteases accounts for the observed proteolytic activity. In a first step, concentrated *H. pylori* lysates were separated by zymography under non-reducing conditions followed by protein elution of proteins from the negatively stained protein bands I and II (Figure 2A). Then, eluted proteins were concentrated and separated by a denaturing SDS PAGE

Table 1. Results of the Blastp search for proteases genes, of the localization prediction and calculated molecular weights (MW).

locus tag	description	matrix	E-Value	bitscore	% id.	class	Calculated MW (kDa)	CELLO	PA-SUB	Phobius	PSORTb	SecP	SignalP
hp0289	toxin-like outer membrane protein	blo62	0.009	38.9	22.3	A	31.12	EX	EX	+	EX	+	+
hp0506	conserved hypothetical secreted protein	blo80	8×10^{-14}	72.8	40.7	A	45.97	OM	EX	1	U		+
hp0608	hypothetical protein	pam250	7×10^{-4}	36.9	23.6	A	17.94	OM	-		IM		
hp0609	hypothetical protein	pam250	0.006	37.9	17.7	A	13.51	EX	EX, OM		U		+
hp0657	processing protease (ymxG)	blo45	4×10^{-9}	56.7	20.7	A	48.80	OM	EX	+	U		+
hp0922	toxin-like outer membrane protein	blo80	0.080	35.7	30.5	A	27.46	EX	EX	+	EX		+
hp0980	conserved hypothetical secreted protein	pam30	3×10^{-12}	65.5	38.6	A	11.41	PP	EX, IM	+	IM		+
hp1012	protease (pqqE)	blo62	1×10^{-20}	95.5	23.8	A	50.33	OM, PP	EX	+	OM		+
hp1019	serine protease (htrA)	blo80	4×10^{-82}	299	42.2	A	47.98	EX	EX, PP		PP		+
hp1037	hypothetical protein	pam70	0.061	33.3	30.7	A	40.80	CP	CP, EX		CP		
hp1350	protease	blo80	7×10^{-81}	295	39.7	A	50.55	CP	EX	+	U		+
hp1543	toxR-activated gene (tagE)	blo80	1×10^{-16}	81.5	43.5	A	35.63	OM	EX	1	U		
hp1544	toxR-activated gene (tagE)	blo62	5×10^{-8}	52.8	34.0	A	34.93	CP	EX,OM,PP	1	U		
hp1584	sialoglycoprotease (gcp)	blo80	1×10^{-39}	158	36.4	C	37.80	CP	EX		EX		

Locus tags and descriptions were taken from the corresponding GenBank entries. The columns "matrix", "E-Value", "bitscore" and "% id." list the alignment data of the according highest scoring alignment. The column "class" refers to our definition of alignment classes. Molecular weight was calculated by a program hosted on the ExPASy server. The columns "CELLO", "PA-SUB", and "PSORTb" give the classifications according to the prediction software. The column "Phobius" gives the number of transmembrane helices and a plus (+) sign if a signal peptide was found. Columns "SecP" and "SignalP" contain a plus sign for a SecretomeP output ≥ 0.5 or prediction of a signal peptide, respectively. CP = cytoplasm, IM = inner membrane, PP = periplasm, OM = outer membrane, EX = extracellular, U = unknown. doi:10.1371/journal.pone.0003510.t001

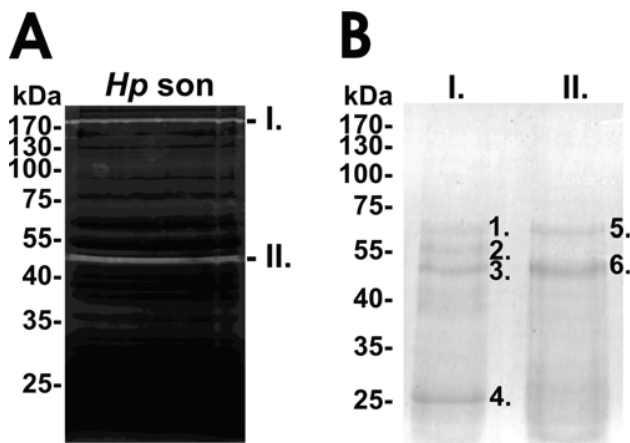


Figure 2. Identification of *H. pylori* proteases. (A) For a preparative analyses, 18×10^9 bacteria were lysed and analyzed by zymography. The upper (1) and lower (2) negatively stained protease bands were excised, proteins were eluted and separated by SDS PAGE and Coomassie staining (B). Indicated protein bands were analyzed by mass-spectrometry.

doi:10.1371/journal.pone.0003510.g002

(Figure 2B). We detected four different proteins in the Coomassie-stained SDS PAGE, which were isolated from protein band I in the zymogram (compare Figure 2A, band I and Figure 2B, lane I). Electrophoretic separation of proteins from protein band II (Figure 2A) resulted in two different proteins (Figure 2B, lane II). The identity of these proteins was determined by MALDI-TOF-MS. The accession number, denomination and a summary of the MS data are presented in Table 2. The results of the MS analyses are shown for a single database entry for each band. However, due to the high degree of sequence identity between proteins isolated from different *H. pylori* strains significant hits were obtained also for other urease and serine proteases, e.g. serine protease from *H. pylori* strain J99 or Ure B from database entry gi/51989332.

Hp1018 encodes a signal peptide for an active Hp1019 protease

Hp1019 has been previously predicted as a secreted *H. pylori* protease with unknown function [14,29]. However, its proteolytic activity had not been demonstrated. Considering the protein sequence of *H. pylori* HtrA, it lacks an annotated N-terminal signal

peptide, in contrast to HtrA of *E. coli*. The gene *hp1019* has an N-terminal overlap with the adjoining gene *hp1018*, which is 147 bases long and in a different reading frame. It has been suggested before that those genes might belong together [30]. Thus, we re-sequenced the gene *hp1018* and aligned it to the published genomic data of *H. pylori* Hp26695 (Figure 3). Here, we demonstrate that *hp1018* reveals a wrongly sequenced guanine at position 1081558 of the published genome of *H. pylori* strain 26695. We conclude from our data that the translation of Hp1018 actually contains a signal peptide-like sequence (SignalP score > 0.99) at its N-terminus, and it is most likely that Hp1018 represents the N-terminal part of Hp1019 resulting in a new sequence with 475 amino acids.

To prove proteolytic activity of Hp1018/19 for the first time, we fused the *hp1018/19* gene lacking the putative signal peptide to the glutathione-S-transferase (*gst*) gene and transformed the construct into *E. coli* BL21 to express the recombinant protein (Figure 4A). Both, induction and enrichment of GST-Hp1018/19 Δ sp protein were analyzed by Coomassie-stained SDS PAGEs (Figure 4B). During GST-Hp1018/19 Δ sp preparation, contaminating proteins were co-purified, which were identified by MALDI-TOF-MS as glutathione-S-transferase and degradation products of HtrA. Accordingly, it had been demonstrated that *E. coli* encoded HtrA is an endopeptidase [31]. To remove the GST tag from the fusion protein, GST-Hp1018/19 Δ sp coupled to GST sepharose was incubated with PreScission protease resulting in the release of Hp1018/19 Δ sp protein (Figure 4B, lane 6).

Purified proteins were then probed for proteolytic activity (Figure 4C). The GST-Hp1018/19 Δ sp proteins were bound to GST sepharose, washed and eluted using reduced glutathione. As a control, we cloned and purified the Hp1018/19 Δ sp^{S205A} protein in which serine-205 was mutated to alanine in the presumable active center of HtrA. As expected, we observed casein degradation by GST-Hp1018/19 Δ sp protein (Figure 4C, lane 3), but not by the GST-Hp1018/19 Δ sp^{S205A} (Figure 4C, lanes 1–2). This finding demonstrates that *H. pylori* HtrA actually is an active protease, which can be inactivated by mutation of serine-205. In parallel, we cloned and purified Hp0506, Hp0657, Hp1012, Hp1037, Hp1543, and Hp0169, which previously had been described as a collagenase [13]. With the exception of Hp1019, we did not detect any proteolytic activities using casein as a substrate in zymography studies (data not shown). Therefore, we conclude that the observed caseinolytic activities were actually mediated by Hp1018/19.

Table 2. Proteins that were identified by mass-spectrometry (cf. Figure 2B).

Band	Accession number	Protein name	Sample	Number of matched peptides	MASCOT Score	Sequence coverage
1	gi 19338960	urease B subunit [H.pylori]	A ^a	9	69 ^b	25%
			B	8	98	20%
2	n.d. ^c					
3	gi 15645633	Serine protease (htra) [H.pylori 26696]	A	10	91	30%
			B	12	88	29%
4	n.d.					
5	n.d.					
6	gi 15645633	Serine protease (htra) [H.pylori 26696]	A	10	84	30%
			B	10	88	25%

^aresults of two independently processed samples; ^bprotein score was below the level that indicates a *p*-value of <0.05, ^cnot determined.

doi:10.1371/journal.pone.0003510.t002

```

Query: 1      acgattgtttctgctggtatttggtagactatctttacacaaaaagctaata 60
            |||
Sbjct: 1081309 acgattgtttctgctggtatttggtagactatctttacacaaaaagctaata 1081368

Query: 61     aataacararaagatttgakataatcttttcttaattttgaagtttagcaaattttaagg 120
            |||
Sbjct: 1081369 aataacaaaaaagatttgatataatcttttcttaattttgaagtttagcaaattttaagg 1081428

Query: 121    aagtaaccatgatgaaaaaacctttttatctctttggcttagcggttaagcttgaatg 180
            |||
Sbjct: 1081429 aagtaaccatgatgaaaaaacctttttatctctttggcttagcggttaagcttgaatg 1081488
                    M K K T L F I S L A L A L S L N A

Query: 181    cgggcaatatccaaatccagagcatgccaaagttaagagcgagtgagtgtcccctcta 240
            |||
Sbjct: 1081489 cgggcaatatccaaatccagagcatgccaaagttaagagcgagtgagtgtcccctcta 1081548
                    *G N I Q I Q S M P K V K E R V S V P S K
                               V S P L ?

Query: 241    aagacgatac-gatctattcttaccacgattctattaaggactctattaaggcgggtgg 297
            |||
Sbjct: 1081549 aagacgatacgggatctattcttaccacgattctattaaggactctattaaggcgggtgg 1081606
                    D D T D L F L P R F Y Stop
                    K T I R I Y S Y H D S I K D S I K A V V

```

Figure 3. Blastn alignment of the re-sequenced nucleotide sequence (query) with the original genomic sequence (subject) of *hp1019*. The annotated gene *hp1018* is marked in grey. The letter 'r' represents ('a' OR 'g'), while the letter 'k' represents ('t' OR 'g'). The inserted guanidine is printed white on black. Numbers give residue positions. The amino acid translation is given in single letter code for Hp1018, starting at position 1081440, and for Hp1019, starting at position 1081537. The predicted most likely signal peptidase cleavage site between the amino acids LNA and GNI is marked with an asterisk. The underlined part of the amino acid sequences will not be part of the translation if the marked guanidine is removed.

doi:10.1371/journal.pone.0003510.g003

As shown by mass spectrometry, we also co-purified processed HtrA variants with GST-Hp1018/19Δsp (Figure 4B). We detected proteolytic activity of these proteins in casein zymography (Figure 4C). We therefore assume that processed variants of HtrA formed multimers with GST-Hp1018/19Δsp during the purification steps. This suggestion is supported by the finding that removal of the GST tag from GST-Hp1018/19Δsp protein led to the formation of the 170 kDa protease (Figure 4C, lane 4), which was not detected after purification of Hp1018/19Δsp^{S205A} (Figure 4C, lane 2). Together with our analysis showing that HtrA was present in the upper and lower protein bands (Figure 2), we conclude from our data that HtrA might also be active as a multimer.

Conclusions

The complex mechanisms how *H. pylori* strongly induce inflammatory responses and invasive growth leading to the disruption of the human epithelium are still unclear. Although exported proteases of pathogens represent extensively studied virulence factors, not much is known about their involvement in *H. pylori*-associated pathogenesis. This comprehensive analysis of the *H. pylori* strain 26695 genome by sequence analysis and activity prediction methods revealed several genes coding for putative proteases. Among those, we identified the HtrA from *H. pylori* as a secreted enzyme exhibiting proteolytic activity. We also found that HtrA forms proteolytically active multimers, which is consistent with an earlier report of Windle and colleagues who demonstrated that *H. pylori* secretes a metalloprotease with a native molecular size of approximately 200 kDa and speculated whether this metalloprotease activity may be involved in proteolysis of a variety of host

proteins *in vivo* and thereby contributes to gastric pathology [12]. The *E. coli* homologue HtrA functions as a heat shock protein, although it cannot be excluded that Hp1019 represents a so-called “moonlighting” protein [32], serving a function both in the periplasm in heat shock degradation and the extracellular matrix as a virulence factor. In fact, a secreted collagenase Hp0169 was identified as an important virulence factor for *H. pylori* colonization [13]. Although the biological function of Hp0169 and the recently detected extracellular proteases Hp0657, Hp1012, and Hp1350 [29] are unknown, it underscores the potential importance of secreted bacterial proteases in *H. pylori* mediated pathogenesis, which represent attractive vaccine and drug target candidates.

Materials and Methods

Homology search

Proteases were compiled for the reference data set by selecting all entries from the UniProtKB/SwissProt database (version 6.7) [18] containing the keyword “protease”, but lacking the phrases “inhibitor*”, “probable*”, “fragment*”, “hypothetical*”, “putative*”, “possible*” or “predicted*” in the keyword and description fields, where the asterisk is a wildcard for any arbitrary suffix. The NCBI BLAST package was employed for pairwise sequence alignment [19]. The Blastp program was used for protein-protein comparison. Tblastn was used to compare the whole DNA sequence of *H. pylori* strain 26695 with the protease sequence set. The substitution matrices PAM30, PAM70, PAM250, BLOSUM45, BLOSUM62 and BLOSUM80 were used with default parameter settings (e.g. scoring penalties, window size).

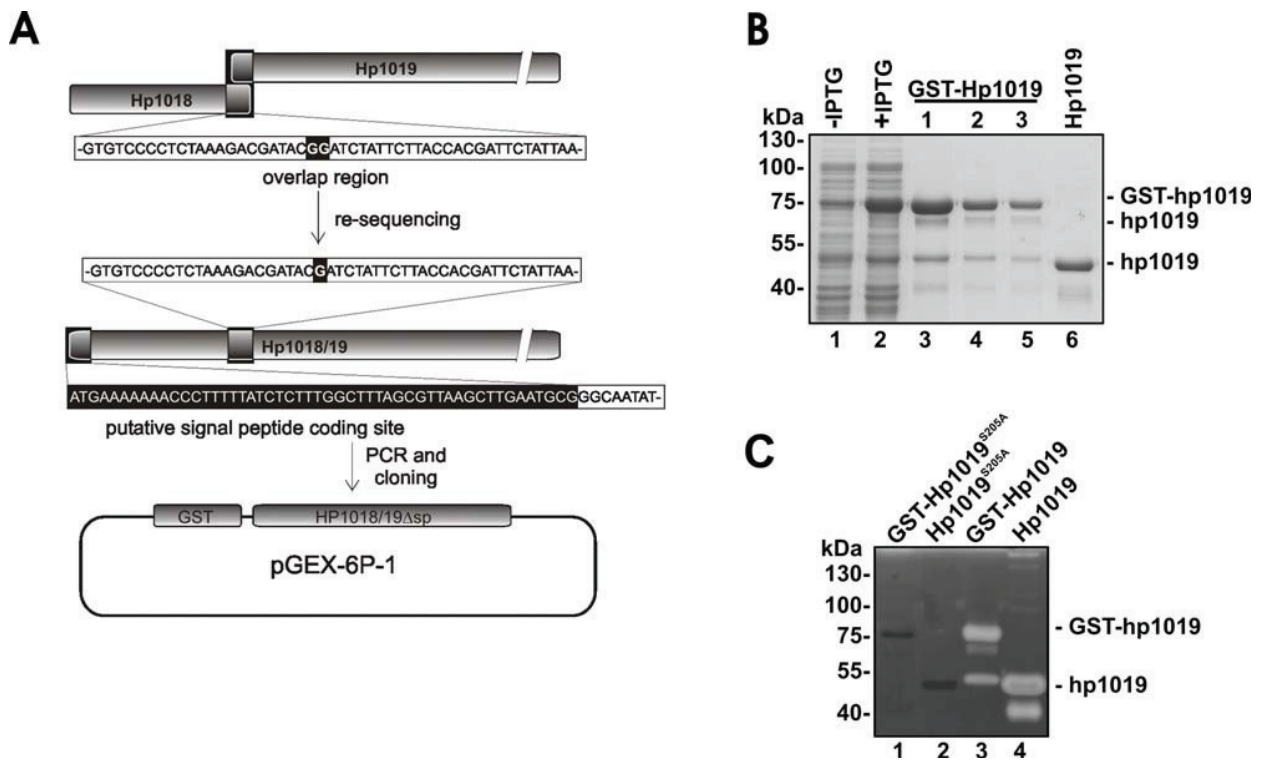


Figure 4. Proteolytic activity of the Hp1018/19 protein. (A) For the construction of the GST-Hp1018/19 Δ sp fusion protein, the re-sequenced Hp1018/19 gene was amplified without the putative signal peptide and cloned into the pGEX-6P-1 vector. (B) The *gst-hp1018/19 Δ sp* construct was transformed in *E. coli* for overexpression and total protein extracts from untreated (lane 1) and IPTG-induced *E. coli* (lane 2) were separated by SDS PAGE. Overexpressed GST-Hp1018/19 Δ sp was precipitated using glutathione sepharose and released by three elution steps (lanes 3–5). To remove the GST tag, GST-Hp1018/19 Δ sp bound to glutathione sepharose were treated with PreScission protease and 30 μ g protein of the supernatant containing the Hp1018/19 (lane 6) were loaded on a SDS PAGE followed by Coomassie staining. (C) Three μ g of purified GST-Hp1018/19 Δ sp^{S205A} (lane 1), GST-Hp1018/19 Δ sp (lane 3), PreScission protease-treated Hp1018/19 Δ sp^{S205A} (lane 2) and Hp1018/19 Δ sp (lane 4) were analyzed by casein zymography for proteolytic activity.
doi:10.1371/journal.pone.0003510.g004

Bacteria

H. pylori wildtype strains 26695 and P12, its isogenic mutant strains Δ VacA, and Δ PAI had been described before [11,33]. Bacteria were grown in protein-free liquid brain heart infusion (BHI) medium (Merck, Darmstadt, Germany) supplemented with β -cyclodextrin for 48 hours, which has been previously optimized for minimal autolysis of *H. pylori* cells [14]. Lysates of *H. pylori* were obtained by sonification in PBS containing 0.1% Triton X-100. Supernatants of *H. pylori* BHI cultures were sterilized by filtration (pore size 0.22 μ m).

Amplification and sequencing of hp1018

The gene *hp1018* was amplified from the genomic DNA of *H. pylori* strain 26695 by standard PCR using the Pfx DNA-polymerase (Invitrogen, Karlsruhe, Germany). The following primers were used: *hp1018*for: 5'-GGC TAT GGA TAA GGA TCA ACG C-3', *hp1018*rev: 5'-CCA CCG CCT TAA TAG AGT CCT T-3'. The PCR product, having a calculated length of 333 bases, was submitted to a commercial provider (GENterprise, Mainz, Germany) for sequencing.

Cloning, mutation and purification of HtrA

The construct Hp1018/19 Δ sp was amplified from genomic DNA of *H. pylori* strain 26695 using the primers 5'-aaggatccgg-caatatccaaatccagagcatg-3' and 5'-aagaattcgaccaccacctatcattacc-3' with Pfx DNA polymerase in supplied buffer with 2 \times PCR

Enhancer (Invitrogen). The amplified BamHI/EcoRI flanked PCR product was then ligated into the pGEM-T Easy plasmid (Promega), subcloned into the pGEX-6P-1 plasmid DNA (GE Healthcare Life Sciences) and transformed in *E. coli* BL21. The construction of the protease-inactive Hp1018/19 Δ sp^{S205A} protein, serine 205 was mutated to alanine using the QuikChange[®] Lightning Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions. For heterologous overexpression and purification of GST-Hp1018/19 Δ sp, transformed *E. coli* was grown in 500 ml TB medium to an OD₅₅₀ of 0.6 and the expression was induced by the addition of 0.1 mM isopropylthiogalactosid (IPTG). The bacterial culture was pelleted at 4000 \times g for 30 minutes and lysed in 25 ml PBS by sonification. The lysate was cleared by centrifugation and the supernatant was incubated with glutathione sepharose (GE Healthcare Life Sciences) at 4 $^{\circ}$ C over night. The fusion protein was either eluted with 10 mM reduced glutathione for 10 minutes at room temperature or cleaved with 180 U PreScission Protease for 16 h at 4 $^{\circ}$ C (GE Healthcare Life Sciences). Elution and cleavage products were analyzed by SDS PAGE and zymography.

Zymography and protein elution

Undiluted aliquots were loaded onto 8% SDS-PAGE containing 0.1% casein (Invitrogen, Germany) and separated by electrophoresis. After separation, the gel was re-natured in 2.5% Triton X-100 solution at room temperature for 60 min with

gentle agitation, equilibrated in developing buffer (50 mM Tris-HCl, pH 7.4; 200 mM NaCl, 5 mM CaCl₂, 0.02% Brij35) at room temperature for 30 min with gentle agitation, and incubated overnight at 37°C in fresh developing buffer. Transparent bands of caseinolytic activity were visualized by staining with 0.5% Coomassie Blue R250. For identification of proteases present in zymographies, negatively stained bands were excised from a preparative casein zymogram and proteins were eluted twice for 6 hours *via* D-Tube™ Dialyzers Maxi MWCO 6–8 kDa (Novagene). Eluated proteins were desalted and concentrated using Vivaspin columns from Sartorius (Germany).

Mass spectrometry

Eluated proteins from zymograms were separated by means of SDS-PAGE and stained with Coomassie for subsequent MS analysis in two independent experiments (A+B). In gel digestion was performed as previously described with several minor modifications [34]. Peptide mixtures were additionally purified and concentrated by using ZipTipC-18 tips (Millipore) according

to the manufacturers' instructions. The identity of HtrA and urease B was proven by mass spectrometry as described [35,36]. Briefly, Samples were mixed with peptide standard (peptide standard II, Bruker) and matrix (a saturated solution of HCCA in 50% ACN+0,5%TFA) at a ratio of 1:1:2; v:v:v), and with matrix only at a ratio of 1:1; v:v, and transferred on a ground steel target. Mass analysis was done on a Bruker Reflex II mass spectrometer with predefined default instrument settings. Proteins were identified by running MASCOT (<http://www.matrixscience.com>) against the entire NCBI database. Peptide tolerance was set to 50 ppm and a maximum of one missed cleavage site was allowed. A hit was considered as significant at a probability value of $p < 0.05$.

Author Contributions

Conceived and designed the experiments: ML ASP SW GS. Performed the experiments: ML CW AR. Analyzed the data: ML CW DM AR ASP SW GS. Contributed reagents/materials/analysis tools: DM AR. Wrote the paper: ML SW GS.

References

- Blaser MJ, Atherton JC (2004) *Helicobacter pylori* persistence: biology and disease. *J Clin Invest* 113: 321–333.
- Peek Jr RM, Crabtree JE (2006) *Helicobacter* infection and gastric neoplasia. *J Pathol* 208: 233–248.
- Rieder G, Fischer W, Haas R (2005) Interaction of *Helicobacter pylori* with host cells: function of secreted and translocated molecules. *Curr Opin Microbiol* 8: 67–73.
- Backert S, Feller SM, Wessler S (2008) Emerging roles of Abl family tyrosine kinases in microbial pathogenesis. *Trends Biochem Sci* 33: 80–90.
- Hatakeyama M (2006) *Helicobacter pylori* CagA - a bacterial intruder conspiring gastric carcinogenesis. *Int J Cancer* 119: 1217–1223.
- Okamoto T, Akaike T, Suga M, Tanase S, Horie H, et al. (1997) Activation of human matrix metalloproteinases by various bacterial proteinases. *J Biol Chem* 272: 6059–6066.
- Min D, Moore AG, Bain MA, Breit SN, Lyons JG (2002) Activation of macrophage promatrix metalloproteinase-9 by lipopolysaccharide-associated proteinases. *J Immunol* 168: 2449–2455.
- Oggioni MR, Memmi G, Maggi T, Chiavolini D, Iannelli F, et al. (2003) Pneumococcal zinc metalloproteinase ZmpC cleaves human matrix metalloproteinase 9 and is a virulence factor in experimental pneumonia. *Mol Microbiol* 49: 795–805.
- Gooz M, Gooz P, Smolka AJ (2001) Epithelial and bacterial metalloproteinases and their inhibitors in *H. pylori* infection of human gastric cells. *Am J Physiol Gastrointest Liver Physiol* 281: G823–G832.
- Piotrowski J, Slomiany A, Slomiany BL (1997) Suppression of *Helicobacter pylori* protease activity towards growth factors by sulglycyotide. *J Physiol Pharmacol* 48: 345–351.
- Schmitt W, Haas R (1994) Genetic analysis of the *Helicobacter pylori* vacuolating cytotoxin: structural similarities with the IgA protease type of exported protein. *Mol Microbiol* 12: 307–319.
- Windle HJ, Kelleher D (1997) Identification and characterization of a metalloprotease activity from *Helicobacter pylori*. *Infect Immun* 65: 3132–3137.
- Kavermann H, Burns BP, Angermuller K, Odenbreit S, Fischer W, et al. (2003) Identification and characterization of *Helicobacter pylori* genes essential for gastric colonization. *J Exp Med* 197: 813–822.
- Bumann D, Aksu S, Wendland M, Janek K, Zimny-Arndt U, et al. (2002) Proteome analysis of secreted proteins of the gastric pathogen *Helicobacter pylori*. *Infect Immun* 70: 3396–3403.
- Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388: 539–547.
- Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl Acids Res* 33: D501–D504.
- Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T (2002) Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature* 416: 455–459.
- Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). *Nucl Acids Res* 33: D154–D159.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25: 3389–3402.
- Bendtsen JD, Kiemer L, Fausboll A, Brunak S (2005) Non-classical protein secretion in bacteria. *BMC Microbiol* 5: 58.
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783–795.
- Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338: 1027–1036.
- Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Prot Sci* 13: 1402–1406.
- Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, et al. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 20: 547–556.
- Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, et al. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 21: 617–623.
- Cover TL, Blanke SR (2005) *Helicobacter pylori* VacA, a paradigm for toxin multifunctionality. *Nat Rev Microbiol* 3: 320–332.
- Henderson IR, Navarro-Garcia F, Nataro JP (1998) The great escape: structure and function of the autotransporter proteins. *Trends Microbiol* 6: 370–378.
- Abdullah KM, Lo RY, Mellors A (1991) Cloning, nucleotide sequence, and expression of the *Pasteurella haemolytica* A1 glycoprotease gene. *J Bacteriol* 173: 5597–5603.
- Smith TG, Lim JM, Weinberg MV, Wells L, Hoover TR (2007) Direct analysis of the extracellular proteome from two strains of *Helicobacter pylori*. *Proteomics* 7: 2240–2245.
- Pflock M, Dietz P, Schär J, Beier D (2004) Genetic evidence for histidine kinase HP165 being an acid sensor of *Helicobacter pylori*. *FEMS Micro Letters* 234: 51–61.
- Lipinska B, Zylicz M, Georgopoulos C (1990) The HtrA (DegP) protein, essential for *Escherichia coli* survival at high temperatures, is an endopeptidase. *J Bacteriol* 172: 1791–1797.
- Jeffery CJ (2003) Moonlighting proteins: old proteins learning new tricks. *Trends Genet* 19: 415–417.
- Wessler S, Hocker M, Fischer W, Wang TC, Rosewicz S, et al. (2000) *Helicobacter pylori* activates the histidine decarboxylase promoter through a mitogen-activated protein kinase pathway independent of pathogenicity island-encoded virulence factors. *J Biol Chem* 275: 3629–3636.
- Shevchenko A, Wilm M, Vorm O, Mann M (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* 68: 850–858.
- Hillenkamp F, Karas M (1990) Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Methods Enzymol* 193: 280–295.
- Albrecht M, Alessandri S, Conti A, Reuter A, Lauer I, et al. (2008) High-level expression, purification and physico-chemical characterisation of recombinant Pen a 1 - a major allergen of shrimp. *Mol Nutr Food Res*: in press.

Anhang G: Prediction of Type III Secretion Signals in Genomes of Gram-negative Bacteria

Parallel zu der vorliegenden Arbeit wurde eine bioinformatische Analyse von Sekretionssignalen für bakterielle Typ III Sekretionssysteme durchgeführt und publiziert. Der entsprechende englischsprachige Artikel (Löwer und Schneider, 2009) ist im Folgenden zu finden.

Zusammenfassung auf Deutsch

Hintergrund: Pathogene Bakterien können sowohl Tiere als auch Pflanzen infizieren. Sie benutzen verschiedene Mechanismen um Virulenzfaktoren über die eigenen Zellmembranen zu transportieren und Proteine in die Wirtszelle abzugeben. Das Typ III Sekretionssystem (T3SS) ermöglicht letzteres. Protein, die über dieses System transportiert werden (Effektorproteine) müssen von allen anderen Proteinen der bakteriellen Zelle unterschieden werden, die nicht auf diesem Weg exportiert werden. Ein spezielles Signal für diesen Exportweg wurde bisher in der Literatur am N-Terminus der Effektorproteine beschrieben, allerdings sind die genauen Eigenschaften unbekannt.

Methoden und Ergebnisse: In dieser Studie wird gezeigt, dass die Signale die in den Sequenzen der T3SS Effektoren durch Maschienenlernverfahren einheitlich erkannt werden können. Bekannte Effektoren wurden aus der Literatur und Sequenzdatenbanken zusammengestellt und dienen als Trainingsdaten für künstliche neuronale Netzwerke und Support Vektor Maschinen. Die gemeinsamen Sequenzeigenschaften waren besonders in den ersten 30 Aminosäuren der Effektorsequenzen zu finden. Die Genauigkeit der Klassifikation erreicht einen kreuzvalidierten Matthews Korrelationskoeffizienten von 0,63 und ermöglicht es, potentielle T3SS Effektoren in 705 proteobakteriellen Genomen vorherzusagen (12% durchschnittlicher Anteil). In 213 untersuchten Genomen aus der Abteilung *Firmicutes* beträgt der durchschnittliche Gehalt nur 7%.

Fazit und Signifikanz: Eine neue Methode zur Vorhersage von T3SS Effektoren wird zusammen mit einer umfassenden Begutachtung von 918 bakteriellen

Genomen vorgestellt. Die Studie zeigt, dass das analysierte Signal und die entsprechenden Eigenschaften über eine große Anzahl von Spezies verbreitet sind. Weiterhin wird eine wichtige Grundlage für die Identifikation von exportierten Proteinen pathogener Bakterien als Ziele für eine zukünftige therapeutische Intervention.

Prediction of Type III Secretion Signals in Genomes of Gram-Negative Bacteria

Martin Löwer, Gisbert Schneider*

Johann Wolfgang Goethe-University, Chair for Chem- and Bioinformatics, Frankfurt, Germany

Abstract

Background: Pathogenic bacteria infecting both animals as well as plants use various mechanisms to transport virulence factors across their cell membranes and channel these proteins into the infected host cell. The type III secretion system represents such a mechanism. Proteins transported *via* this pathway (“effector proteins”) have to be distinguished from all other proteins that are not exported from the bacterial cell. Although a special targeting signal at the N-terminal end of effector proteins has been proposed in literature its exact characteristics remain unknown.

Methodology/Principal Findings: In this study, we demonstrate that the signals encoded in the sequences of type III secretion system effectors can be consistently recognized and predicted by machine learning techniques. Known protein effectors were compiled from the literature and sequence databases, and served as training data for artificial neural networks and support vector machine classifiers. Common sequence features were most pronounced in the first 30 amino acids of the effector sequences. Classification accuracy yielded a cross-validated Matthews correlation of 0.63 and allowed for genome-wide prediction of potential type III secretion system effectors in 705 proteobacterial genomes (12% predicted candidates protein), their chromosomes (11%) and plasmids (13%), as well as 213 *Firmicute* genomes (7%).

Conclusions/Significance: We present a signal prediction method together with comprehensive survey of potential type III secretion system effectors extracted from 918 published bacterial genomes. Our study demonstrates that the analyzed signal features are common across a wide range of species, and provides a substantial basis for the identification of exported pathogenic proteins as targets for future therapeutic intervention. The prediction software is publicly accessible from our web server (www.modlab.org).

Citation: Löwer M, Schneider G (2009) Prediction of Type III Secretion Signals in Genomes of Gram-Negative Bacteria. *PLoS ONE* 4(6): e5917. doi:10.1371/journal.pone.0005917

Editor: Debbie Fox, The Research Institute for Children at Children’s Hospital New Orleans, United States of America

Received: March 20, 2009; **Accepted:** May 15, 2009; **Published:** June 15, 2009

Copyright: © 2009 Löwer, Schneider. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was supported by the Centre for Membrane Proteomics (www.cmp.uni-frankfurt.de) and the Beilstein-Institut zur Förderung der Chemischen Wissenschaften (www.beilstein-institut.de). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gisbert.schneider@modlab.de

Introduction

There are six known types of secretion systems in Gram-negative bacteria [1]. Among these, several prediction systems are available for the *sec* pathway that can be used to recognize N-terminal secretion signals (signal peptides) [2]. Predicting proteins that are secreted *via* other pathways has recently become a major goal of bioinformatics research [3]. The multi sub-unit type III secretion systems (T3SS) contribute to flagellar biosynthesis [4] and interaction with eukaryotic cells (Figure 1a) [5] and are therefore often involved in pathogenicity of the corresponding bacterial species, e.g. *Yersinia pestis*, *Salmonella enterica*, and *Escherichia coli* [6,7].

Substrate specificity of the T3SS relies on two distinct signals. Most T3SS effector proteins contain an N-terminal secretion signal, which is believed to be generic for the T3SS from different species [6]. Cellular decoding of this signal is achieved by a family of cytosolic chaperones which bind the effector sequences and are recognized by the secretion machinery [6]. Usually, there is one chaperone *per* effector protein, but chaperones targeting several effectors have also been described [6]. The genes encoding the

corresponding effector proteins and their chaperones are often organized in direct vicinity on the coding DNA sequence [8]. The function of these chaperones is not entirely clear; however, experimental results support a role as antifolding factors since fully folded effector proteins are too big for the translocation channel, and stabilizers of effector proteins, which are rapidly degraded in the absence of the corresponding chaperone [5]. Also, they are thought to provide a secondary secretion signal which is somehow involved in the prioritization and order of effector secretion [5].

Analyses of known effector sequences have revealed characteristic properties, such as an overall amphipathic amino acid composition, an over-representation of serine and glutamine, and the absence of acidic residues [9]. The actual secretion signal is believed to be contained in the first 50 amino acids, although synthetic signals with as few as eight residues have been shown to promote type III secretion in *Yersinia* [10]. Furthermore, some evidence has been collected that the signal might be encoded on RNA level rather than on protein level [11]. Figure 1b presents the typical structure of a classic signal peptide [12] compared to T3SS signals.

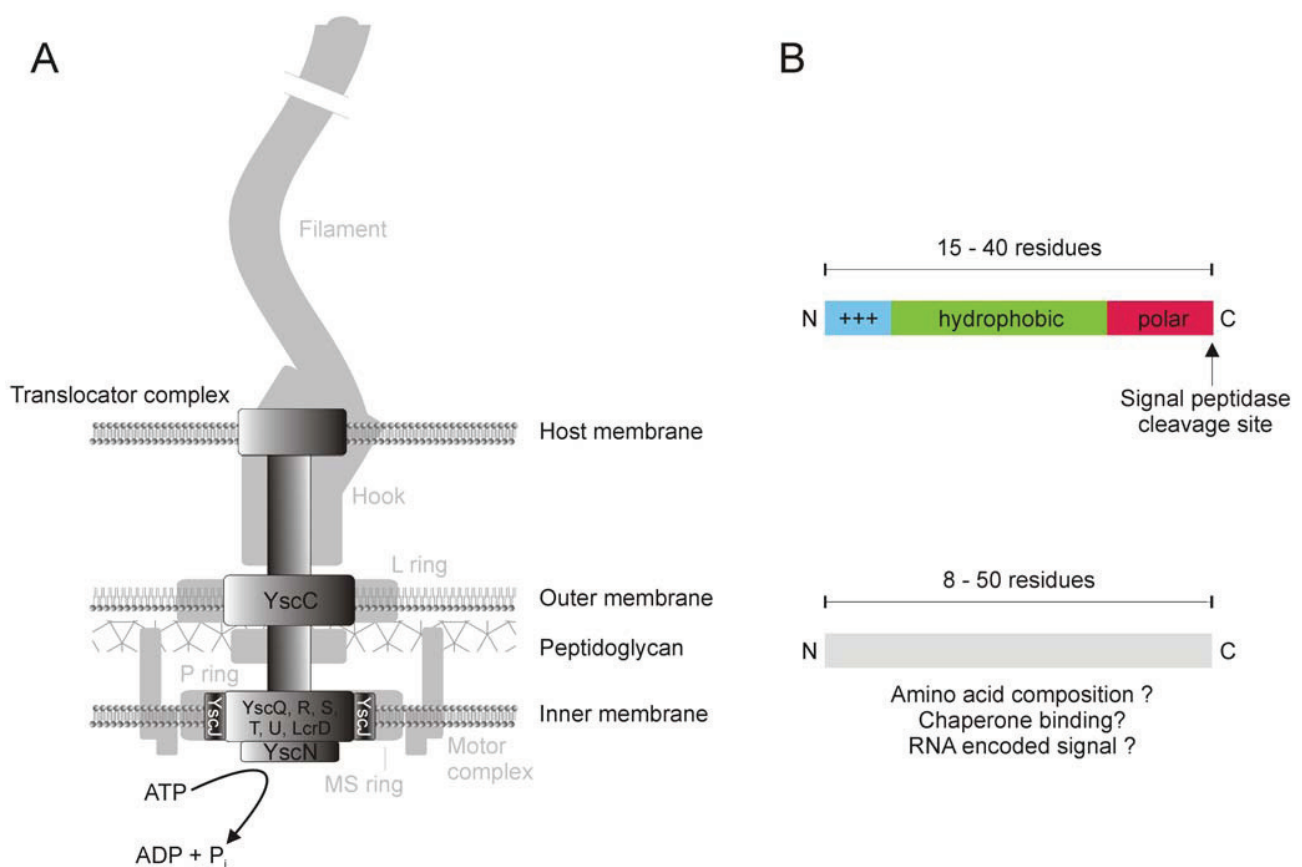


Figure 1. The bacterial type III secretion system (T3SS) forms a translocator complex spanning the bacterial and the host cell membranes for protein translocation. (a) Schematic T3SS structure together with a flagella apparatus (shaded in light grey). The nine components being conserved among T3SS are named in *Yersinia* nomenclature. In flagella apparatus, proteins of the axial structure are exported via a T3SS, e.g. flagellins. Note that T3SS injection needle and translocator complex are not present in flagella (adapted from Sheng *et al.* [5] and Pallen and Matzke [4]). (b) Comparison of the features of classic signal peptides (top) [12] and the proposed features of T3SS signals (bottom). Both kinds of signals are located at the N-terminal end of a protein.
doi:10.1371/journal.pone.0005917.g001

Recent sequence-based bioinformatics approaches to finding new effector proteins utilize consensus sequence patterns of the N-terminal secretion signals [9], similarity-based comparison to known effectors [13], the genomic organization of the effectors by identifying genes in vicinity to chaperone homologues [14], and amino acid composition rules [15]. Here we present a new machine learning approach to identify potential T3SS effectors by their N-terminal amino acid sequence using a sliding window procedure in combination with artificial neural networks (ANN, feedforward type) [16] and support vector machine (SVM) classifiers [17], together with a comprehensive prediction of potential T3SS effectors for 918 bacterial genomes.

Materials and Methods

Data preparation

We collected a raw data set containing a total of 1,860 protein sequences (979 positive, 881 negative samples) from various literature and database sources. Included were sequences from the SwissProt [18] and *Pseudomonas syringae* Hop [19] databases and from a dataset published by Tobe and coworkers [13]. The negative data consisted of 881 cytoplasmic sequences and secreted proteins from Gram-negative organisms. The publicly available SignalP [20] and SeretomeP [21] training sets were

included. Each of the sequences of the secreted proteins contains an N-terminal secretion signal for the *sec* pathway. Possible redundancy of both datasets was reduced by using the PISCES implementation of the Hobohm algorithm [22]. Sequences with fewer than 100 amino acids were removed. The maximum pairwise identity of the sequences was 90% after the reduction, resulting in a final set of 575 positive and 685 negative sequence examples. The complete data set is available in FASTA-format [23] as Supplementary Material.

Then, sequences were analyzed using the sliding-window technique. The sliding window procedure divides a sequence in a number of overlapping subsequences. Starting from the N-terminal residue position, as many residues were read as determined by the window size, then the window was moved one residue position towards the C-terminus. The procedure was repeated until the C-terminus is reached. For each subsequence a score value (probability) was calculated by a machine learning classifier. For classifier training, the sequences were prepared by removal of the N-terminal amino acid (a methionine in most cases) and keeping only the N-terminal portions of length L . For each sequence stretch of length L , the appropriate number of windows with a width W was computed. Each amino acid residue of a single window was encoded into a unitary bit string of length 20, where a bit was set (value = 1) if its position in the string corresponds to the

position of the amino acid residue and zero otherwise [24]. As a result, a sequence window of length W was encoded by a bit string containing $W \times 20$ bits with exactly W bits set to 1 and all other bits zero.

The input data for the machine learning algorithms consisted of $(L-1)-W$ such bit vectors. Additionally, $575 \times (L-1)-W$ encoded sequence windows were randomly sampled from the C-terminal portions (starting at sequence position 51) of the positive sequence set and included as *pseudo-negative* training samples. The values of the length cut-off L and the window size W were systematically varied between 10 or 7 and 50 or 49, respectively.

Machine learning classifiers

We used MATLAB version R2007a [25] and SVMlight version 6.02 [26] software for training of the classifier models. The ANNs had feed-forward architecture with a single hidden neuron layer (Figure 2). All neurons in the hidden layer and the single output neuron had sigmoidal activation [16]. We used gradient descent backpropagation learning with momentum and an adaptive learning rate, as described previously [16]. Early termination of the training process was implemented by splitting the training data into two smaller training and validation sets, and stopping the training when the calculated error for the validation data rose for a predefined number of training cycles. For each set of training data, the number of hidden neurons was systematically varied from one to ten. For binary (yes/no) classification, the output of the ANN was converted to binary value using a threshold value of $\theta = 0.5$. The overall function modelled by the implemented ANN is given by Eq. (1).

$$f(\mathbf{x}) = \text{logsig} \left(\sum_h v_h \text{logsig} \left(\sum_d w_d \xi_d + \vartheta_h \right) + \Theta \right), \quad (1)$$

where *logsig* is a sigmoidal transfer function (activation function) limiting the neuron output to the interval $[0,1]$, v and w are the connection weights, ϑ the hidden neurons' bias values, and Θ the bias of the output neuron.

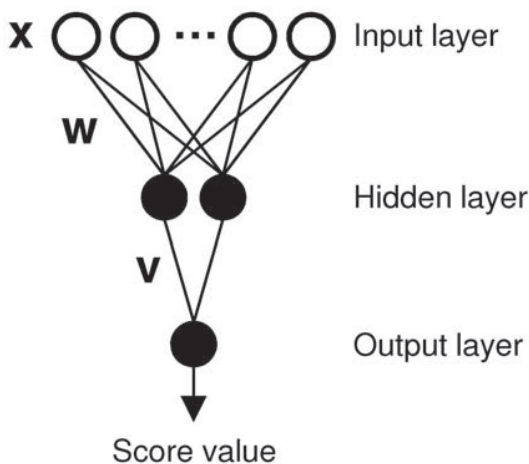


Figure 2. Three-layered feedforward neural networks were trained on the prediction of T3SS effector proteins. In this schematic, artificial neurons are drawn as circles (white: fan-out neuron; black: sigmoidal activation). For clarity, not all neurons are shown. The output neuron computes values between 0 and 1, which can be interpreted as the probability of an input sequence window being part of a T3SS effector signal.

doi:10.1371/journal.pone.0005917.g002

The SVMs used soft margins and a radial basis function (RBF) kernel (Eq. 2). A grid search in logarithmic space was performed to find optimal values for the complexity parameter C and RBF parameter γ , as described [17]. The prediction of a trained SVM classifier used in this study is given by Eq. 2.

$$f(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}, \mathbf{x}_i^{SV}) + b, \quad (2)$$

where $K(\mathbf{x}, \mathbf{x}_i^{SV}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}_i^{SV}\|^2\right)$.

The greater f the higher is the probability for a compound to belong to the positive class (here: T3SS signals), \mathbf{x} and \mathbf{y} are sequence descriptor vectors, \mathbf{x}^{sv} are support vectors, *i.e.* data vectors that define the exact shape of the separating SVM hyperplane. The kernel function K defines the complexity of the surface that will be constructed. Here, we used the RBF kernel. No optimization of the choice of K was performed.

Training performance of both the ANNs and SVMs was evaluated by ten-fold cross-validation (leave 50% out) and calculation of the average Matthews correlation coefficient (Eq. 3) [27]

$$mcc = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}, \quad (3)$$

where TP , TN , FN and FP denote the true-positive, true-negative, false-negative and false-positive prediction counts, respectively.

During the training process, each sequence window was considered as an individual training example and given a score, *i.e.* the ANN or SVM output. For application of the classifiers to protein sequences (obtained from bacterial genome data), an average score was computed from the individual window scores.

To compare our results to other approaches, two previously applied sets of classification rules [10,15] were re-implemented in the programming language Python [28].

The final SVM and ANN prediction models are publicly available *via* our web server (<http://www.modlab.de>).

Results and Discussion

Our study consisted of two subsequent steps: i) training of machine learning classifiers on the prediction of T3SS effectors, and ii) application of the trained classifiers on known or hypothetical proteins from available bacterial genomes, chromosomes, and plasmids.

Machine learning and prediction performance

The starting point for both classification methods is a vector representation of the training data. Thus each training example represents a point in a vector space. During the training process, both the ANN and SVM approximate a function (hyperplane) in this vector space, which is intended to separate the positive and the negative training examples. This function can be used to classify new data points in the vector space. The multilayer perceptron used in this study employed multiple layers of artificial neurons (Figure 2) to non-linearly map the input vector to a binary classifier value. The parameters defining this mapping (weights and threshold values) are learned during the classifier training by minimizing an error function. In contrast to such ANNs, support vector machines use a so-called “kernel function” to map the training examples into a higher dimensional feature space where

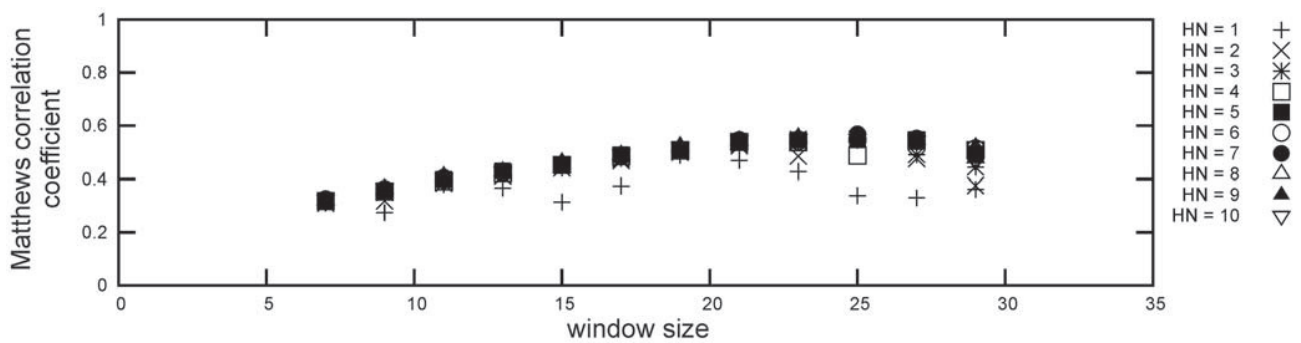


Figure 3. T3SS effector proteins contain a targeting signal in their N-terminal sequence portion. Performance results of the first round of neural network cross-validation for sequence length 30 and varying numbers of hidden neurons (HN) in the neural network classifiers and window sizes are shown. Values are averaged over the cross-validation folds. The data for lengths 10, 20, 40 and 50 can be found in Supplementary Figure S1. doi:10.1371/journal.pone.0005917.g003

the examples can be separated by a hyperplane. The task of finding such a plane for a given kernel function with the constraint of maximizing the distance of the plane to the training data can be formulated as a convex optimization problem and computed efficiently [29,30,31].

For machine learning, it was important to realize that other transport mechanisms than T3SS also rely on N-terminal sequence signals, e.g. the *Sec* dependent pathway. Our dataset reflects the need to differentiate between T3SS signals and other signals, as all transportation pathways may coexist in a single species. Included are sequences with *Sec* signals, cytoplasmic proteins, and proteins exported by unknown pathways. In addition, the C-terminal sequence portions of the collected T3SS effectors were included in the negative training set. This excludes a possible general sequence bias which might be shared among the species providing the positive training data.

In order to reduce the theoretical number of 6,242,600 possible parameter sets, which results as the product of the number of sequence lengths L , possible window sizes W per sequence length, number of hidden neurons in the ANNs, and cross-validation shuffles, several attempts were made to reduce the parameter space: First, a minimal window size of $W=7$ residues with an increment by two was used. Second, we employed a straightforward optimization protocol for the sequence length cut-off, starting with a first round of calculations using the lengths $L=10, 20, 30, 40$ and 50 only. In the following rounds the cut-off value interval around the best performing value of the previous round was investigated in more detail. We wish to point out that due to this optimization protocol, only a single performance

maximum (a “practical optimum”) can be found and it bears the risk of missing the absolute optimum.

Maximal average cross-validation performance was achieved for $L=30$ (Figure 3), $W=25$ and seven hidden neurons in the ANN ($mcc=0.57\pm 0.04$), although all results with more than four hidden neurons are comparable. Two more training rounds were executed (Supplementary Figures S2 and S3), using $L=25$ and $L=35$ for the second, and $L=31$ to 34 for the third pass. Neither of these calculations yielded a higher performance than the maximum for $L=30$, so the respective parameter values were employed by the final model, which was obtained by 100 training runs with randomly shuffled training data and early stop validation but no cross-validation. The performance of the best model on the complete training data is presented in Table 1. The higher accuracy likely results for three reasons: i) more data was included in the training, ii) randomized training allows for finding other performance optima, and iii) the scoring of individual sequence windows was changed to the average score over all windows.

We also studied the influence of the most N-terminal part of the training examples on the performance of ANN training. However, cleaving N-terminal parts of varying size off the training sequences reduced the performance (cf. Supplementary Figure S4). This suggests that the N-terminal part of the training sequences holds important information for distinguishing T3SS effectors.

The ANN model bears an adjustable parameter, the threshold θ , which is the decision boundary for classification of the network output. It was set to 0.5 during training, but the influence of this parameter on the performance of the final model can be studied by a Receiver Operating Characteristic (ROC) threshold test [32].

Table 1. Performance of the prediction systems and sequence patterns on the complete training data (re-classification).

model	prediction for positive data (T3SS effectors)		prediction for negative data (non-effectors)		mcc
	Positive (TP)	Negative (FN)	Positive (FP)	Negative (TN)	
ANN	423 (0.74)	152 (0.26)	12 (0.02)	673 (0.98)	0.75
SVM	569 (0.99)	6 (0.01)	0 (0.0)	685 (1.0)	0.99
P1	468 (0.81)	107 (0.19)	476 (0.69)	209 (0.31)	0.14
P2	200 (0.34)	375 (0.66)	107 (0.15)	578 (0.85)	0.22

Given are absolute values and relative values in brackets. *TP*, *TN*, *FN* and *FP* denote the true-positive, true-negative, false-negative and false-positive prediction ratios, respectively. P1 and P2 indicate rule sets for prediction of type III secretion system effectors (T3SS) published by Petnicki-Ocwieja *et al.* and Vencato *et al.* [7,12]. ANN: artificial neural network; SVM: support-vector machine. doi:10.1371/journal.pone.0005917.t001

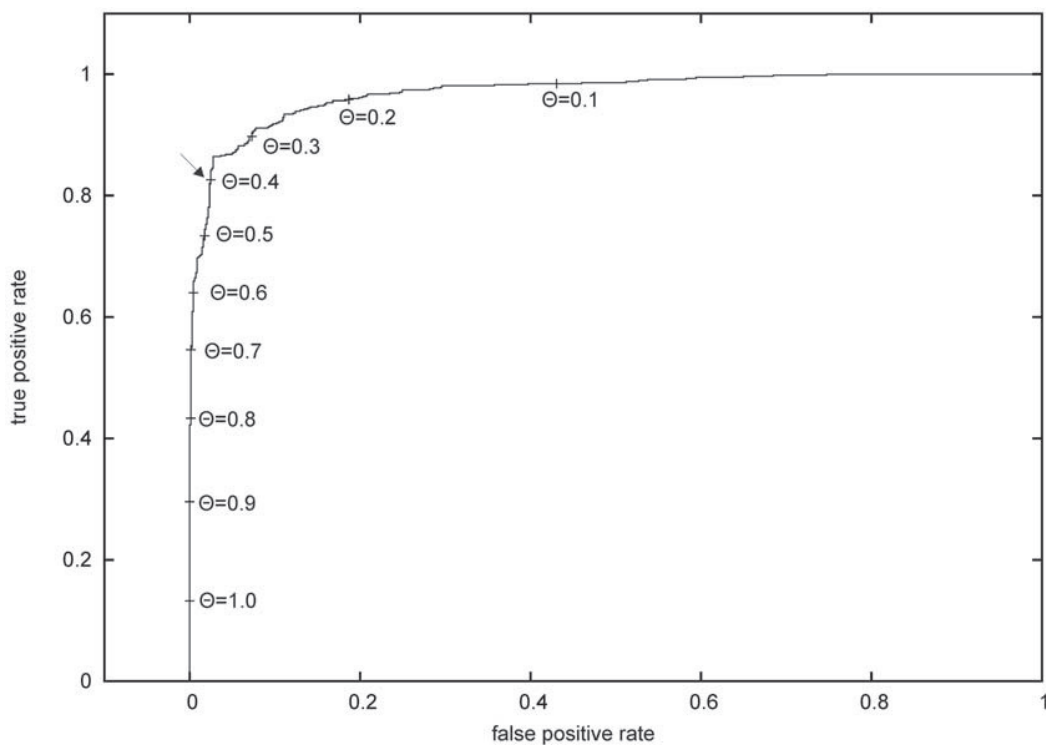


Figure 4. The best neural network classifier was determined by receiver-operator characteristic (ROC) analysis. The plot results from a threshold test with the final neural network model. Threshold values Θ for the predicted score ranged from 0.1 to 1.0. The threshold value of the final model ($\Theta = 0.4$) is marked by an arrow.
doi:10.1371/journal.pone.0005917.g004

The ROC curve is shown in Figure 4. The sudden flattening of the curve at a true positive ratio of about 0.85 suggests a selection of θ between 0.4 and 0.3 to optimize the true positive/false positive ratio tradeoff. For genome/predicted proteome analysis, we used the final ANN model with $\theta = 0.4$.

Employing this parameter value for re-classification of the training data yielded an increased Matthews correlation of $mcc = 0.82$. The final classifier has a *sensitivity* of 83%, a *specificity* of 97%, and an *accuracy* of 91% [33]. As a control, we also trained neural networks on a sequence set randomly picked from the SwissProt database [18] and of the same size as our training data. A second control was done by training neural networks on the collected training data randomly divided into positive and negative examples (*Y-scrambling* test). In both experiments, no correlation between the actual and predicted class labels was observed ($mcc = 0.0 \pm 0.0$, and $mcc = 0.003 \pm 0.018$, respectively).

In addition to the neural network classifier, we trained a preliminary SVM with $L = 30$ and $W = 25$ input data. The best performing model had a complexity value of $C = 1000$ and a kernel gamma of $\gamma = 0.01$. Average cross-validation performance yielded $mcc = 0.63 \pm 0.02$. Results for the complete training data are given in Table 1. In both cases, the SVM apparently outperformed the ANN model. However, concerning its “true” predictive capabilities, it might be more appropriate to compare the SVM cross-validation performance to the ANN final model performance, as in both cases the training algorithm used only 90% of the available data (10% were employed for determination of the forced stop time point during training). In addition, the great number of support vectors (5,144 support vectors among 7,340 training vectors) in combination with the comparably large gamma value, suggest a limited generalization ability of this

particular SVM model [34]. This is why we used only the ANN classifier for productive genome analysis in this study, while the SVM model served as secondary classifier.

We wish to stress that it is unlikely that the ANN will outperform an SVM solution once a good kernel will have been identified [35]. This technical optimization of the SVM kernel function was not part of our study, and is currently under investigation by us. Profile Hidden Markov Models (HMM) might also represent a method of choice for the given prediction task [36]. The present analysis was intended to provide a first cross-genomic prediction of potential T3SS effectors and certainly leaves room for future improvement. This will also have to address the interpretation of the decisive feature vector used by the machine learning classifier.

Compared to recently published residue motif rules (Table 1, rows P1 and P2) [7,12] – whose performance was optimized by allowing for some rule violations – the performance of the ANN and SVM models is clearly superior. It should be kept in mind, however, that these rule sets were derived from a far smaller dataset and not intended for predictive purposes.

Genome analysis and protein prediction

We applied the ANN classifier to two groups of genomes collected from the RefSeq database [37]. The groups include the phylum *Proteobacteria* as Gram-negative examples and the phylum *Firmicutes* as Gram-positive examples. BLAST (BLOSUM62 substitution matrix [38], e -value $< 10^{-5}$) [39] was used to divide the genomes in groups depending on their possession of a homologue of the *YscN* gene from *Yersinia pseudotuberculosis* (UniProt ID YSCN_YERPS), which is known to be an integral part of a functional T3SS in *Yersinia* [40]. Notably, for all examined genomes at least one significant alignment was found, which is

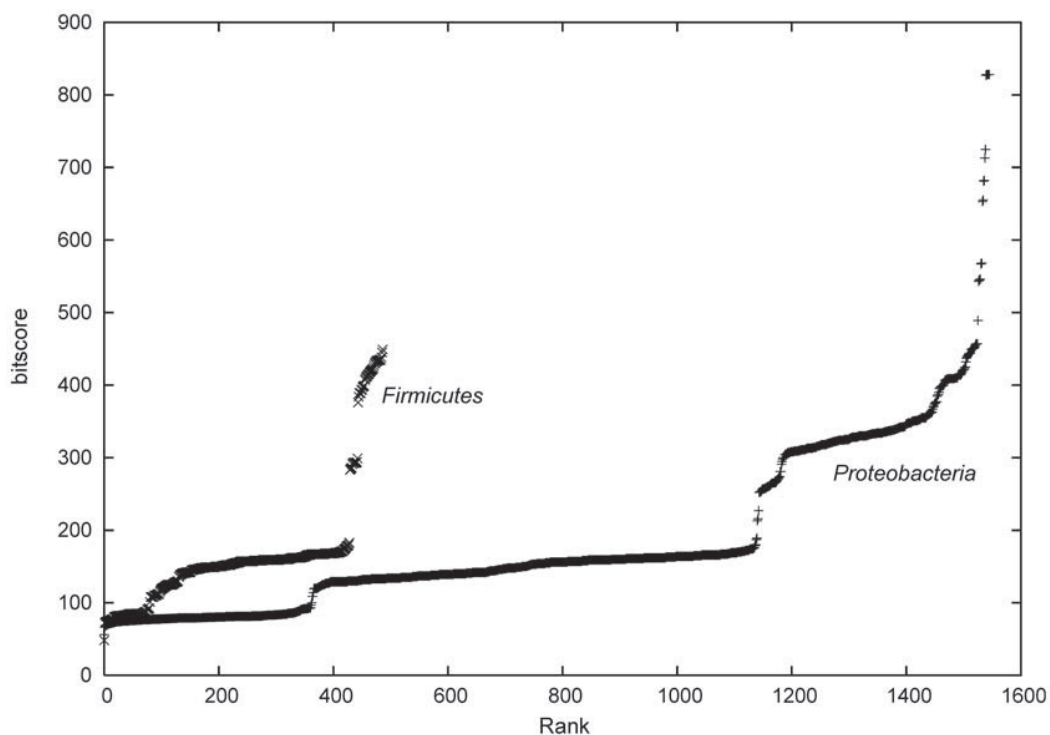


Figure 5. Ranking of the most significant protein alignments from all genomes was done according to their BLAST bitscore (BLOSUM62 substitution matrix, e -value $< 10^{-5}$). The query protein was of the *YscN* gene from *Yersinia pseudotuberculosis* (UniProt ID YSCN_YERPS).

doi:10.1371/journal.pone.0005917.g005

not expected for the Gram-positive genomes. As *YscN* is believed to be an ATPase, other enzymes with the same activity might be the reason for this finding. Consequently the BLAST bitscore threshold was set to 200 bit, as a plot of the scores suggests an inflection point around this value (Figure 5). Furthermore, sequence data from proteobacterial plasmids were separately evaluated, as only 17 plasmids seem to code for an *YscN* like protein, and these plasmids often encode virulence determinants including T3SSs, e.g. the *Shigella* plasmids [41].

Table 2 presents the main results of this screening exercise. All of the examined proteobacterial genomes have a comparable percentage of positive predictions (approx. 11%), which seems to be unbiased by the presence of a potential *YscN* protein, as the averages are comparable when the genomes are divided accordingly (not shown). Noticeable is a high standard deviation for the plasmid data, which might be caused by the pronounced length variation of the examined plasmids. The complete list of results shows that with regard to the relative amount of positive predictions, plasmid

sequences occupy the highest ranks (cf. Supplementary Table S1). Many belong to genera including animal pathogenic species such as *Shigella*, *Yersinia*, *Escherichia*. Several plant pathogens are listed, e.g. *Pseudomonas syringae*, *Xanthomonas campestris*. All of the 17 plasmids holding an *YscN* homologue are present among the first 19% of the list entries. This observation clearly supports the robustness of our predictions and justifies the selection of the particular bitscore threshold applied in this study.

The *Firmicutes* yield a lower overall content of *YscN* homologues relative to *Proteobacteria*. This is expected as only flagella but no standalone T3SSs exist in this phylum [42]. The average positive predictions suggest again that the T3SS signal appears to be widely spread. On the other hand, the ordering of the genomes by positive prediction content is insightful. For example, different *Clostridium* species yield a high content of positive predictions, and are also known to have flagella (cf. Supplementary Table S2).

The plasmid of *Yersinia* species is known to code for several virulence determinants including a T3SS and at least twelve

Table 2. Prediction results for the genomes (*in silico* translated sequences) of *Proteobacteria* and *Firmicutes*.

group	Number of genomes	Number of <i>YscN</i> containing genomes	Average % positive prediction (standard deviation in brackets)
<i>Proteobacteria</i>	705	284	11.5 ($\sigma = 7.5$)
proteobacterial chromosomes	405	267	10.5 ($\sigma = 2.7$)
proteobacterial plasmids	300	17	12.9 ($\sigma = 10.8$)
<i>Firmicutes</i>	213	58	6.9 ($\sigma = 5.6$)

doi:10.1371/journal.pone.0005917.t002

effector proteins named “*Yersinia* outer membrane proteins” (Yops) [43]. Note that the proteins encoded on this plasmid were not included in the training data. Out of the 72 proteins encoded on the plasmid of *Yersinia enterocolitica* subsp. *enterocolitica* 8081 [44], 16 are predicted to have a T3SS targeting signal (*cf.* Supplementary Table S3). Ten of these proteins are Yops and thus correctly identified. The two missing Yops are *yopQ* and *lerV*, which received a neural network score of 0.22 and 0.3, respectively. Among the remaining six positive predictions are the chaperone of *yscN* and the *repA* and *spyB* proteins, which are involved in plasmid replication and partition [43]. These proteins are clearly false-positive predictions. Also, there are *yscP* and *yscM*, which are known to be secreted [38]. The last predicted T3SS effector is *yscW*, which is a chaperone of the T3SS component *yscC* and enables the outer membrane localization of *yscN* [45]. As *yscN* has no predicted T3SS targeting signal and *yscW* is described to be the “pilot protein” for *yscN* [45], the predicted signal of *yscW* might be responsible for the transport of both proteins.

We then took a closer look at one of the examined species, *Helicobacter pylori* 26695 (RefSeq ID NC_000915), which uses flagella to propel itself and therefore has a functioning T3SS [4]. As expected, an *YscN* homologue is found, but the content of positive predictions is relatively low (6.5%). Only 93 sequences are predicted to actually contain a T3SS signal. Twelve of them are annotated as being associated to the flagellar complex, and 38 sequences are marked as “hypothetical” or lack a functional annotation (*cf.* Supplementary Table S4).

We also applied the SVM model to these *Helicobacter* data, yielding 77 candidate proteins of which 37 are annotated as “hypothetical” (not shown). 18 of these hypothetical protein sequences are shared with the ANN predictions (Table 3). BLAST [34] was used to compare these sequences with the non-redundant (nr) database of the NCBI [46]. For most of the sequences it is not possible to infer a putative function. As an exception, the sequence

Hp0906 is distantly related to a putative flagellar hook protein of *Campylobacter jejuni* (alignment length = 113 residues, 36% identities).

While the flagellum associated positive predictions can be regarded as biologically plausible and the hypothetical proteins might be effectors of a T3SS, some of the predicted signal-containing proteins are metabolic enzymes, *i.e.* the citrate synthase or biotin synthetase, which are not expected to be exported. Chromosomes of the other two strains of *Helicobacter pylori*, for which fully sequenced genomes are available (HPAG1 [47] and J99 [48]), obtain a similar predicted percentage of T3SS effectors, which also holds for the related species *Helicobacter acinonychis*, being a gastric pathogen of large felines [49]. For each of the three *Helicobacter pylori* strains ten putative flagellar components are predicted to possess a T3SS signal and share the same functional annotation. Also the obvious false-positive predictions (citrate synthase and biotin synthetase) occur for all strains.

Conclusions

In this study we present evidence for the existence of common sequence features in the N-terminal portion (30 residues) of T3SS effectors. The existence or absence of these features can be predicted with reasonable accuracy. A low number of false positive predictions of our classifiers is an important feature, as it might help preventing unnecessary experiments when applied to selecting candidates for an experimental survey. Moreover, the predicted features seem to be universally distributed among sequences of a wide range of both Gram-negative and Gram-positive bacteria, regardless of the existence of a T3SS. Thus, we cannot be completely sure that the machine learning classifiers actually extracted directly T3SS-related and secretion-inducing features. Additional and different types of machine learning classifiers will have to be developed to address this point. In particular, we expect that thorough SVM classifier training will provide improved predictions and help understand the actually

Table 3. Predicted proteins from *Helicobacter pylori* strain 26695 that might be exported via a Type 3 Secretion System and were predicted by both ANN and SVM classifiers.

No.	Database accession codes/loci (Genbank, NCBI)	Annotation	<i>H. pylori</i> gene identifier
1	gi 15644743 ref NP_206913.1	Hypothetical protein	HP0113
2	gi 15644939 ref NP_207109.1	Hypothetical protein	HP0311
3	gi 15644995 ref NP_207165.1	Hypothetical protein	HP0367
4	gi 15645055 ref NP_207225.1	Hypothetical protein	HP0427
5	gi 15645292 ref NP_207462.1	Hypothetical protein	HP0668
6	gi 15645302 ref NP_207472.1	Hypothetical protein	HP0678
7	gi 15645498 ref NP_207673.1	Hypothetical protein	HP0879
8	gi 15645522 ref NP_207698.1	Hypothetical protein	HP0906
9	gi 15645579 ref NP_207755.1	Hypothetical protein	HP0963
10	gi 15645605 ref NP_207781.1	Hypothetical protein	HP0990
11	gi 15645679 ref NP_207856.1	Hypothetical protein	HP1065
12	gi 15645756 ref NP_207933.1	Hypothetical protein	HP1142
13	gi 15645847 ref NP_208025.1	Hypothetical protein	HP1233
14	gi 15646001 ref NP_208182.1	Hypothetical protein	HP1391
15	gi 15646018 ref NP_208199.1	Hypothetical protein	HP1408
16	gi 15646039 ref NP_208221.1	Hypothetical ATP-binding protein	HP1430
17	gi 15646108 ref NP_208290.1	Hypothetical protein	HP1499
18	gi 15646129 ref NP_208311.1	Hypothetical protein	HP1520

doi:10.1371/journal.pone.0005917.t003

relevant sequence features. During the reviewing process of this paper two other articles [50,51] were published which address the same problem of the prediction of T3SS effectors using a similar methodology. Interestingly, both studies lead to similar conclusions regarding the length of the putative signal on the primary protein structure and the spread of the signal among different species. Arnold *et al.* developed a naïve Bayes classifier by which up to 12% potential T3SS effectors were predicted for whole genomes [50], which is in perfect agreement with our results. These authors also demonstrate that in some cases *in silico* frame-shift mutations do not affect the predictions which might be an explanation for the hypothetical RNA encoded signal [11]. We wish to point out that our prediction system has the highest *specificity* among the presented approaches, which is an important property for prioritizing biochemical and cell biological experiments. This might be a result of the larger training data set and especially the composition of the negative training data used in our study.

Most interestingly, according to our analysis flagella T3SS and standalone T3SS seem to share the same kind of signal. Viewed from an evolutionary perspective, one might speculate that the signal evolved independently from the T3SS, maybe even without having any particular targeting function, and eventually the signal pattern was adopted by the developing T3SS for effector tagging. On the other hand, we stress that the predictions contain apparent errors, as we predict obvious cytoplasmic proteins to have a T3SS export signal. This observation leaves room for further improvements, for example by modifying the training data composition. In this context one has to keep in mind that there are certain chaperones that promote type III secretion [4], but it has not yet been determined whether both signal components (the actual sequence feature and the chaperone) are required for protein translocation or if one alone might be sufficient under certain conditions.

Supporting Information

Figure S1 The plots present the performance results for the first round of ANN cross-validation for sequence lengths 10 (A), 20 (B), 30 (C), 40 (D) and 50 (E) and varying numbers of hidden neurons and window sizes. The data values are averaged over the cross validation folds, standard deviation is not shown for clarity.
Found at: doi:10.1371/journal.pone.0005917.s001 (0.05 MB PDF)

Figure S2 The graphs present performance results for the first round of ANN cross-validation for sequence lengths 25 (A) and 35 (B) and varying numbers of hidden neurons and window sizes. The data values are averaged over the cross-validation folds, standard deviations are not shown for clarity.
Found at: doi:10.1371/journal.pone.0005917.s002 (0.03 MB PDF)

Figure S3 The plot presents the performance results for the first round of ANN cross-validation for sequence lengths 31 (A), 32 (B), 33 (C) and 34 (D) and varying numbers of hidden neurons and window sizes. The data values are averaged over the cross-validation folds, standard deviations are not shown for clarity.

References

- Bingle LEH, Bailey CM, Pallen MJ (2008) Type VI secretion: a beginner's guide. *Curr Opin Microbiol* 11: 3–8.
- Schneider G, Fechner U (2004) Advances in the prediction of protein targeting signals. *Proteomics* 4: 1571–1580.
- Casadio R, Martelli PL, Pierleoni A (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Briefings in Functional Genomics and Proteomics* 7: 63–73.
- Pallen MJ, Matzke NJ (2006) From The Origin of Species to the origin of bacterial flagella. *Nat Rev Microbiol* 4: 784–790.
- Sheng YH, Nomura K, Whitam TS (2004) Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim Biophys Acta* 1694: 181–206.
- Galán JE, Wolf-Watz H (2006) Protein delivery into eukaryotic cells by type III secretion machines. *Nature* 444: 567–573.

Found at: doi:10.1371/journal.pone.0005917.s003 (0.04 MB PDF)

Figure S4 The length of the N-terminal sequence portion used for classifier training has an influence on neural network performance. Results are presented for three different lengths L . The x -axis is scaled to the fraction of removed sequence (cutoff values divided by the overall length). The performance values presented are averaged over the number of hidden neurons, the number of cross-validation shuffles, and different window sizes. Error bars denote the standard deviation. For length $L = 30$ the most N-terminal 10 and 20 residues were removed and for $L = 40$ and $L = 50$ the most N-terminal 10, 20 and 30 residues were removed. For better visualisation, this is expressed as fraction in the plot. In all cases a decrease in performance can be observed when compared to Figure S1.

Found at: doi:10.1371/journal.pone.0005917.s004 (0.02 MB PDF)

Table S1 Complete list of examined protein sequence sets of *Proteobacteria*. Given is the genome name, the NCBI Refseq database identification string, the existence of an *YscN* homologue, the number of positive predictions (P), the number of negative predictions (N) and the relative number of positively predicted protein sequences (%). The list is sorted according to decreasing fractions of predicted proteins.

Found at: doi:10.1371/journal.pone.0005917.s005 (0.59 MB DOC)

Table S2 Complete list of examined protein sequence sets of *Firmicutes*. Given is the genome name, the NCBI Refseq database identification string, the existence of an *YscN* homologue, the number of positive predictions (P), the number of negative predictions (N) and the relative number of positively predicted protein sequences (%). The list is sorted according to decreasing fractions of predicted proteins.

Found at: doi:10.1371/journal.pone.0005917.s006 (0.20 MB DOC)

Table S3 Predicted proteins from the *Yersinia enterocolitica* strain 8081 virulence plasmid that might be exported via a Type 3 Secretion System. Higher score values indicate more reliable predictions.

Found at: doi:10.1371/journal.pone.0005917.s007 (0.04 MB DOC)

Table S4 Predicted proteins from *Helicobacter pylori* strain 26695 that might be exported via a Type 3 Secretion System. Higher score values indicate more reliable predictions.

Found at: doi:10.1371/journal.pone.0005917.s008 (0.08 MB DOC)

Acknowledgments

We are grateful to Dr. Silja Wessler and Dr. Jan A. Hiss for helpful discussion.

Author Contributions

Conceived and designed the experiments: ML GS. Performed the experiments: ML. Analyzed the data: ML GS. Wrote the paper: ML GS.

7. Pallen MJ, Beatson SA, Bailey CM (2005) Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS Microbiol Rev* 29: 201–229.
8. Page AL, Parsot C (2002) Chaperones of the type III secretion pathway: jacks of all trades. *Mol Microbiol* 46: 1–11.
9. Petnicki-Ocwieja T, Schneider DJ, Tam VC, Chancey ST, Shan L, et al. (2002) Genomewide identification of proteins secreted by the Hrp type III protein secretion system of *Pseudomonas syringae* pv. *tomato* DC3000. *Proc Natl Acad Sci USA* 99: 7652–7657.
10. Lloyd SA, Sjöström M, Andersson S, Wolf-Watz H (2002) Molecular characterization of type III secretion signals via analysis of synthetic N-terminal amino acid sequences. *Mol Microbiol* 43: 51–59.
11. Sorg JA, Miller NC, Schneewind O (2005) Substrate recognition of type III secretion machines – testing the RNA signal hypothesis. *Cell Microbiol* 7: 1217–1225.
12. von Heijne G (1985) Signal sequences. The limits of variation. *J Mol Biol* 184: 99–105.
13. Tobe T, Beatson SA, Taniguchi H, Abe H, Bailey CM, et al. (2006) An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci USA* 103: 14941–14946.
14. Panina EM, Mattoo S, Griffith N, Kozak NA, Ming H, Miller JF (2005) A genome-wide screen identifies a *Bordetella* type III secretion effector and candidate effectors in other species. *Mol Microbiol* 58: 267–279.
15. Vencato M, Tian F, Alfano JR, Buell CR, Cartinhour S, et al. (2006) Bioinformatics-Enabled Identification of the HrpL Regulon and Type III Secretion System Effector Proteins of *Pseudomonas syringae* pv. *phaseolicola* 1448A. *Molecular Plant-Microbe Interactions* 19: 1193–1206.
16. Russell S, Norvig P (2003) *Artificial Intelligence – A Modern Approach* (second edition). New Jersey: Pearson Education Inc.
17. Vapnik V (1995) *The nature of statistical learning theory*. New York: Springer.
18. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33: 154–159.
19. Lindeberg M, Stavrinides J, Chang JH, Alfano JR, Collmer A, et al. (2005) Unified nomenclature and phylogenetic analysis of extracellular proteins delivered by the type III secretion system of the plant pathogenic bacterium *Pseudomonas syringae*. *Molecular Plant-Microbe Interactions* 18: 275–282.
20. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783–795.
21. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S (2005) Non-classical protein secretion in bacteria. *BMC Microbiology* 5: 58.
22. Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19: 1589–1591.
23. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science* 227: 1435–1441.
24. Qian N, Sejnowski TJ (1998) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 202: 865–884.
25. MATLAB, Version 2007b, The MathWorks, Natick, Massachusetts (USA).
26. Joachims T (1999) Making large-Scale SVM Learning Practical. In: Schölkopf B, Burges C, Smola A, eds. *Advances in Kernel Methods - Support Vector Learning*. Cambridge: MIT Press.
27. Mathews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405: 442–451.
28. van Rossum G (1995) *Python Reference Manual*. Amsterdam: CWI Report CS-R9525.
29. Duda RO, Hart PE, Stork DG (2001) *Pattern Classification*. New York: John Wiley & Sons.
30. Schölkopf B, Smola AJ (2002) *Learning with Kernels*. Cambridge: MIT Press.
31. Byvatov E, Schneider G (2003) Support vector machine applications in bioinformatics *Appl Bioinformatics* 2: 67–77.
32. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recongn Lett* 27: 861–874.
33. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783–795.
34. Chin KK (1998) *Support Vector Machines applied to Speech Pattern Classification*. Dissertation. Available: http://svr-www.eng.cam.ac.uk/~kcc21/thesis_main/thesis_main.html. Accessed March 2, 2009.
35. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, et al. (2008) Machine learning methods for predictive proteomics. *Brief Bioinformatics* 9: 119–128.
36. Eddy S (1998) Profile Hidden Markov Models. *Bioinformatics* 14: 755–762.
37. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33: 501–504.
38. Henikoff S, Henikoff G (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89: 10915–10919.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
40. Hueck CJ (1998) Type III Protein Secretion Systems in Bacterial Pathogens of Animals and Plants. *Microbiol Mol Biol Rev* 62: 379–433.
41. Yang F, Yang J, Zhang X, Chen L, Jiang Y, et al. (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucl Acids Res* 33: 6445–58.
42. Jarrell KF, McBride MJ (2008) The surprisingly diverse ways that prokaryotes move. *Nat Rev Microbiol* 6: 466–476.
43. Cornelis GR, Boland A, Boyd AP, Geuijen C, Iriarte M, et al. (1998) The Virulence Plasmid of *Yersinia*, an Antihost Genome. *Microbiol Mol Biol Rev* 62: 1315–1352.
44. Thomson NR, Howard S, Wren BW, Holden MTG, Crossman L, et al. (2006) The Complete Genome Sequence and Comparative Genome Analysis of the High Pathogenicity *Yersinia enterocolitica* Strain 8081. *PLoS Genet* 2: e206.
45. Burghout P, Beckers F, de Wit E, van Boxtel R, Cornelis GR (2004) Role of the Pilot Protein YscW in the Biogenesis of the YscC Secretin in *Yersinia enterocolitica*. *J Bacteriol* 186: 5366–5375.
46. National Center for Biotechnology Information (NCBI), <http://www.ncbi.nlm.nih.gov/>.
47. Oh JD, Kling-Bäckhed H, Ginnakis M, Xu J, Fulton RS, et al. (2006) The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proc Natl Acad Sci USA* 103: 9999–10004.
48. Alm RA, Ling LS, Moir DT, King BL, Brown ED, et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397: 176–80.
49. Eppinger M, Baar C, Linz B, Raddatz G, Lanz C, et al. (2006) Who ate whom? Adaptive *Helicobacter* genomic changes that accompanied a host jump from early humans to large felines. *PLoS Genet* 2: e120.
50. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, et al. (2009) Sequence-based prediction of type III secreted proteins. *PLoS Pathog* 5: e1000376.
51. Samudrala R, Heffron F, McDermott JE (2009) Accurate Prediction of Secreted Substrates and Identification of a Conserved Putative Secretion Signal for Type III Secretion Systems. *PLoS Pathog* 5(4): e1000375.

11. 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 Löwer)

12. Lebenslauf

Persönliche Daten

Martin Löwer

Yorckstraße 13

65195 Wiesbaden

Telefon 0611/4475066

Email martin.loewer@gmx.de

Geburtstag 22.10.1980

Geburtsort Frankfurt am Main

Akademische Ausbildung

- 08/2006 - heute Doktorand
- Arbeitsgruppe: Prof. Dr. Gisbert Schneider (Leiter), Beilstein Stiftungsprofessur für Chemie- & Bioinformatik, Goethe Universität, Frankfurt a.M., <http://www.modlab.de>
- Laboraufenthalte: Dr. Silja Wessler (Leiterin), Nachwuchsforschungsgruppe für Pathogen- Wirts Beziehungen, Paul-Ehrlich Institut, Langen; Prof. Dr. Anna Starzinski-Powitz (Leiterin), Humangenetik, Goethe Universität, Frankfurt a.M.
- Forschungsgebiete: Virulenzfaktoren von *Helicobacter pylori*, bakterielle Transportmechanismen, struktur-basiertes virtuelles Screening
- Thema der Dissertation: "Virtuelles Screening nach Inhibitoren der Protease HtrA aus *Helicobacter pylori*"
- 10/2001 - 06/2006 Student der Bioinformatik
- Universität: Goethe Universität, Frankfurt a.M.
- Abschluss: Diplom in Bioinformatik
- Note: "sehr gut" (1.4)
- Hauptfächer: Theoretische und praktische Bioinformatik, Organische Chemie, Biochemie
- Thema der Diplomarbeit: "*In silico* – Suche nach Protease-Genen im Genom von *Helicobacter pylori*"

Arbeitserfahrung

- 10/2007 - heute Wissenschaftlicher Mitarbeiter der Johann-Wolfgang-Goethe Universität
Aufgaben: Forschung in Chemie- und Bioinformatik, Betreuung der Praktika "Chemie für Mediziner"
- 11/2004 - 03/2005 Studentische Hilfskraft (Softwareentwicklung)
- 04/2004 - 09/2004 Studentische Hilfskraft (Betreuung von Programmierpraktika)
- 08/2000 - 06/2001 Zivildienst

Stipendien

- 10/2006 - 9/2007 Center for Membrane Proteomics (Goethe Universität)
Nachwuchsforscherstipendium

Veröffentlichungen

- Löwer M, Weydig C, Metzler D, Reuter A, Starzinski-Powitz A, Wessler S, Schneider G (2008) Prediction of Extracellular Proteases of the Human Pathogen *Helicobacter pylori* Reveals Proteolytic Activity of the Hp1018/19 Protein HtrA. PLoS ONE 3(10).
- Löwer M, Tankrikulu Y, Weisel M, Schneider G (2008) Fuzzy Virtual Ligands for Virtual Screening. 4. German Conference for Cheminformatics, Goslar, November 9-11.
- Löwer M, Schneider G (2009) Prediction of Type III Secretion Signals in Genomes of Gram-negative Bacteria. PLoS ONE 4(6).
- Löwer M, Tankrikulu Y, Weisel M und Schneider G (2009) Structure Based Virtual Screening for Inhibitors of the Novel Protease HtrA of *Helicobacter pylori*. *In Vorbereitung*.
- Weydig C, Löwer M, Hoy B, Carra G, Backert S, Schneider G, Wessler S (2009) *Helicobacter pylori* HtrA cleaves E-Cadherin to disrupt intercellular adhesion. *Eingereicht*.