Optimiertes Design kombinatorischer Verbindungsbibliotheken durch Genetische Algorithmen und deren Bewertung anhand wissensbasierter Protein-Ligand Bindungsprofile

Dissertation

zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem

Fachbereich Pharmazie

der Philipps-Universität Marburg

vorgelegt von

Patrick Pfeffer

aus Frankfurt am Main

Marburg an der Lahn 2009

Vom Fachbereich Pharmazie der Philipps-Universität Marburg

als Dissertation angenommen am:

Erstgutachter: Prof. Dr. G. Klebe

Zweitgutachter: Prof. Dr. H. Gohlke

Tag der mündlichen Prüfung:

Die Untersuchungen zur vorliegenden Arbeit wurden auf Anregung von Herrn Prof. Dr. G. Klebe am Institut für Pharmazeutische Chemie des Fachbereichs Pharmazie der Philipps-Universität Marburg in der Zeit von Februar 2006 bis April 2009 durchgeführt.

„Ich bin wahrlich kein ehrgeiziger Mensch – ich will nur Weltmeister werden“

Ernest Hemingway

Inhaltsverzeichnis

1 Einleitung und Problemstellung

Eine der entscheidenden Herausforderungen im Frühstadium des Wirkstoffentwicklungsprozesses ist die zielgerichtete Suche nach neuen Leitstrukturen. Zum Auffinden solcher Strukturen werden derzeit neben experimentellen Methoden wie dem so genannten High Throughput Screening zwei computergestützte Strategien verfolgt: das Durchmustern von Substanzbibliotheken (virtuelles Screening) und die rechnergestützte Substanzentwicklung (de-novo Design).

Im Fall von computergestützter Wirkstoffentwicklung wird zwischen zwei einander ergänzenden Vorgehensweisen unterschieden, dem ligand- und dem strukturbasierten Moleküldesign. Im ersten Fall steht die Selektion und Validierung wirkstoffartiger Moleküle durch den Vergleich mit bekannten Liganden im Vordergrund. Solche Verfahren werden eingesetzt, wenn keine Strukturinformationen des jeweiligen Rezeptors zur Verfügung stehen. Beim strukturbasierten Moleküldesign gilt es, Selektion und Validierung mithilfe vorhandener Rezeptor-Strukturinformation durchzuführen. Hierbei können Hits zum Beispiel mit virtuellen Hochdurchsatz-Verfahren zum Einpassen von wirkstoffähnlichen Molekülen in Rezeptor-Bindetaschen (High Throughput Docking-Verfahren) gefunden werden [1]. Die Ergebnisse einer solchen Dockingsimulation liefern wichtige Informationen über die mögliche Orientierung und Affinität von Wirkstoffkandidaten in der Bindetasche des Zielmoleküls.

Von großer Bedeutung sind dabei die Bewertungsfunktionen. Sie bilden eine Entscheidungsgrundlage dafür, ob durch die Einpassung eines Liganden eine möglichst gute Approximation des nativen Bindemodus erzielt werden kann, und wie gut dessen Affinität vorhersagbar sein wird. Da in der Regel zwischen eingenommener Molekülgeometrie und daraus abgeleiteter Affinität eine Korrelation besteht, ist es das Ziel einer Dockingsimulation, den nativen Bindemodus des niedermolekularen Liganden zu rekonstruieren. Die verwendeten Strukturinformationen sind meist röntgenkristallographischen Daten entnommen. Die Herausforderung hierbei besteht in der Entwicklung von Bewertungsmethoden, die die Eigenschaften der untersuchten Systeme ausreichend genau abbilden. Es darf vermutet werden, dass sich durch solche Bewertungsmodelle potente Leitstrukturen auffinden lassen. Möglicherweise muss aber auch damit gerechnet werden, dass ein

Wirkstoff nicht aufgefunden bzw. negativ bewertet wird, weil er seine biochemische Aktivität über Mechanismen entfaltet, die in der Modellierung des betrachteten Systems nicht berücksichtigt wurden.

Ziel der vorliegenden Arbeit ist es, die rechnergestützte Suche nach neuen Leitstrukturen zu verbessern. Der Fokus liegt hierbei auf neuartigen Verfahren und Methoden, um Modelle von Wirkstoffkandidaten *in silico* zu erzeugen und zu bewerten. Das folgende Kapitel liefert zunächst einen Überblick der bekannten Verfahren zur Bewertung von Protein-Ligand Komplexen. Anschließend wird sowohl die Methodik als auch die *in silico*- und experimentelle Validierung der in dieser Arbeit entwickelten Bewertungsfunktion DrugScore Fingerprint behandelt. Diese hat ihren Ursprung in dem wissensbasierten Ansatz DrugScore [2]. Mit DrugScore können Bindungsgeometrien von Liganden in Protein-Bindetaschen bewertet und Bindungsaffinitäten vorhergesagt werden. Basierend auf diesem Ansatz wurde eine Modifikation der DrugScore-Methode vorgenommen, um eine benutzerdefinierte Anzahl von Protein-Ligand-Komplexen zu einem Rezeptor-Profil zusammenzufassen. Für jeden computergenerierten Bindungsmodus einer niedermolekularen Verbindung lässt sich ein entsprechendes Profil erzeugen. So können mithilfe von Ähnlichkeitsberechnungen Distanzen zwischen *in vivo*- und *in silico*-Profilen mit dem Ziel ermittelt werden, neuartige Leitstrukturen zu identifizieren, die bereits bekannte Interaktionsmuster in Form aktiv identifizierter Liganden aufweisen.

Ist eine Strategie zur Bewertung von Ligandgeometrien in Rezeptorbindetaschen gefunden, fällt der optimalen Auswahl mehrerer Liganden als Mitglieder einer nach kombinatorischen Prinzipien sowie rezeptorspezifisch aufgebauten Verbindungsbibliothek eine wichtige Aufgabe zu. Kapitel 3 behandelt daher das optimierte Design kombinatorischer Verbindungsbibliotheken. Nach einer kurzen Erläuterung der bereits bekannten Suchstrategien wird das in dieser Arbeit entwickelte Programm GARLig vorgestellt, welches unter Verwendung eines selbst-adaptiven Genetischen Algorithmus für eine bestmögliche Auswahl chemischer Seitenketten bei wirkstoffähnlichen Grundgerüsten sorgen soll. Zielsetzung ist hier die Zusammenstellung einer rezeptorspezifisch optimierten Verbindungsbibliothek, welche eine benutzerdefiniert große Untermenge aller möglichen chemischen Modifikationen Ligand-ähnlicher Grundgerüste beinhaltet. Als zentrales

Qualitätskriterium für die einzelnen Vertreter der Verbindungsbibliothek dienen durch Docking erzeugte Geometrien und deren Bewertungen durch sämtliche in dieser Arbeit behandelten Protein-Ligand-Bewertungsfunktionen.

2 Bewertung niedermolekularer Verbindungen anhand wissensbasierter Protein-Ligand Bindungsprofile

Die Erzeugung nativ-ähnlicher Bindungsgeometrien von Liganden in der Bindetasche eines untersuchten Zielproteins ist die erste Voraussetzung für den Erfolg virtueller Screening-Ansätze. Das so genannte Dockingproblem wird prinzipiell als gelöst angesehen [3], jedoch limitieren Protein-Ligand Bewertungsfunktionen oftmals die Algorithmen, die die Ligandgeometrie erzeugen. Somit können virtuell erzeugte Geometrien durch eine schlechte Bewertung bestraft werden, obwohl diese nativen Charakter besitzen. Aus diesem Grund stellt die Bewertung der virtuell erzeugten Ligandgeometrien die entscheidende Herausforderung in diesem Forschungsgebiet dar [4]. In diesem Kontext ist auch die im Rahmen dieser Arbeit entwickelte Bewertungsfunktion DrugScore Fingerprint ($DrugScore^{FP}$) zu verstehen, welche als Weiterentwicklung des SIFT-Ansatzes [5] unter Verwendung der von Velec *et al.* publizierten $DrugScore^{CSD}$ Methodik [6] angesehen werden darf. Die aus Kristallstrukturdaten niedermolekularer Verbindungen abgeleiteten Paarpotentiale werden verwendet, um für eine gegebene Bindetasche eine benutzerdefinierte Anzahl von Protein-Ligand-Komplexen zu einem Rezeptor-Profil in vektorieller Form zusammenzufassen. Dieser so genannte Referenzvektor bildet ein Bindungsprofil auf der Basis bereits bekannter und aktiver Liganden. Dieser Referenzvektor, der zu jedem Atom der Bindetasche die entsprechenden Potentialwerte enthält, kann nun mit Vektoren verglichen werden, die Informationen aus virtuell erzeugten Ligandgeometrien enthalten. Letztlich ist so die Ähnlichkeit zwischen virtuellen und nativen Bindungsmoden bewertbar bzw. können so Affinitätsabschätzungen getroffen werden.

2.1 Literaturbekannte Ansätze zur Bewertung von Protein-Ligand Komplexen

Die Suche nach neuen Leitstrukturen für einen gegebenen Rezeptor stellt eine große Herausforderung im Entwicklungsprozess von Wirkstoffen dar. Ansätze wie virtuelles Screening verwenden daher oftmals Methoden, um Bindungsmodi und -affinitäten von Liganden in einer Bindetasche möglichst effizient zu bewerten. Aus diesem Grund sollen Bewertungsfunktionen über die folgenden Eigenschaften verfügen:

- sie sollen schnell viele Komplexe bewerten
- sie sollen die computergenerierten Ligandgeometrien favorisieren, welche hinsichtlich der chemischen als auch der sterischen Eigenschaften möglichst komplementär zum Zielmolekül sind
- sie sollen generierte Lösungen korrekt bewerten (d.h. die auf Rang 1 vorgeschlagene Lösung sollte eine geringe geometrische Abweichung zur experimentell bestimmten Struktur des Liganden aufweisen)
- sie sollen die freie Bindungsenthalpie ΔG korrekt vorhersagen

Die Bindungsaffinität eines Rezeptors (R) und eines Liganden (L) kann experimentell über die Inhibitionskonstante K_i [mol/l] in wässriger, elektrolythaltiger Lösung ermittelt werden:

$$K_i = \frac{[L][R]}{[LR]} \qquad \textbf{(1)}$$

Bei einer Temperatur T = 310 K entspricht ein K_i = 10^{-9} M einer freien Bindungsenthalpie ΔG von 51 kJ/mol.

ΔG ist ein Maß für die Neigung eines Moleküls zur Assoziation mit einem anderen Molekül und setzt sich aus enthalpischen und entropischen Beiträgen zusammen, wobei T die absolute Temperatur und R die allgemeine Gaskonstante bezeichnet:

$$\Delta G = -RT \ln K_i = \Delta H - T\Delta S \qquad \textbf{(2)}$$

In enthalpische Beiträge (Δ*H)* gehen hauptsächlich Anteile durch die Ausbildung von Wasserstoffbrücken und van der Waals-Wechselwirkungen ein. In den entropischen Beiträgen (ΔS) kommt es zum Verlust von Freiheitsgraden bezüglich Rezeptor und Ligand, sowie zur Freisetzung von Wasser von hydrophoben Oberflächen in die Volumenphase.

Die Abweichung einer im Computer erzeugten Molekülgeometrie gegenüber der experimentell bestimmten kann mit dem *rmsd*-Wert (**r**oot **m**ean **s**quare **d**eviation) ermittelt werden. Er stellt ein Maß für die mittlere quadratische Abweichung der kartesischen Koordinaten $\vec{X}$ der Atome L_k zwischen generierten und nativen Konformationen dar, wobei Wasserstoffatome in der Regel nicht berücksichtigt werden:

$$rmsd = \sqrt{\sum_{l \in L_k} \frac{(|\vec{X}_l^{\,nativ} - \vec{X}_l^{\,generiert}|)^2}{||L_k||}} \qquad \textbf{(3)}$$

Die drei bekanntesten Arten von Bewertungsfunktionen sind Kraftfeld-basierte, empirische (auch „regressionsbasierte Ansätze“ genannt) und wissensbasierte Bewertungsfunktionen. Hinzukommen nun als wesentliche Bestandteile der vorliegenden Arbeit die so genannten Fingerprint-basierten Bewertungsansätze, welche über implizite Wechselwirkungsinformationen hinaus einen Einfluss expliziter und benutzerdefinierter Zusatzinformationen zulassen, um so den Grad der Vorhersagegenauigkeit bisheriger Bewertungsansätze steigern zu können. Alle soeben erwähnten Verfahren werden in den nächsten Abschnitten erläutert.

2.1.1 Kraftfeld-basierte Bewertungsfunktionen

Kraftfeld-basierte Bewertungsfunktionen beinhalten Terme der Molekülmechanik, wobei die Summe dieser Terme, die Protein-Ligand-Wechselwirkung und interne Ligandenenergie beschreiben, eine Abschätzung der Affinität einer niedermolekularen Verbindung zu seinem Zielprotein erlaubt. Da in allen folgenden Betrachtungen ein rigides Protein vorausgesetzt wird, kann hier auf die Berechnung interner Proteinenergie verzichtet werden. Die Protein-Ligand-Wechselwirkungen werden hierbei durch eine Kombination von Elektrostatik- und van-der-Waals-Termen beschrieben. Das elektrostatische Potential verwendet eine distanzabhängige Dielektrizitätsfunktion und beschreibt die paarweise Summe der Coulomb-Interaktionen:

$$E_{Coul}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}, \qquad \textbf{(4)}$$

wobei *N* der Anzahl der Atome in Molekül A bzw. B und *q* der Ladung eines Atoms entspricht. Die van-der-Waals-Energie zur Beschreibung der nicht-kovalenten Wechselwirkungen wird oftmals durch das Lennard-Jones-Potential modelliert (hier in der 12-6-Form dargestellt):

$$E_{vdW}(r) = \sum_{i=1}^{N} \sum_{j=1}^{N} 4\varepsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \qquad \textbf{(5)}$$

Variiert man die Potenzen dieser Gleichung, lässt sich die ‘Schärfe‘ der aus Gleichung 5 resultierenden Bewertungsfunktion regulieren. Ein 8-4-Lennard-Jones-Potential toleriert beispielsweise eher geringere Distanzen zweier Atome als ein 12-6-Lennard-Jones-Potential. Die intramolekularen Terme zur Beschreibung der Ligandenenergie verwenden ebenfalls van-der-Waals- und elektrostatische Terme und sind denen intermolekularer Wechselwirkungen sehr ähnlich. Neben DOCK [7] ist die in dieser Arbeit betrachtete Funktion AutoDock Score [8] eine etablierte

Bewertungsfunktion, welche sich beispielsweise der Parameter des Amber-Kraftfeldes bedient [9].

2.1.2 Empirische Bewertungsfunktionen

Bei empirischen Bewertungsfunktionen wird die durch Böhm *et al.* formulierte Idee verfolgt, die Neigung einer niedermolekularen Verbindung zur Assoziation mit einem Rezeptor durch die Summe einzelner, additiver Terme zu beschreiben [10]. Dabei werden Informationen und Gesetzmäßigkeiten aus experimentell bestimmten Affinitäts- oder Strukturdatensätzen zu Gewichtungsfaktoren verdichtet, die mit Verfahren der multiplen linearen Regression ermittelt werden können.
Die zur Modellierung solcher Bewertungsfunktionen verwendeten Terme können auch die bereits erwähnten Kraftfeld-basierten Terme wie das Lennard-Jones- oder das Coulomb-Potential beinhalten. Desweiteren fließen oftmals molekulare Deskriptoren wie die Anzahl an Wasserstoffbrücken, hydrophobe Wechselwirkungen, frei drehbare Bindungen, Wasserstoffbrücken-Donoren und –Akzeptoren oder das Molekulargewicht in das Bewertungsmodell mit ein.
Die Ableitung solcher Funktionen und die anschließenden Affinitätsabschätzungen bei Protein-Ligand Komplexen können zwar schnell ausgeführt werden, es muss aber meist ein Abstrich bei der Güte der gefundenen Lösung gemacht werden, denn die Allgemeingültigkeit empirischer Bewertungsfunktionen hängt sehr stark von der Art und der Anzahl der Trainingskomplexe ab. Generell reicht empirisches Wissen zur hinreichenden Erkenntnis von Gesetzmäßigkeiten eines zu untersuchenden Systems nicht aus.
Prominente Vertreter dieser Klasse sind die in LUDI verwendete Bewertungsfunktion [11], F-Score [12] und ChemScore [13] aus den Docking-Programmen FlexX und GOLD.

2.1.3 Wissensbasierte Bewertungsfunktionen

Bei wissensbasierten Bewertungsfunktionen wird aus einer Quelle mit ausreichend vorhandenem Wissen eine Statistik abgeleitet, welche Auskunft über Häufigkeiten von Beobachtungen geben soll. Im Kontext des computergestützten Wirkstoffdesigns werden Atom-Atom-Wechselwirkungen in der Regel aus experimentell bestimmten Protein-Ligand Kristallstrukturen gezählt. Beobachtungen in der Nähe von Häufigkeitsmaxima werden als energetisch günstig angesehen und dementsprechend favorisiert. Aus dieser der Strukturinformation entnommenen Statistik lassen sich unter Anwendung eines Referenzzustandes sowie unter Annahme der Gültigkeit des inversen Boltzmann'schen Gesetzes für den betrachteten Datensatz Paarpotentiale zur Beschreibung atomarer Wechselwirkungen ableiten [14]. Die Bewertung der Affinität einer niedermolekularen Verbindung zu seinem Rezeptor findet durch die Ermittlung und anschließende Summation der Potential-Einzelbeiträge von Protein-Ligand Wechselwirkungen statt, wobei sich die aus einer Berechnung mit wissensbasierten Bewertungsfunktionen resultierenden Werte nicht direkt zur Affinitätsvorhersage eignen.

Die Unterschiede einzelner Verfahren liegen hauptsächlich in der Wahl des zugrundeliegenden Datensatzes, der verwendeten Atomtypen sowie des Referenzzustandes und der betrachteten Wechselwirkungs-Distanzen.

Bei wissensbasierten Ansätzen werden zwar wie bei empirischen Bewertungsfunktionen aus einem Trainingsdatensatz allgemeine Rückschlüsse gezogen, jedoch ist bei ersteren die Trainingsmenge größer und somit als detailreichere Modellierung der Wechselwirkungsmechanismen zwischen niedermolekularen Verbindungen und Rezeptormolekülen anzusehen.

Erfolgreiche Ansätze sind aus der Proteinfaltungsvorhersage bekannt [15]. Um prominente Vertreter von Protein-Ligand Bewertungsfunktionen zu nennen, seien PMF [16], DrugScore und SMoG [17] erwähnt.

2.2 *DrugScore^{FP}: Profiling Protein-Ligand Interactions*

DrugScore Fingerprint (DrugScore^{FP}) is a novel vector-based scoring function based on statistical pair potentials as derived by the DrugScore^{CSD} formalism. In contrast to DrugScore^{CSD} the overall score is partitioned into a per-protein-atom score vector. Simple distance metrics allow the determination of similarities between fingerprints of docked compounds and reference fingerprints derived from e.g. crystal structures. Available information about known ligands can be regarded as a reference by generating a weighted consensus fingerprint, resulting in a protein-based binding profile. In contrast to the program SIFT, which is also capable of generating protein-based fingerprints, DrugScore^{FP} binding profiles do not only capture similarities, but also consider dissimilarities with respect to a given drug target. In recognizing near-native docking poses for the Wang data set, DrugScore^{FP} showed improved results compared to DrugScore^{CSD} and SIFT. We performed cross-validation studies for trypsin and HIV-1 protease using compounds from the National Cancer Institute Diversity Set (NCI) which were docked into the targets by the program GOLD. Here DrugScore^{FP} offers better enrichments compared to GOLDScore and DrugScore^{CSD}. The approach also suggests two novel fragments actively inhibiting trypsin. Additionally, a virtual screening for trypsin- and HIV-1 protease was performed using the more challenging DUD dataset. Our results implicate that DrugScore^{FP} is especially useful to identify fragment-like compounds in a computer screening. As a practical proof-of-concept, we performed a second virtual screening run on two targets, tRNA-guanine transglycosylase (TGT) and thermolysin. For this purpose, we used the fragment-like subset of the ZINC database along with a fragment-like in-house database. For one of the discovered screening hits *N*-benzoyl-β-alanine, the crystal structure in complex with thermolysin could be determined. The structure proves that a docking geometry close to the subsequently determined crystal structure has been selected by DrugScore^{FP} demonstrating its superior performance for fragment-based virtual screening. Furthermore, several hits discovered by this procedure could be confirmed as active against TGT and thermolysin.

2.2.1 Introduction

Structure-based drug design is increasingly used for the discovery of novel lead structures along with their subsequent optimization [18]. This process is frequently supported by computational approaches using virtual screening. In this context, docking is applied as a crucial step. The *in silico* methods try to generate and identify near-native binding poses of small molecule candidates in the binding pocket of the target protein [19]. Docking solutions are generated exploiting information about the binding pocket and the residues available to form interactions with a ligand. Usually multiple solutions with alternative ligand poses are generated. These solutions have to be ranked by sophisticated scoring functions to suggest the most probable docking mode to the user. Most docking methods neglect in a rather unsatisfactory way possibly given a priori knowledge about ligands known to bind to the target protein. Better tailored approaches are desirable, which combine the statistical information about atom-atom pair-potentials as comprised e.g. in the knowledge-based scoring function DrugScore^{CSD} with additional target-specific structural data.

One general hypothesis in structure-based design assumes that structurally related molecules give similar biological responses. In early days of drug discovery when any knowledge about the 3D structure of the targeted macromolecule was absent, this similarity principle has formed the basis of QSAR methods [20]. Subsequently it was further exploited in approaches such as feature trees [21] or Catalyst (Accelrys Software, Inc., San Diego, CA) or software for database retrieval.

In the field of molecular docking, similarity-driven approaches have been suggested that consider ligand information in terms of a spatial pharmacophore. Such pharmacophores have been derived from ligand data or derived from the protein [22, 23]. Fradera *et al.* modified DOCK 4.0 using a ligand similarity score to weight the DOCK energy score [24]. Hindle *et al.* introduced the docking on predefined pharmacophore patterns into FlexX [25] and Cross *et al.* reported a FlexS/FlexX hybrid docking approach which is guided by a FlexS superimposition of the test ligand onto an experimentally determined reference ligand [26]. Radestock *et al.* have described an improvement of docking solutions by using the AFMoC approach which is based on protein-specifically adapted DrugScore potential fields and also allows for the integration of information about known binders [27, 28].

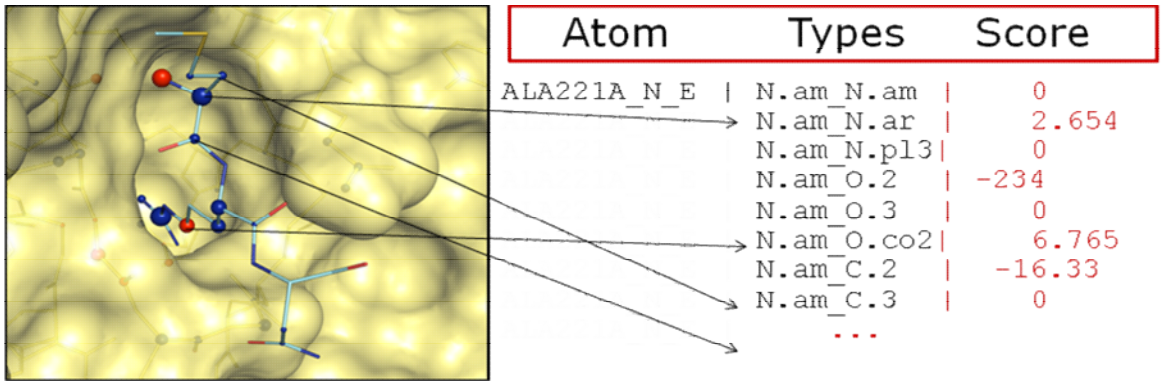


Figure 1: Principle of partitioning an overall DrugScore ranking into per-atom scores resulting in a DrugScore Fingerprint vector. Blue/red balls scaled by their size denote favorable/unfavorable contributions to the protein-ligand atom-atom interactions under consideration to the overall binding score.

The docking program GOLD uses spatial constraints such as receptor-based pharmacophores allowing only for docking poses which satisfy the predefined pharmacophore criteria.

Our strategy is based on the DrugScore methodology [2, 29] combined with the recently introduced SIFT-like fingerprint approach [5, 30, 31]. DrugScore^{CSD} is a knowledge-based scoring function that operates on statistical pair potentials derived from small molecule crystal data [6]. This function was shown to reliably recognize near-native poses out of a set of widespread decoys. SIFT and similar approaches compute a residue-based protein-ligand interaction fingerprint to elucidate criteria for inhibitor selectivity [32, 33]. Recently, other methods have been reported which calculate molecular fingerprints and gain improved retrieval rates of active compounds in virtual screenings by clustering similar molecules [34, 35].

Our novel approach DrugScore^{FP} (DrugScore Fingerprint) exploits structural information about known protein-ligand complexes and encodes this information in simple fixed-length vectors which capture individual site-specific atom-atom interaction values and encode the original DrugScore in a delocalized fashion as shown in Figure 1.

A DrugScore fingerprint can be computed from at least one- up to sets of multiple crystal structures. Distance metrics such as Euclidean- or Manhattan distance can be used to capture similarities between different fingerprints and a weighting function can be applied to increase the impact of certain protein-ligand interaction types.

2.2.2 Theory

The DrugScore^{FP} fingerprints are vectors in $\Re^n$, where *n* depends on the number of protein atoms *p* representing the binding pocket and the number of possible atom type combinations considered in the applied set of pair potentials.

Evaluation is proceeded in two steps: First a consensus fingerprint $\underline{f}$ is calculated for a given set of reference structures (e.g. X-ray structures) and subsequently similarities between fingerprints $\underline{x}$ of probes (e.g. docking solutions) and the reference are calculated. For a given set *K* of protein-ligand complexes (of the same target) each protein atom within a distance *d* (a default value of 6.1 Å is used) to any ligand atom *l* is considered to be a part of the binding pocket *D*.

The input for the computation of the consensus vector is a list of protein structures together with their corresponding ligands. In principle, the same protein geometry can be used for all ligands, particularly in case, the individual binding pockets are structurally highly conserved across the data set.

Each pocket atom contributes a number of components to $\underline{f}$ corresponding to the number of different pair potentials $W_{i,j}$ for its atom type *i* and all possible ligand atom types *j*, where each component is simply the mean of the according potential value over all input complexes:

$$f_{i(p)\_j} = \sum_{k} \sum_{p_k : p_k \in D} \sum_{l_k : j(l_k)=j} \frac{W_{i(p_k),j}(dist(p_k, l_k))}{|K|} \qquad \textbf{(6)}$$

The components for the fingerprint $\underline{x}$ of a single complex are calculated similarly:

$$x_{i(p)\_j} = \sum_{p : p \in D} \sum_{l : j(l)=j} W_{i(p),j}(dist(p,l)) \qquad \textbf{(7)}$$

To quantify the similarities of different binding poses, we use either Euclidean or Manhattan distances between the corresponding probes and the reference.

A short distance denotes that the probe structure shows very similar interactions in its adopted conformation compared to the reference. Once a pose with a similar binding mode has been detected, high probability is given that the probe structure interacts with the target in a same fashion as the reference structure(s).

The individual contributions to the fingerprint vector will be of deviating importance. A strong anchor group for example, that is present in all reference structures should be much more determinant for the binding mode compared to an interaction that contributes to the consensus fingerprint with varying strength across the reference structures. Therefore it is also possible to calculate the distances using weights χ for the components of $\underline{f}$, depending on their variances among the reference structures:

$$\chi_{i(p)\_j} = \sqrt{\frac{\sum_k \sum_j (f_{i(p_k)\_j} - f_{i(p)\_j})^2}{|K| - 1}} + 1$$

$$sim_{euclidean,weighted} = \sqrt{\sum_{p:p\in D} \sum_j \frac{(x_{i(p)\_j} - f_{i(p)\_j})^2}{\chi^2_{i(p)\_j}}} \qquad \textbf{(8)}$$

To generate reasonable weights for the consensus fingerprint vector, of course a higher number of input structures is required. For interactions that are never observed in $\underline{f}$, the lowest weight is assigned (as it is possible to observe this interaction for a probe structure). Finally the similarity values are normalized to a range between zero and one, accordingly identical fingerprints will result in a value of 1.0 and the most dissimilar fingerprint will correspond to a value of 0.0.

Crucial to the approach is the setting of pair potentials or, more precisely, the partitioning in a reasonable set of atom types. The original CSD-derived DrugScore uses a set of 18 different types, based on the Sybyl atom type notation. This partitioning results in a huge number of 324 different atom-atom pair combinations. Within the data set it might happen that an H-bond is formed to the same protein

acceptor atom by two different donor functionalities in the ligands, thus creating dissimilar fingerprints simply because different types of pair potentials are used. Ideally for each residue only one type of potential should be used to describe a specific type of interaction. Thus, a more general definition of DrugScore atom types is necessary. The determination of the best suited definition is still a project of current research. For the results presented in this contribution we applied a potential set which allowed for best performance for the present studies. This set, in contrast to the original CSD-potentials, uses an asymmetric description of atom types with respect to protein and ligand. For the protein we still use the original atom type definitions, however for the ligand we assign only the following 11 types: ar(aromatic), hp(hydrophobic), don(H-donor), acc(H-acceptor), da(H-donor or acceptor), am(nitrogen in amides), F, Cl, Br, I and Met (Ca, Fe, Zn).

2.2.3 Programs, Datasets and Materials

Comparison with 14 different scoring functions. To examine the performance of DrugScore^{FP} in recognizing near-native docking poses, we used a data set of protein-ligand complexes published by Wang *et al.* This data is frequently used as benchmark and has been ranked by many of the currently available scoring functions. Furthermore, DrugScore^{CSD} has previously been validated using this data set. Thus, a direct comparison with the original DrugScore^{CSD} of our fingerprint-based approach is possible. The Wang data set consists of 100 crystallographically determined protein-ligand complexes, each supplemented by 100 docking poses computed with the program AutoDock [36]. The docking parameters were set in a way to cover a broad range of geometries, such that also docking poses largely deviating from the crystal structure were obtained. The overall computed geometries cover a range of up to 20 Å rmsd from the native crystal structure. DrugScore^{FP} was applied to this data set and compared to the published results obtained by the 13 different scoring functions and the fingerprint-based method SIFT.

Leave-One-Out (LOO) Cross-validation Study. A data set of 56 trypsin complexes (Figure 2) was assembled using the PDBBIND database [37], considering only crystal structures with a resolution ≤ 2 Å.

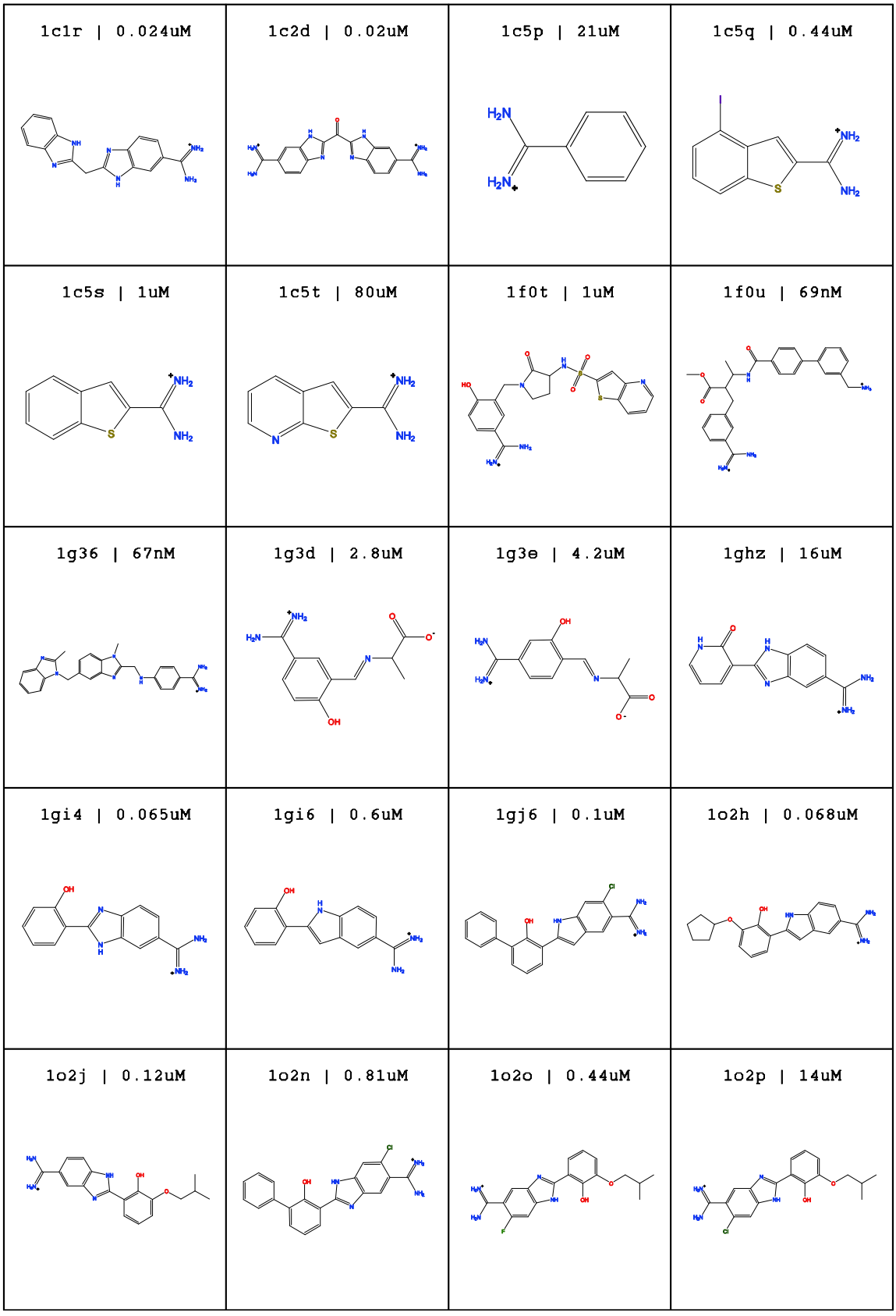


Figure 2.1: Composition of the trypsin data set (56 active compounds indicated by their corresponding pdb-codes). Affinity data is given as K_i-value.

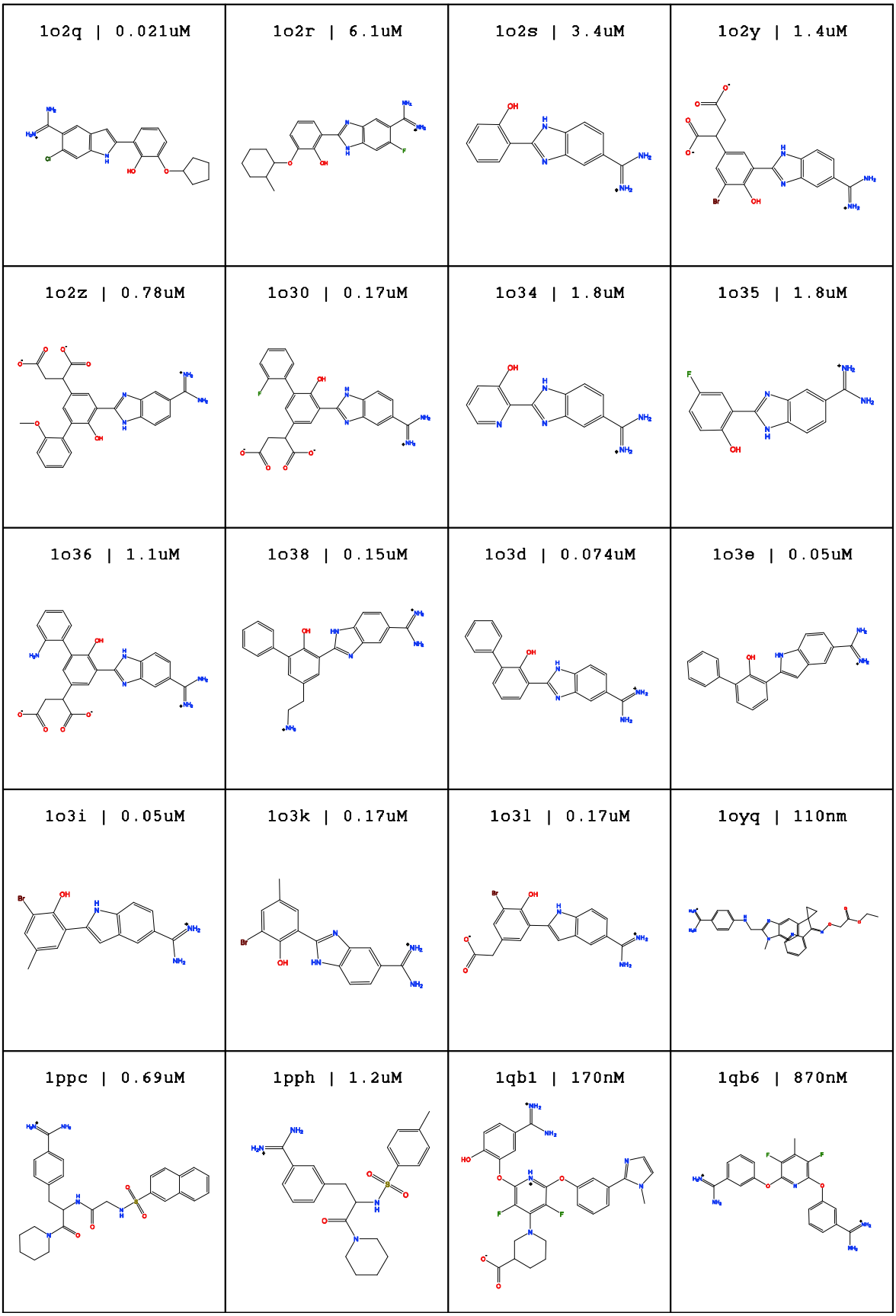


Figure 2.2: Composition of the trypsin data set (56 active compounds indicated by their corresponding pdb-codes). Affinity data is given as K_i-value.

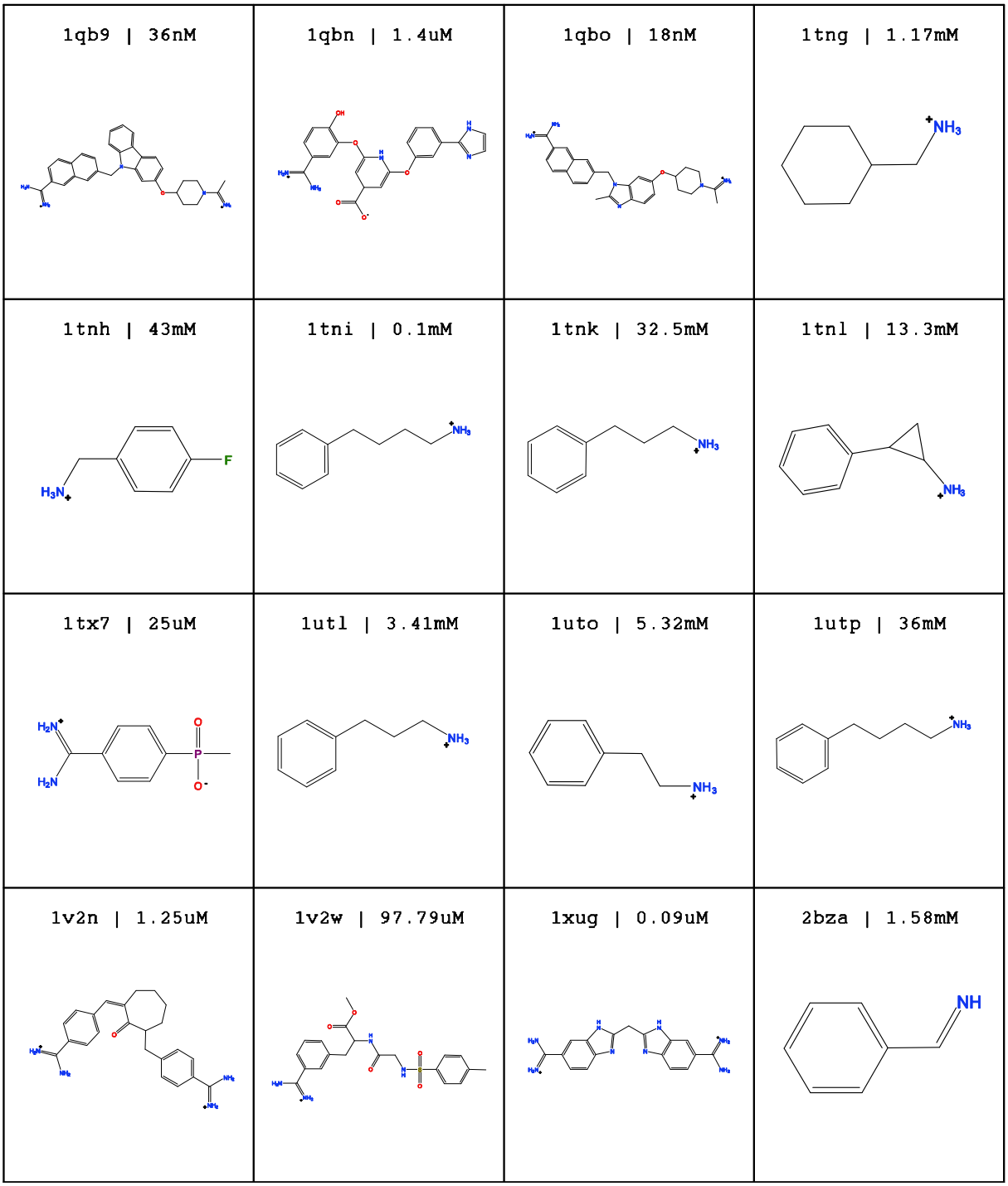


Figure 2.3: Composition of the trypsin data set (56 active compounds indicated by their corresponding pdb-codes). Affinity data is given as K_i-value.

For each complex, 10 docking solutions were computed using the docking program GOLD [38]. Standard parameters have been assigned to the docking run. For all compounds, standard protonation states were assumed, i.e. carboxylate groups were considered as deprotonated, and aliphatic amines and amidino-/guanidino groups

were handled as protonated. The protonation state of the protein was predicted at pH 8.0 using the program MOE [39]. A protein-based C_α-alignment was applied to the entire data set to obtain a consistent superimposition of all inhibitors. Subsequently, the MAB force field [40] was used to minimize all native ligands in the rigid binding pocket of trypsin (PDB code: 1pph). This protein reference was also selected for the subsequent docking calculations. DrugScore consensus fingerprints were computed for different subsets considering two up to 55 complexes of the crystal structures. The method was tested by checking whether a ligand pose is retrieved that falls close in geometry to the native pose among the remaining docking decoys on a top rank of the similarity list. We varied the number of reference ligands considered in the reference fingerprint to detect possible dependencies between chemical composition (homogeneous and diverse ligand subsets) and the predictive power of the approach. Furthermore, we tested the influence of our DrugScore^{FP} weighting function in a LOO-experiment. To select reasonable subsets either composed by similar or dissimilar ligands, MACCS keys [41] have been computed to rate similarities among all 56 inhibitors.

Virtual Screening using the National Cancer Institute diversity set. A virtual screening was performed on trypsin and HIV-1 protease. In the first case, the docking program GOLD was used to dock 1800 compounds of the NCI diversity set into trypsin. Default parameters were set as suggested for docking calculations with GOLD. The 56 known binders from the cross-validation study were used as active reference compounds and the 1800 docked candidate molecules from the NCI diversity set were considered as inactive compounds. To obtain reasonable results from our virtual screening run, the data set had to be prefiltered to exhibit an equal distribution of chemical properties either in the data set of active and inactive compounds. Therefore, Lipinski's Rule-of-5 [42] was applied to the NCI data set originally consisting of more than 3000 compounds. This selection should help to avoid artificial enrichments [43]. For HIV-1 protease, a similar validation scenario was defined. 22 active compounds (Figure 3) were retrieved from the PDBBIND database with a minimal resolution of ≤ 2 Å.

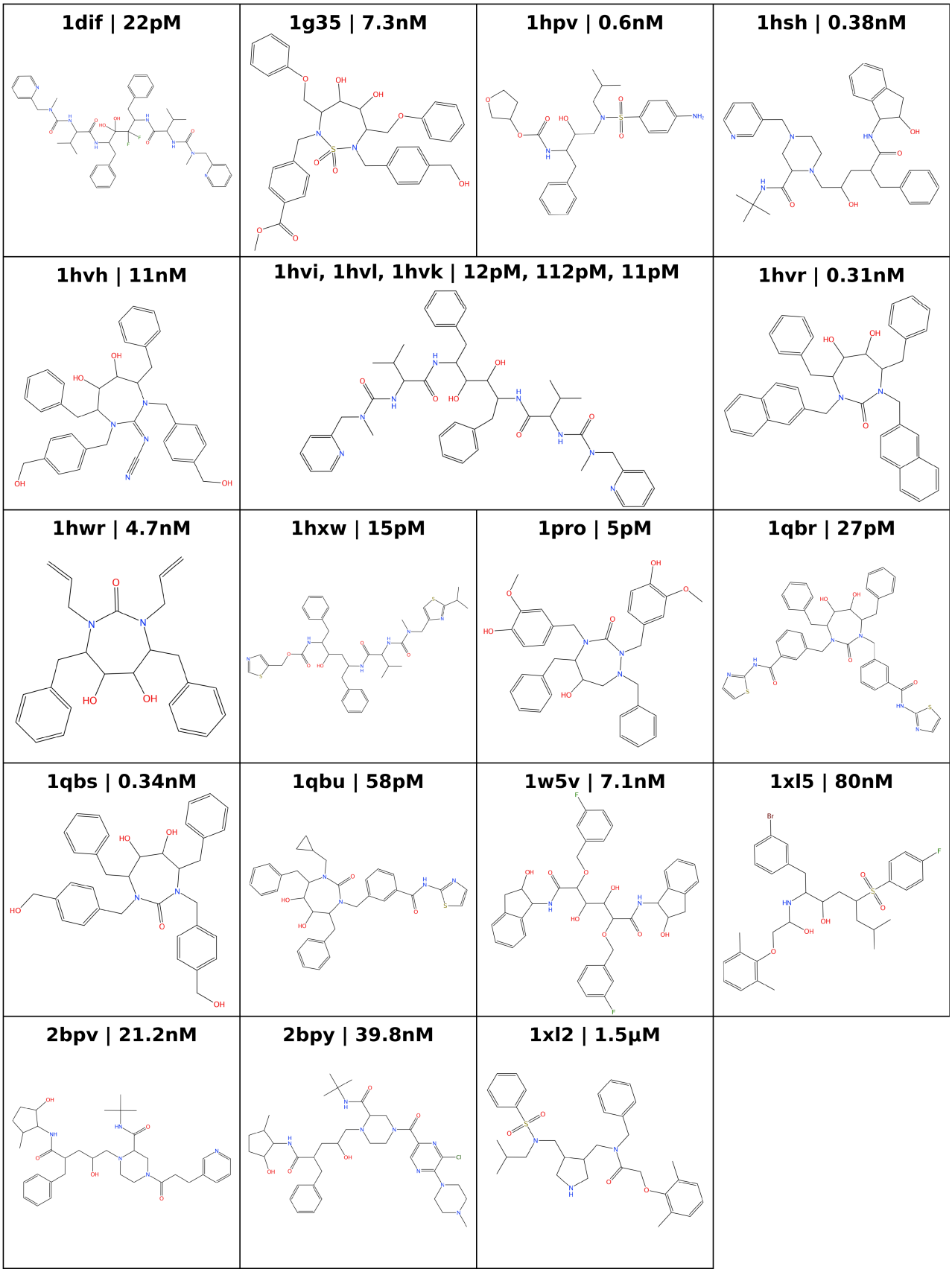


Figure 3: Composition of parts of the HIV-1 protease data set (active compounds indicated by their corresponding pdb-codes). Affinity data is given as K_i-value.

After a protein-based C_α-alignment of all entries with respect to the PDB-entry 1dif, all 22 inhibitors were minimized in the binding pocket using the MAB force field. Furthermore, we docked 70 compounds (assumed inactives) from the NCI diversity set into HIV-1 protease using GOLD. The standard settings of the program were applied. The assignment of protonation states was treated as mentioned above. Here, only compounds with a molecular weight ranging from 500 to 700 Da were considered from the NCI diversity set in order to reveal a similar distribution of chemical properties compared to the data set of known HIV-1 protease inhibitors. The results of both virtual screening studies were visualized using Receiver Operating Characteristic (ROC) curves and quantified using the Area Under the Curve (AUC) values. Again, we varied the size and the chemical composition of the DrugScore reference fingerprint in both cases. Furthermore, we tested the effect of our DrugScore^{FP} weighting function on the enrichment. Two different validation strategies were followed: (1) several crystal structures of known binders were merged with the docked geometries of the NCI diversity set. (2) Docking geometries of known binders were seeded into the docked NCI diversity set. The AUC values of DrugScore^{FP} were compared to DrugScore^{CSD}, GOLDScore and the residue-based interaction fingerprint PLIF (**P**rotein-**L**igand **I**nteraction **F**ingerprint), which has recently been implemented in MOE 2007.09. The PLIF fingerprint calculations were performed using the standard atom type notation as parameterized within MOE. PLIF consensus fingerprints were generated by averaging the interaction bits over different protein-ligand subsets which also have been used to calculate the DrugScore fingerprints. This procedure mimics best our DrugScore consensus fingerprint calculations. Finally, contact statistics between ligand atoms and protein residues have been computed to provide insight into similarities of interactions generated by the native ligands and the docked compounds of the NCI diversity set.

Virtual Screening using the Directory of Useful Decoys (DUD). We performed another virtual screening on trypsin and HIV-1 protease, this time using the DUD instead of the NCI data set. For trypsin, this dataset contains 43 active compounds and 1548 decoys. Crystal structures for 10 out of the 43 active compounds were selected from the PDB and used to derive a reference fingerprint (PDB codes: 1a0j, 1cit, 1f0t, 1k1i, 1k1l, 1o2p, 1o2r, 1oyq, 1qb6, 1qb9). Consistently, GOLD was used to generate docking solutions for the remaining 33 actives and the decoys. In the case

of HIV-1 protease, the DUD set holds 52 active compounds and 1872 decoys. Nine crystal structures were retrieved from the PDB to generate the reference fingerprint (PDB codes: 1g2k, 1hpo, 1hvh, 1hvr, 1hwr, 1ohr, 1qbs, 1rq9, 1upj). Again the remaining 43 actives together with the decoys were docked using GOLD. For both targets DrugScore^{FP}, DrugScore^{CSD} and GOLDScore were applied to rank all docking solutions. In the case of DrugScore^{FP}, no weighting was applied. Ten docking solutions were generated for each compound. AUC values for the resulting ROC curves were calculated for comparison.

Virtual Screening using the Fragment-like Subset of ZINC and an In-house Database. A promising target in structure-based drug design is tRNA-guanine transglycosylase (TGT), a protein involved in the pathogenicity mechanism of *Shigella flexneri*, the causative agent of Shigellosis. The enzyme exchanges guanine in the wobble position of tRNA $^{Asn,\ Asp,\ His,\ Tyr}$ against a modified base [44, 45]. Here, the docking program GOLD was used to dock 67489 compounds of the ZINC fragment-like subset and ~2000 fragment-like molecules taken from an in-house database into TGT. Default parameters were set as suggested for docking calculations with GOLD and protonation states were assigned as mentioned above.

As a second case study we used thermolysin, a thermostable endoprotease from *Bacillus thermoproteolyticus* with a zinc ion in the catalytic center. The enzyme is well studied and it is considered to be a prototype for zinc metalloproteases belonging to the gluzincin family. It often serves as a role model for other metalloproteinases. The docking programs GOLD and AutoDock were applied to produce reasonable binding poses. Since AutoDock performed best in retrieving native geometries from a set of 25 thermolysin crystal structures, it was used as docking engine for the present study. Protonation states were assigned as mentioned above. To avoid artificial ligand – zinc contacts during docking, the charge on the zinc ion in the binding pocket was modified from the default value of +2 to +0.5. The number of energy evaluations of the genetic algorithm in AutoDock was set to 1.5×10^6 and 20 solutions were generated for each fragment.

DrugScore Fingerprint Clustering. Finally, the calculated DrugScore fingerprints of the docked NCI diversity set and the native geometries of the known binders were used as input for the clustering program Cluto [46]. The Trypsin and HIV-1 protease

data set were evaluated, respectively. Different clustering algorithms were tested with the aim to automatically group similar fingerprint representations of binding modes together in order to produce meaningful dendrograms, separating known binders from inactive compounds. Finally, we decided to use a nonhierarchical k-means clustering algorithm [47] to maximize the pairwise similarities of the DrugScore fingerprints within a cluster.

2.2.4 Results and Discussion

Reranking of Ligand Poses. Our approach has been validated using the Wang data set to evaluate the reliability in recognizing near-native binding modes out of a set of widespread docking decoys. Our results will be faced to those obtained by other scoring functions as compiled by Velec *et al.* To perform the rescoring, a DrugScore fingerprint has been generated based on the crystallographically determined binding geometry. Table 1 lists for each scoring function the fraction of docked geometries with rmsd values ≤ 0.5 Å and ≤ 1 Å on the first scoring rank. The recovery rate of the crystal structures is not displayed here as this would be a trivial task for our method. Using DrugScore^{FP} with the chosen atom-type definition, in 94 % of the evaluated 100 protein-ligand complexes, a docked geometry with an rmsd value ≤ 0.5 Å is ranked best out of the total set of all 100 docking poses. DrugScore^{FP} shows with respect to the SIFT fingerprint method an improvement of 18 %. In this case, SIFT might be too general as it only considers residue-based interaction fingerprints. Furthermore, bitstring comparisons use the Tanimoto index which reveals only the similarity in the presence of features whereas our distance metric computes the presence as well as absence of features crucial for comparing protein-ligand complexes. If we relax conditions to the recovery rates of retrieved docking solutions with an rmsd value ≤ 1 Å, DrugScore^{CSD} performs slightly better than our fingerprint method. This might be an indication that DrugScore^{FP}'s performance relies more strongly on the quality of the produced docking poses.

Table 1: The Table corresponds to Table 1 from [6].

	success rate	
scoring function	rmsd ≤ 0.5Å	rmsd ≤ 1.0Å
DrugScore^{FP}	94%	79%
DrugScore^{CSD}	88%	83%
Cerius2/LigScore	88%	64%
DrugScore^{PDB}	81%	63%
SIFT	76%	60%
Cerius2/PLP	75%	63%
AutoDock	69%	34%
SYBYL/F-Score	63%	56%
Cerius2/LUDI	63%	43%
X-Score	50%	40%
Cerius2/PMF	38%	40%
Lennard Jones 12-6[a]	38%	65%
SYBYL/G-Score	25%	24%
SYBYL/ChemScore	6%	12%
SYBYL/D-Score	0%	8%

The success rate is given as percentage with respect to all complexes analyzed, allowing rmsd deviations as indicated. Scoring functions are sorted according to their success rates at rmsd values ≤ 0.5Å. Percentages denote results obtained when excluding the crystal structure geometry. [a]Scoring function is a standard Lennard-Jones 12-6 potential.

LOO Cross-validation Study. Trypsin is well suited for validation of our approach, as a considerable amount of structural information is available. The results of the LOO cross-validation are shown in Table 2. The native geometry of the known binder, which has been omitted to derive the DrugScore consensus fingerprint is ranked best in 75 % of the cases. Therefore, the 24 most similar ligand structures in terms of their MACCS keys were included in the fingerprint query. A recovery rate of 72 % could be achieved by using ligand structures, which are, according to MACCS keys, most dissimilar across the data set. In total, we fail in 25 % to recover the omitted native pose on rank 1 using a fingerprint composed by the similar ligands. This increases to 28 % once a fingerprint compiled from the dissimilar compounds is used. However, in all of these cases, poses that were placed on the best rank showed an rmsd value ≤ 0.5 Å with respect to the crystal structure. Obviously in some case, the near-native poses agrees slightly better to the fingerprint model.

Table 2: Percentages denote recognition rates for the “left-out” crystal structure out of docking decoys on similarity rank 1 obtained for different sizes of similar DrugScore Fingerprint compositions in the leave-one-out (LOO) cross-validation study. Values in parentheses denote results when using DrugScore Fingerprints composed of most dissimilar compounds.

Model Size	Recognition Rate
2 (-1)	51(46) %
3 (-1)	67(49) %
4 (-1)	66(54) %
5 (-1)	67(57) %
10 (-1)	69(62) %
25 (-1)	**75(72)** %
56 (-1)	70(70) %

Dissimilarity in ()

However, deviations of ≤ 0.5 Å should not be discussed as “significant difference” considering the accuracy of crystal structures and uncertainties introduced by the mutual fit of reference geometries.

An increasing improvement of the recovery rates is observed once we analyze the results of DrugScore^{FP} with respect to a growing number of reference ligands (varied from 1 to 24 inhibitors). Surprisingly, if the most comprehensive fingerprint model (55 inhibitors) is used, the recovery rate of the known binders is slightly worse (70 %) compared to a consensus fingerprint based on only 24 structures. Possibly, some degree of overfitting is given, as the reference data set is clearly not unbiased with respect to recursive features in its chemical composition. Obviously, the set of 24 ligands gives the best representation. Furthermore, no improvement could be recognized by applying our fingerprint weighting function. Obviously, the trypsin data is quite homogeneous with respect to the spatial distribution and scatter of the key interactions.

Virtual Screening using the NCI Diversity Set. Table 3 shows the results for the screening of trypsin, where the best performing DrugScore fingerprint query was constructed of only three out of 56 known binders.

Table 3: The recovery rate of known trypsin binders is given as AUC value. Different DrugScore Fingerprint compositions have been evaluated considering the fingerprint model size and the chemical composition (similar and dissimilar fingerprint models) as well as the influence of the DrugScore Fingerprint weighting function.

MAXIMUM DISSIMILARITY		
	AUC Score	
Model Size	weighting	no weight.
2	99.8 %	99.8 %
3	**99.9 %**	**99.9 %**
4	88.6%	99.5 %
5	87.5 %	99.4 %
10	81.3%	99.3 %
25	91.6 %	99.4 %
56	99.2%	99.6 %

MAXIMUM SIMILARITY		
	AUC Score	
Model Size	Weighting	no weight.
2	99.2 %	99.2 %
3	99.3%	99.3%
4	94.7 %	99.3%
5	93.6 %	99.4 %
10	92.3%	99.3%
25	92.3 %	99.5 %
56	99.2 %	**99.6 %**

Here, DrugScore^{FP} performs best in recovering the native geometries out of the docked NCI compounds (assumed non-binders) with an AUC value of 99.9 %. Figure 4a shows the compounds which have been used for computing the fingerprint query and Figure 4b the corresponding ROC curve.

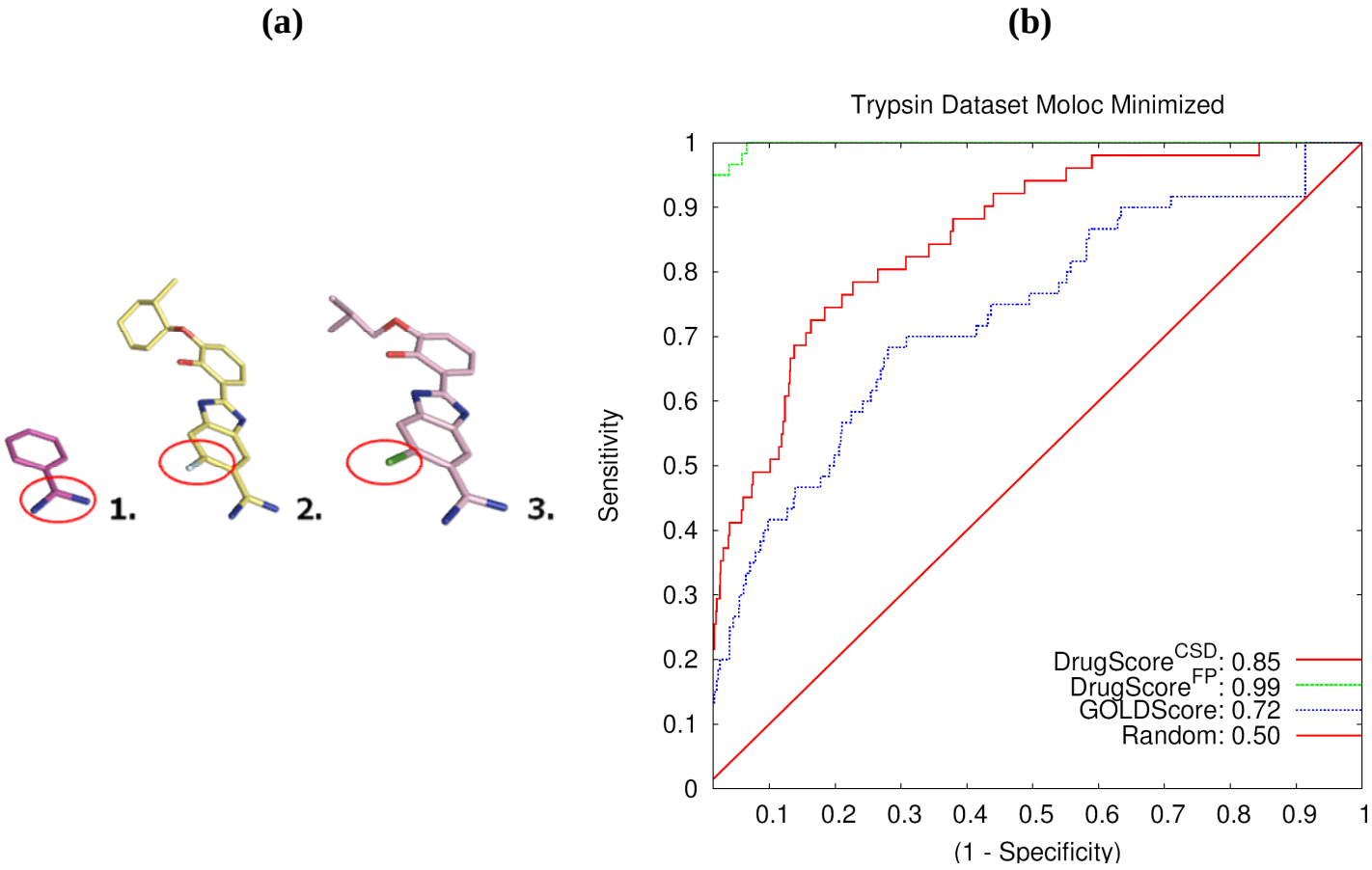


Figure 4: (a) An alignment of these three known trypsin binders have been used to compute the DrugScore Fingerprint. Red circles mark difference in the atomic composition of these scaffolds. (b) ROC plot for GOLD docking of 1800 compounds in the X-ray structure of trypsin (1pph) using DrugScore Fingerprint similarity score, DrugScore^{CSD} and GOLDScore to discriminate the 56 true active compounds from 1800 decoys.

Obviously in the trypsin case, a small consensus fingerprint is sufficient, likely because the trypsin data set is very homogeneous. For a trial we used the three most dissimilar compounds as reference and we still cover the range of possible key interactions in this data set, as it is sufficient to reliably retrieve nearly all 56 known binders at the top of the similarity list. DrugScore^{FP} performs slightly better than PLIF either derived with a full consensus fingerprint (56 ligands) or a fingerprint consisting of the same three ligands, which were used to calculate the best DrugScore fingerprint (PLIF AUC: 99.0 % in both cases, 1 % deviation reveals ~18 positions difference in the similarity list). This may be due to the fact that our method stores per-atom instead of per-residue information in the fingerprint vectors and takes advantage of the implicit incorporation of the knowledge-based CSD-potentials. DrugScore^{CSD} achieves only an AUC value of 85.0 %.

Table 4: The recovery rate of known HIV-1 protease binders is given as AUC value. Different DrugScore Fingerprint compositions have been evaluated considering the fingerprint model size and the chemical composition (similar and dissimilar fingerprint models) as well as the influence of the DrugScore Fingerprint weighting function.

MAXIMUM DISSIMILARITY

	AUC Score	
Model Size	weighting	No weight.
2	69.5 %	69.5%
3	77.3%	77.3%
4	63.2%	81.0%
5	63.7%	85.6%
10	85.9%	91.3%
22	**99.1%**	92.4 %

MAXIMUM SIMILARITY

	AUC Score	
Model Size	weighting	No weight.
2	57.4 %	57.4%
3	76.7%	76.7 %
4	80.0%	84.2%
5	81.0%	86.5%
10	83.1%	89.3%
22	**99.1%**	92.4 %

This figure indicates the impact of including structural information to the original version of the scoring function. GOLDScore performs worst with an AUC value of 72.0 %.

Table 4 shows the results of a similar screening for HIV-1 protease. In Figure 5a all compounds, which have been used for computing the DrugScore fingerprint and Figure 5b the corresponding ROC curve are displayed.

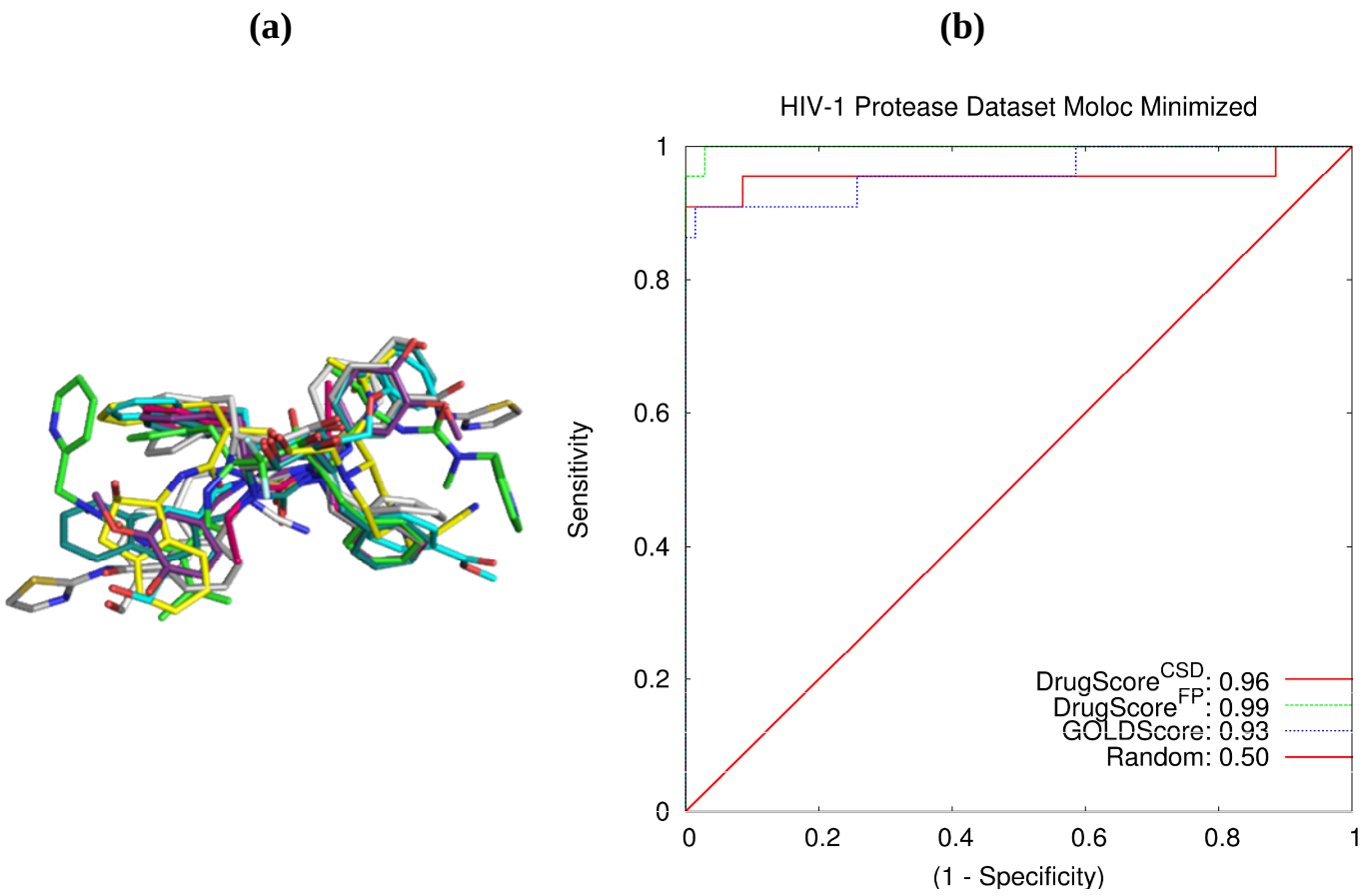


Figure 5: (a) Alignment of the 22 known HIV-1 protease binders used to compute the DrugScore Fingerprint. (b) ROC plot for GOLD docking of 70 compounds to the X-ray structure of HIV-1 protease (1dif) using DrugScore Fingerprint similarity score, DrugScore^{CSD} and GOLDScore to discriminate the 22 true active compounds from 70 decoys.

Again, DrugScore^{FP} outperforms the other scoring functions. Here, the virtual screening was carried out with a fingerprint composed of all 22 known HIV-1 protease binders to achieve the best AUC value of 99.1 %. In this case, obviously a complete consensus fingerprint performs better, likely because the HIV-1 protease data set exhibits a more pronounced molecular diversity compared to the trypsin case. Some protein-ligand interactions in the HIV-1 protease data set differ strongly among the different binders, and rather high standard deviations are observed for these DrugScore contributions. Accordingly, it appears reasonable to reduce the influence of such interactions in the reference fingerprint. In fact, applying our weighting function results in an improvement of about 7 % compared to the unweighted AUC value.

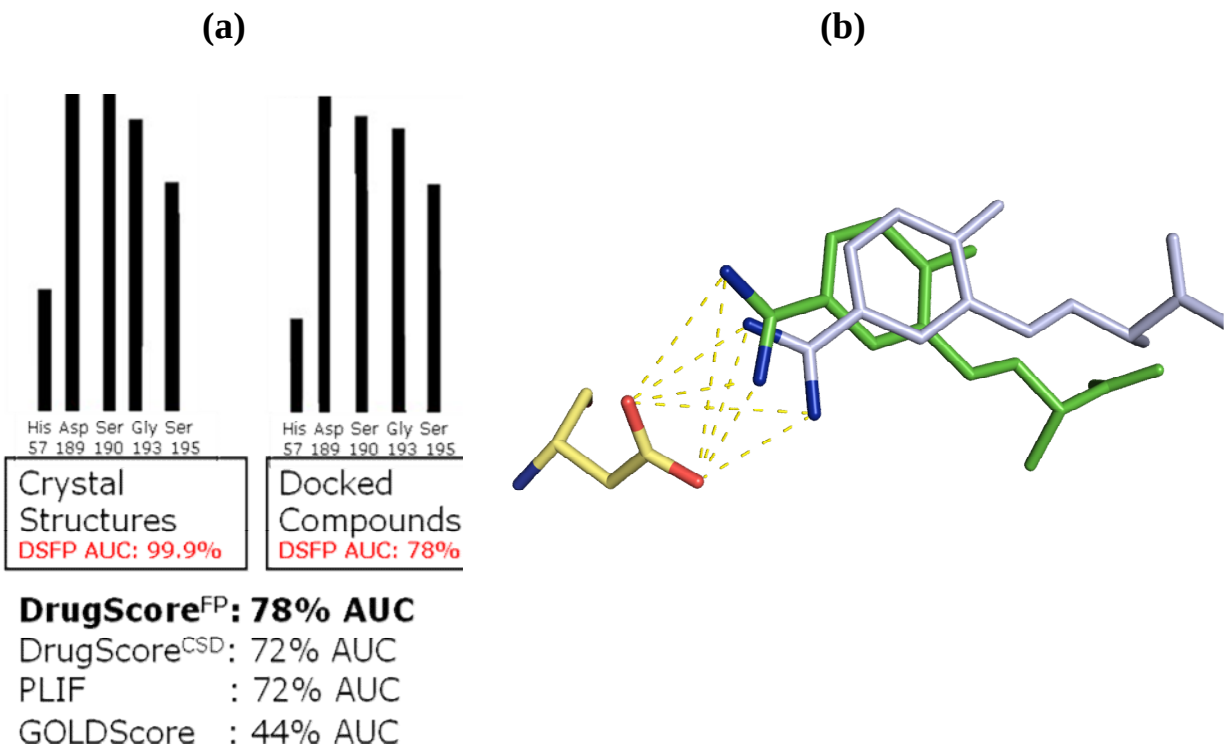


Figure 6: (a) Histogram showing the similarity of the interaction profiles between crystal structures and the corresponding docking poses of compounds interacting with the core residues of trypsin. Below, the recovery rate of docking poses of 54 active compounds seeded into 1800 inactives is given for the scoring functions DrugScore^{FP}, DrugScore^{CSD}, PLIF and GOLDScore. (b) Crystal structure (light blue) and docked geometry (green, 2.2Å rmsd) of a trypsin inhibitor. The computed binding mode shows similar key interactions to the protein as observed for the native geometry guaranteeing a successful virtual screening with DrugScore^{FP}. The docked compound is scored on similarity rank 43 of 1854 using DrugScore^{FP} and can only be found on rank 412 of 1854 using GOLDScore.

Furthermore, we performed a virtual screening on trypsin and HIV-1 protease again using the described DrugScore reference fingerprints to rank the docking hits. This time, it is not the issue to seed crystal structures but docking solutions of the known binders into the NCI diversity set used for docking. At first, it becomes clear that the recovery rates of the known binders strongly depend on the quality of the achieved docking solutions, which have a mean rmsd value of 2.8 Å compared to the crystal structures and a standard deviation of 2.8 Å in the case of the trypsin set. Here, DrugScore^{FP} performs better than the other scoring functions because the key interactions between the native and the docked geometries of the known binders are quite similar (Figure 6).

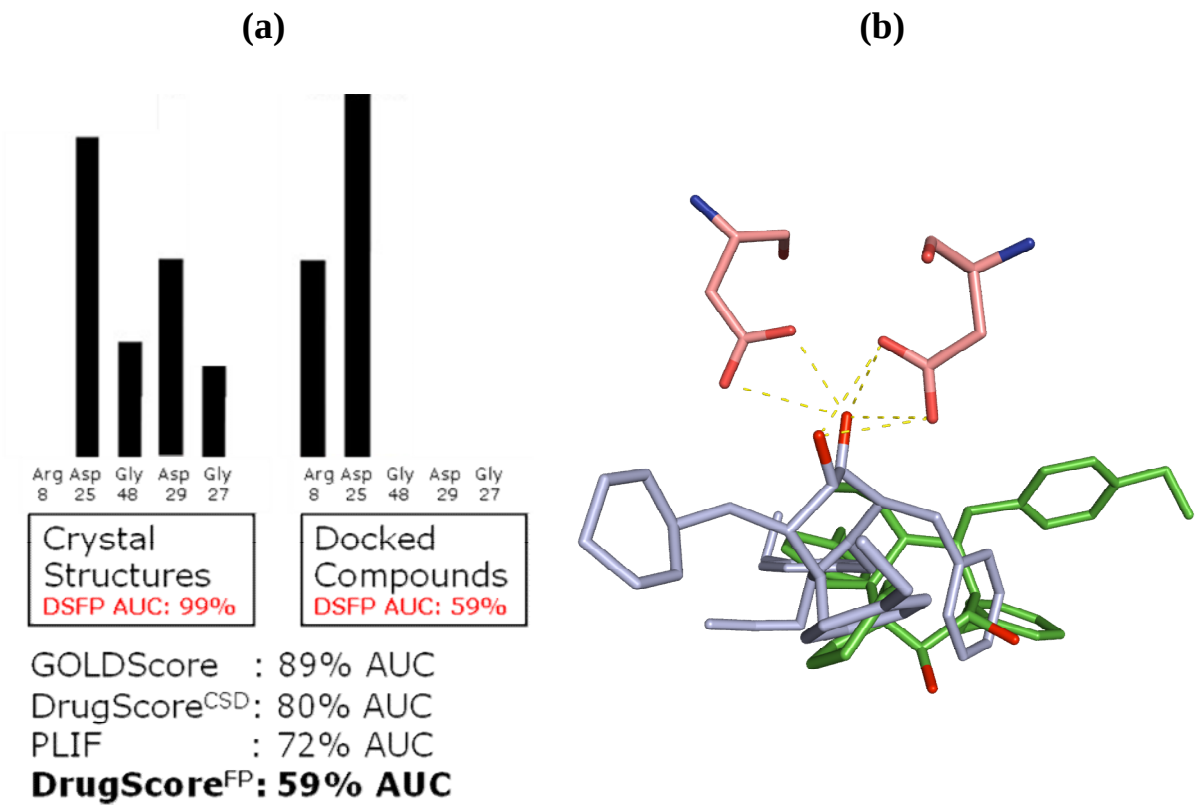


Figure 7: (a) Histogram showing the similarity of the interaction profiles between crystal structures and the corresponding docking poses of compounds interacting with the core residues of HIV-1 protease. Below, the recovery rate of docking poses of 20 active compounds seeded into 70 inactives is given for the scoring functions DrugScore^{FP}, DrugScore^{CSD}, PLIF and GOLDScore. (b) Crystal structure (light blue) and docked geometry (green, 8.15Å rmsd) of a HIV-1 protease inhibitor. The computed binding mode shows interactions formed to the protein not observed experimentally. In such a case, a successful virtual screening using DrugScore^{FP} will not be successful as expected. The docked compound is scored on similarity rank 87 of 90 using DrugScore^{FP} and can be surprisingly found on rank 8 of 90 using GOLDScore.

If near-native interactions are computed by docking programs, DrugScore^{FP} reaches high AUC values and performs better than the other scoring functions in recovering known binders at the top of the similarity list.

The docking solutions of the 20 native HIV-1 protease inhibitors had a mean rmsd value of 3.8 Å and a standard deviation of 2.4 Å. Here, DrugScore^{FP} shows minor performance with an AUC value of only 59.0 % whereas GOLDScore performs best with an AUC value of 89.0 %. Taking a closer look into the interaction profiles of the 20 docking solutions and the native geometries (Figure 7), strong differences among both cases can be observed, denoting that the docking program encounters interactions not observed in the native structures.

DrugScore^{FP} must rank these solutions low, as no match with the trained fingerprint is given. In fact, it is the question, why GOLDScore performs so well with an AUC value of 89.0 %. The intention of a scoring function should be to retrieve geometries which strongly correlate to the native one. GOLDScore places most of the known binders on top of the scoring list, even though the ranking is based on docking modes significantly deviating from the crystallography observed ones. Thus, the drop of DrugScore^{FP}'s AUC values underlines its reliability to correctly retrieve near-native docking geometries.

Virtual Screening using the Directory of Useful Decoys (DUD). In the case of trypsin DrugScore^{FP} with an AUC of 79.9 % again outperforms DrugScore^{CSD} with 51.4 % and GOLDScore with 52.0 %. The result of the latter two functions nearly matches a random enrichment. In the case of HIV-1 protease DrugScore^{CSD} with an AUC of 77.0 % performs better than DrugScore^{FP} with 71.3 % and GOLDScore with 65.6 %. The active compounds for HIV-1 protease have a higher molecular weight compared to the trypsin binders. Thus, the influence of key interactions may be overwhelmed by contributions generated by remaining parts of the molecules. In such cases we suggest an optional weighting of the fingerprints. However, our results state that a high number of reference structures is necessary to benefit from the weighting (Table 4). Hence it appears that DrugScore^{FP} is more suitable for small-sized molecules, e.g. fragment-like structures, whereas for ligands of the typical size of drug molecules a more diverse data set for training might be important.

DrugScore Fingerprint Clustering. Figure 8 shows the similarity dendrogram resulting from a k-means clustering using DrugScore^{FP} with respect to the HIV-1 protease data set (used in the NCI screening) as input. The fingerprints include the binding information from the native geometry of the known binders together with the docking solutions of the NCI diversity set. Figure 8 indicates that all 22 known binders of HIV-1 protease can be grouped into one lead cluster; the computed similarity dendrogram from the trypsin data set (dendrogram not shown here) indicates that clustering the trypsin fingerprint can group all 56 active compounds out of 1800 assumed inactive compounds into two clusters.

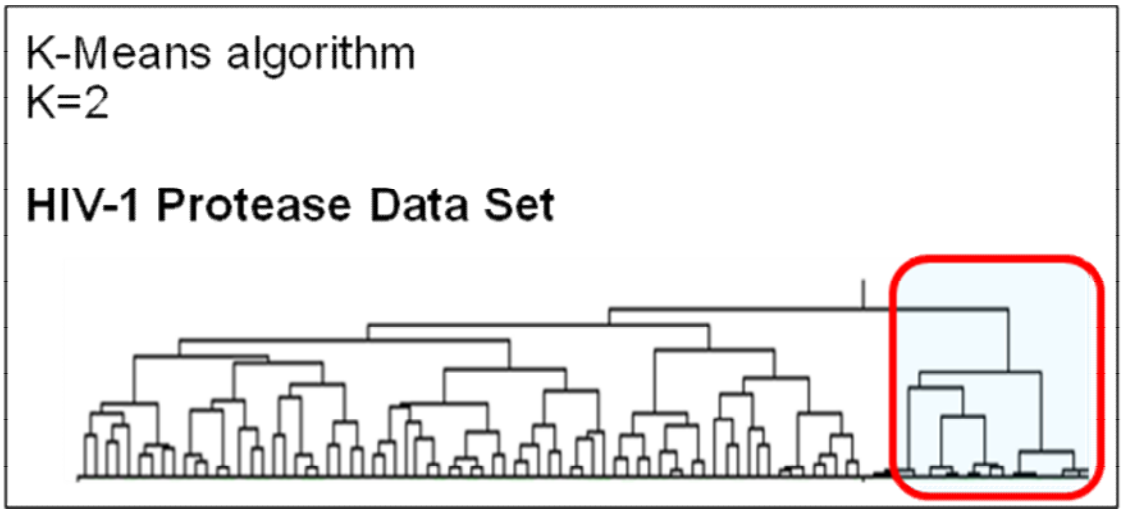


Figure 8: Clustering dendrogram of DrugScore Fingerprints derived from the docking poses of 70 NCI diversity set compounds and the crystal structure of 22 known HIV-1 protease binders. The red square marks the lead cluster covering all 22 active ligands.

Selection of Trypsin Testing Candidates. To discover putative novel binders, we focused on those hits of the NCI diversity set that were found in a cluster together with known binders. By visual inspection, four compounds showing deviating scaffolds from all known binders were selected and tested for trypsin inhibition. We selected particularly compounds with low molecular weight to assess DrugScore^{FP}'s performance with respect to fragment-size ligands in a virtual screening run. Two compounds were found to inhibit trypsin in the low millimolar range (Figure 9). Compound **1** (DrugScore^{FP} rank 13) did not show any binding towards trypsin and **2** (DrugScore^{FP} rank 1) was unfortunately incompatible with our assay conditions. The ligand efficiencies (LE) [48] of the two active compounds **3** (DrugScore^{FP} rank 3) and **4** (DrugScore^{FP} rank 2) are quite satisfying, **3** of LE=1.554 kJ/mol (0.37 kcal/mol) and **4** of LE=1.848 kJ/mol (0.44 kcal/mol).

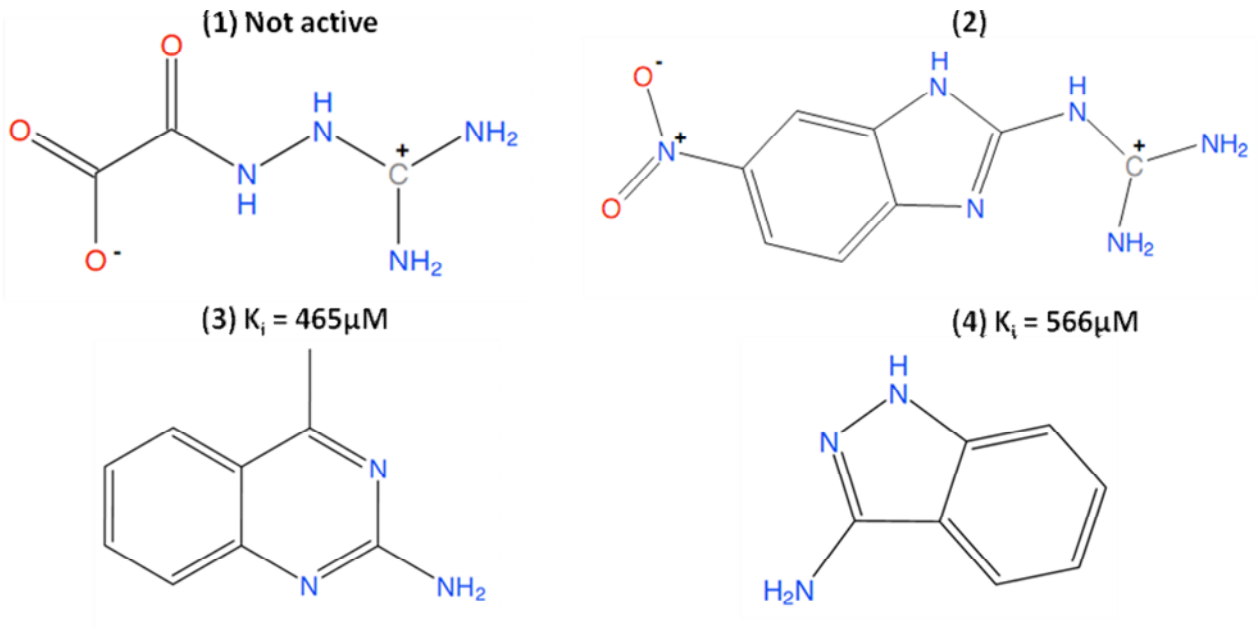


Figure 9: Four fragment-like compounds were selected for testing kinetically the inhibition against bovine trypsin in our laboratory. Affinity data are given as K_i-values.

Virtual Screening using the Fragment-like Subset of the ZINC Database. We decided to perform a similar validation scenario using TGT and thermolysin, which are currently studied in our laboratory. We performed a visual inspection of the top ranked fragments of the ZINC subset showing the highest similarity scores with the fingerprint models. Figure 10 shows four fragments selected for experimental testing in a kinetic TGT enzyme assay. Three compounds showed inhibition constants in the low micromolar range. The ligand efficiencies show very satisfactory values: (1) LE=2.22 kJ/mol (0.53 kcal/mol) (2) LE=2.94 kJ/mol (0.7 kcal/mol) and (3) LE=2.73 kJ/mol (0.65 kcal/mol). Compound (4) is the well-known anti-tuberculosis drug Isoniazid. Unfortunately, it does not inhibit TGT.

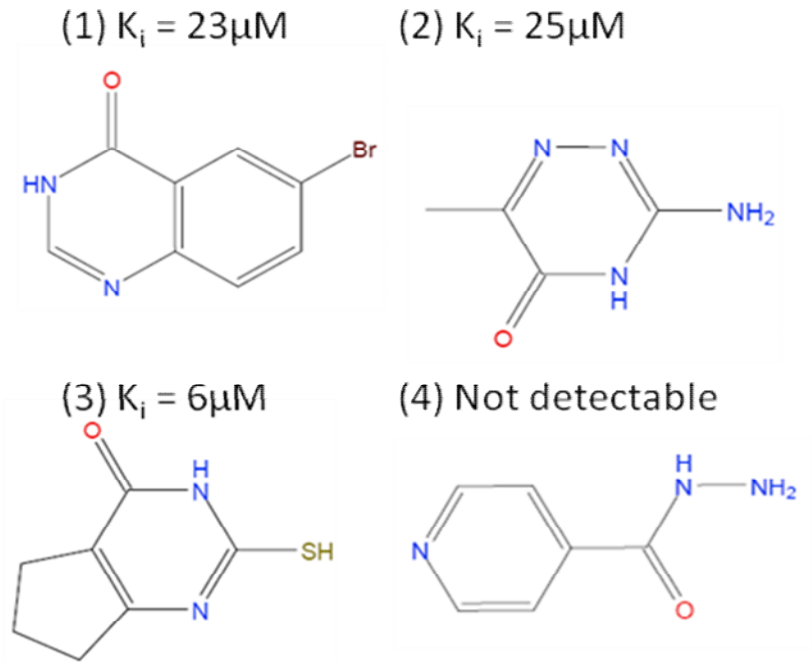


Figure 10: Four fragment-like compounds were selected for testing kinetically the inhibition against TGT in our laboratory. Affinity data are given as K_i-values.

After visual inspection of the 20 top ranked fragments for thermolysin, we selected five compounds for experimental affinity testing. Common features of all compounds are a metal binding group to address the zinc ion and a hydrophobic moiety to putatively fill the hydrophobic S_1'- specificity pocket of thermolysin. Compound (1) shows a K_i–value of 1.6 µM and a ligand efficiency of LE=2.22 kJ/mol (0.53 kcal/mol) (Figure 11); its predicted binding mode is shown in Figure 12. The thiol group of this ligand possibly coordinates the zinc ion. In addition, the positively charged nitrogen can form hydrogen bonds to Asn 112 and the catalytic Glu 143. The hydrophobic indole ring fits nicely into the hydrophobic S1'/S2'-pocket. The binding affinity of compound (2), *N*-benzoyl-β-alanine, was too weak to determine an inhibition constant. Yet we succeeded in solving the three-dimensional crystal structure of this fragment in complex with thermolysin at a resolution of 1.3Å. Important interactions of *N*-benzoyl-β-alanine with thermolysin are the complexation of the zinc ion and the occupation the hydrophobic S1`-pocket. The carboxylate group of the inhibitor coordinates to the zinc ion in a bidentate mode (1.97Å, 2.63Å).

(1) K_i = 1.6µM **(2)**

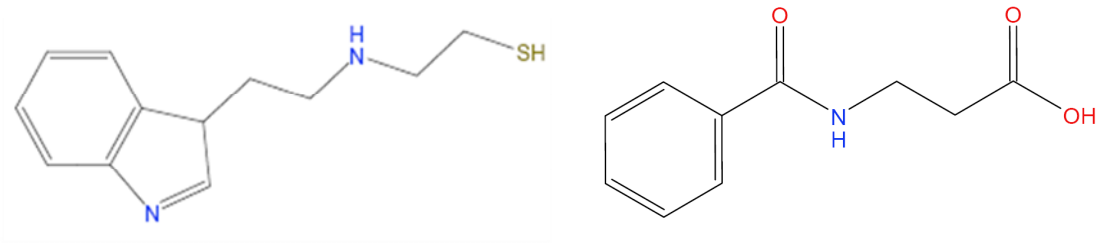


Figure 11: These fragment-like compounds were found to inhibit thermolysin in our virtual screening. Affinity data for (1) is given as K_i-value.

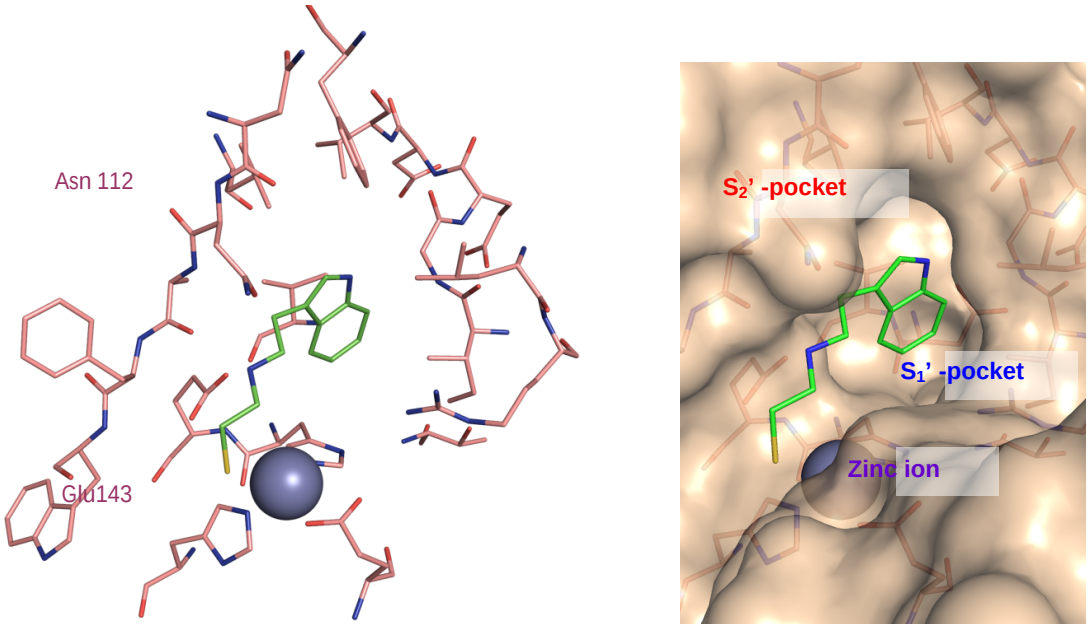


Figure 12: Thermolysin: Possible binding mode of fragment (1).

As known from other thermolysin inhibitors the metal coordinating group seems to mimic a catalytic transition state [49, 50] and forms an additional hydrogen bond to Glu143 (2.69Å). The nitrogen of the amide group is hydrogen-bonded to asparagine 112 whereas the carbonyl functionality is able to make two hydrogen-bonds to the guanidinium nitrogens of arginine 203 (2.96Å, 3.08Å). Van der Waals interactions can be observed between the aromatic ring of the fragment and the mainly hydrophobic S1`-pocket of thermolysin. The amide group is slightly twisted out of plane of the aromatic ring system.

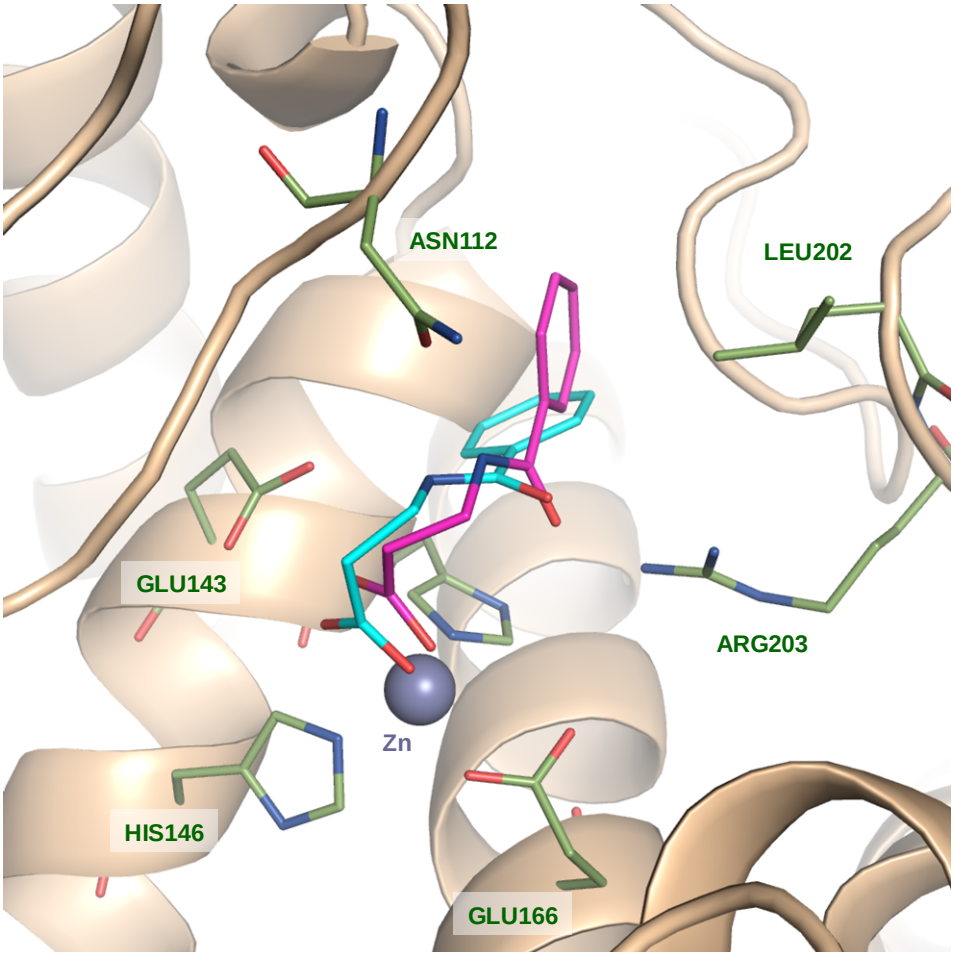


Figure 13: The crystal structure of *N*-benzoyl-β-alanine (compound (2)) with thermolysin is shown in magenta and the computed binding mode is shown in cyan.

The crystal structure of compound (2) in complex with thermolysin confirms the binding mode predicted by AutoDock (Figure 13). Docking solution and native geometry of *N*-benzoyl-β-alanine match convincingly well. As predicted, the inhibitor coordinates with its carboxylate group to the zinc ion and forms hydrogen bonds to Asn112 and Arg203. The only difference between predicted and native geometry is the position of the benzyl-group in the S1`-pocket of thermolysin. Here, the docking solution suggests a binding mode where the ring is perpendicular to the position of the benzyl group observed in the crystal structure. Furthermore, the docking program has predicted the benzyl group to be buried deeper in the S1`-pocket of the enzyme than actually found in the crystal structure. One explanation for this finding could be that in the crystal structure Ile202 adapts slightly to the binding of *N*-benzoyl-β-alanine by moving out of the pocket. During docking, however, the protein was kept rigid and thus docking into the induced-fit adapted geometry was not possible.

2.2.5 Conclusions

This study presents an extension of the well-known scoring function DrugScore^{CSD}, named DrugScore^{FP}. The method can be used for rescoring docking solutions by adding structural information from a user-defined set of protein-ligand complexes resulting in a tailor-made protein-specific scoring function. The new method demonstrates significant improvements in retrieving near-native poses in comparison to 14 established scoring functions. Using four different protein structures as test examples, a large data set of non-binders is reliably discriminated from a set of known binders indicated by a superior AUC value compared to the other scoring schemes. Prerequisite to exploit the power of this fingerprint approach are reliable docking geometries exhibiting near-native interaction profiles. Actually for this scenario, DrugScore^{FP} outperforms other scoring schemes as it only ranks geometries as favorable that agree with experimentally observed interaction patterns. Other scoring functions, summed up to an overall score, might be mislead due to alike interactions ranked similarly, however never observed in this spatial arrangement for the protein under consideration. DrugScore^{FP} is robust with respect to cross-validations and the fingerprint data can be used as input for standard clustering algorithms. The resulting similarity dendrograms can easily be used to detect docking poses which fall next to one cluster populated by known binders. In total we identified six fragment-sized molecules as potential binders of the three investigated drug targets in this study. Furthermore, for one of these fragments, a crystal structure of thermolysin in complex with the top ranked *N*-benzoyl-β-alanine could be solved at a resolution of 1.3Å. These results appear promising and the found binders now serve for further optimization studies with respect to their side chains.

2.2.5 Experimental Section

Experimental Validation - Trypsin. Kinetic inhibition data of bovine trypsin (Sigma-Aldrich Chemie GmbH, Germany) was determined photometrically at 405 nM using the chromogenic substrate Pefachrom tPa (LoxoGmbH, Dossenheim, Germany) according to the protocols described by Stürzebecher *et al.* [51] applying the following conditions: 50mM Tris/HCl, pH 8.0, 154mM NaCl, 5 % DMSO and addition of 10mM $CaCl_2$ at 25 °C using different concentrations of substrate and inhibitor. K_i-values were determined as described by Dixon [52].

Experimental Validation – Thermolysin. Kinetic inhibition data of thermolysin (Calbiochem) was determined fluorimetrically using the quenched fluorescent dipeptide Dabcyl-Ser-Phe-EDANS (2-N-(4-[4'-N',N'-Dimethylamino)phenylazo]-benzoyl-L-serinyl-L-phenylalanylamido)-N''-ethylaminonaphthalene-5-sulfonic acid) (N-Zyme BioTech GmbH) at an excitation wavelength of λ_{ex}=336 nm and an emission wavelength of λ_{em}= 525 nm [53]. The following conditions were used: 100 mM Tris/HCl, pH 7.5, 2 mM $CaCl_2$, 4 % DMSO at 25 °C and different inhibitor concentrations. K_i-values were determined using the program GraFit [54].

Trapping Experiment - TGT. In order to characterize the kinetic properties of the discovered ligands 5 µM *Z. mobilis* TGT, 100 µM *E. coli* $tRNA^{Tyr}$, and 10 mM of the putative hit (dissolved in DMSO) were incubated for 1 h at 25 °C. Afterwards, 10 µl SDS loading buffer were added and incubated for an additional hour at 25 °C. 5 µl of each sample were loaded onto a 15 % SDS gel and stained with 0.1 % Comassie brilliant blue.

Inhibition Constant Determination - TGT. For the determination of the kinetic inhibition constants an assay solution of 150 nM *Z. mobilis* TGT was used and a protocol established by Grädler *et al.* [55] and Meyer *et al.* [56] was applied.

Crystallisation. Native thermolysin (purchased from Calbiochem) crystals were prepared as described by Holmes and Matthews [57] with slight modifications. Thermolysin was dissolved in 0.05M Tris/HCl buffer (pH 7.3), containing DMSO (50% (v/v)) and 1.9M caesium chloride. The final protein concentration was 4.0mM. Crystals were grown at 18°C by the sitting drop vapor diffusion method using water

as reservoir solution. Protein-ligand complex crystals were obtained via soaking crystals in a solution of 0.1M Tris/HCl, 2mM calcium chloride and 10% DMSO. Ligand concentration was 50mM for fragment soaking. Crystals were soaked for 24h before freezing. For cryo-protection crystals were briefly soaked in cryobuffer (10mM Tris/HCl, pH7.3, 10mM $CaCl_2$, 5% DMSO, 20% glycerol).

Data collection, phasing and refinement. Data for *N*-benzoyl-ß-alanine was determined at the synchrotron BESSYII in Berlin on PSF beamline 14.2 equipped with a MAR-CCD detector. Data were processed and scaled with Denzo and Scalepack as implemented in HKL2000 [58]. The coordinates of thermolysin in complex with an *N*-carboxymethyl dipeptide inhibitor (PDB code: 1TMN) were used after removal of ligand, metal ions and water atoms for initial rigid-body refinement of the protein atoms followed by repeated cycles of conjugate gradient energy minimisation, simulated annealing and B-factor refinement using the CNS programme package [59]. The structure refinement was continued with SHELXL-97 [60], for each refinement step at least 10 cycles of conjugate gradient minimisation were performed with restraints on bond distances, angles, and B-values (Table 5). Intermittent cycles of model building were done with the programme COOT [61]. The coordinates have been deposited in the PDB (http://www.rcsb.org/pdb/) with access codes 3FGD.

Table 5: Data collection and refinement statistics for TLN in complex with *N*-benzoyl-β-alanine

Crystal data	
pdb code	**3FGD**
A. *Data collection and processing*	
No. crystals used	1
Wavelength (Å)	0.9184
Space group	$P6_122$
Unit cell parameters	
a,b (Å)	92.6
c (Å)	128.7
B. *Diffraction data*	
Resolution range (Å)	30–1.33
Unique reflections	72415 (3214)*
$R(I)_{sym}$ (%)‡	5.4 (41.5)*
Completeness (%)	96.4 (87.0)*
Redundancy	5.3 (3.2)*
I/σ(I)	26.3 (2.3)*
C. *Refinement*	
Resolution range (Å)	10-1.33
Reflections used in refinement	72226
Final *R* values	
R_{free} (F_o; $F_o > 4\sigma F_o$) $^{\$}$	18.9 (17.1)*
R_{work} (F_o; $F_o > 4\sigma F_o$)°	14.8 (13.4)*
No. of atoms (non-hydrogen)	
Protein atoms	2459
Water molecules	288
Ligand atoms	14
RMSD, angle (deg.)	2.0
RMSD, bond (Å)	0.013
Ramachandran plot‖	
Most favoured regions (%)	87.8
Additionally allowed regions (%)	11.1
Generously allowed regions (%)	0.7

Disallowed regions (%)	0.4
Mean *B*-factors (Å^2)	
Protein atoms	12.5
Metals	11.1
Water molecules	26.5
Ligand atoms	16.1

Table 5 *Values in parenthesis are statistics for the highest resolution shell. ‡$R(I)_{sym} = [\Sigma_h \Sigma_i |I_i(h) - \langle I(h) \rangle| / \Sigma_h \Sigma_i I_i(h)] x100$, where $\langle I(h) \rangle$ is the mean of the $I(h)$ observation of reflection h. ° $R_{work} = \Sigma_{hkl} |F_o - F_c| / \Sigma_{hkl} |F_o|$, $ R_{free} was calculated as for R_{work} but on 5% of the data excluded from refinement. ||From Procheck.

3 Optimiertes Design kombinatorischer Verbindungsbibliotheken unter Verwendung Genetischer Algorithmen

Die Identifizierung bevorzugter Seitenketten niedermolekularer Grundgerüste für einen betrachteten Rezeptor ist eng an die Erkennung energetischer Minima auf einer Energiehyperfläche gebunden. Ein numerischer Vektor bestehend aus einem Satz von Seitenketten für ein Grundgerüst und durch eine Dockingsimulation errechnete kartesische Koordinaten erzeugt jeweils einen Energiewert für eine niedermolekulare Verbindung. Sinnbildlich ist ein Energiewert ein Punkt auf einer Gebirgslandschaft-ähnlichen Energiehyperfläche, welche im Sinne hoher und niedriger Energiewerte Gebirge und Täler aufweist. Für ein wirkstoffähnliches Grundgerüst mit *N* Substitutionspunkten können

$$\prod_{i=1}^{N} N_i \qquad \textbf{(9)}$$

Produkte enumeriert werden. Aus dieser Grundgesamtheit können wiederum $\binom{n}{k}$ viele Teilmengen von Molekülen zusammengestellt werden, wobei dann für jedes aus *M* Atomen bestehende System *3M* kartesische oder *3M-6* interne Koordinaten benötigt, um einen Punkt auf der Energiehyperfläche darzustellen. Bei der Erzeugung von Kleinmolekülen und für deren relevante Geometrien in der Bindetasche ihres Rezeptors sind Minima der Gesamtenergie von Protein und Ligand von besonderem Interesse. Sie beschreiben einen energetisch bevorzugten Zustand des Systems, wobei meist ein Pass mit hohem Energiegehalt zu überwinden ist, um von einem zum anderen Minimum zu gelangen. Für ein Molekül kann es in Abhängigkeit von dessen chemischer Dekoration und Geometrie eine Vielzahl solcher Minima auf der Energieoberfläche geben, wobei es das Ziel ist, das globale Minimum aufzufinden. Zur computergestützten Identifizierung dieser Minima kann die Energielandschaft systematisch abgesucht werden. Dies ist aber aufgrund der Vielzahl an Parametern für je eine untersuchte Struktur ein kombinatorisches Problem. Ein niedermolekulares Grundgerüst mit drei Substitutionspunkten und 1000 möglichen Seitenketten pro Anknüpfungspunkt umfasst bereits einen kombinatorischen Raum von 1.000.000.000 möglichen Produkten, ohne weitere

konformative Freiheitsgrade zu berücksichtigen. Der denkbar einfachste Ansatz zur Vermeidung einer kombinatorischen Explosion wäre die zufällige Enumerierung einer Teilmenge der möglichen Seitendekorationen eines molekularen Grundgerüstes; eine Einpassung in die Bindetasche des ausgewählten Rezeptors wäre dann nur für diese Moleküle nötig. Zwar würde es sich hierbei um eine beeindruckend einfache Methode handeln, aber die Wahrscheinlichkeit ein gutes Minimum zu treffen wäre abhängig von der Anzahl der zufällig enumerierten Moleküle und könnte sehr gering sein. Bessere Möglichkeiten bieten Methoden, welche für eine gegebene Molekülstruktur benachbarte Koordinaten mithilfe von Gradienten berechnen, um günstigere Energiezustände zu identifizieren. Kombiniert man diese Vorgehensweise mit einer zufälligen oder intelligenten Auswahl an Startpunkten auf der Energiehyperfläche, ergibt sich eine globale Suchstrategie.

In den folgenden Kapiteln werden zunächst gängige Verfahren zur Repräsentation von Molekülen im Computer vorgestellt (Kapitel 3.1). Im Anschluss werden Methoden erläutert, welche eine Berechnung der dreidimensionalen Struktur erlauben (Kapitel 3.2) gefolgt von Kapitel 3.3, welches sich mit der Erzeugung von Ligand-Geometrien in Protein-Bindetaschen beschäftigt. Kapitel 3.4 ist der Lösung von Optimierungsproblemen gewidmet. Im Anschluss wird das in dieser Arbeit entwickelte Programm GARLig ausführlich beschrieben und diskutiert (Kapitel 3.5).

3.1 Literaturbekannte Ansätze zur Repräsentation von Molekülen im Computer

In der Computerchemie werden Repräsentationen von Molekülen benötigt, die Informationen über die Atome und deren Konnektivitäten beinhalten. Solche festen Regelsätze wie etwa Atomkonnektivitäten zu einer für den Computer einheitlichen Struktur verdichtet, nennt man Dateiformat. Nun ist es beispielsweise möglich, ein Molekül oder eine chemische Substruktur in einer Datenbank zu suchen.

2D- oder 3D-Molekülformate können desweiteren zur Visualisierung genutzt werden. Darüber hinaus ist aber noch eine Vielzahl anderer Anwendungen möglich, etwa die Berechnung diverser molekularer Deskriptoren für physikochemische Eigenschaften.

Aktuell gibt es eine Vielzahl an Dateiformaten für biochemisch relevante Moleküle. Dabei wird hauptsächlich Wert auf den verwendeten chemischen Informationsgehalt gelegt. Ein Format der einfachsten Stufe beinhaltet lediglich eine Auflistung der Atome. Wird der Informationsgehalt erweitert, kommen noch Informationen über Konnektivitäten, Atomtypen und Stereochemie hinzu. Ein hohes Maß an Information beinhalten 3D-Molekülformate, welche räumliche bzw. konformative Aspekte abdecken.

In den folgenden Kapiteln werden jene Molekülformate behandelt, die eine Relevanz für diese Arbeit besitzen, wobei der strukturelle Aufbau und die verwendete chemische Information des Formats im Vordergrund stehen sollen. Dabei werden neben topologischen 2D-Strukturrepräsentationen wie SMILES (Kapitel 3.1.1) auch topographische 3D-Formate wie Sybyl mol2 (Kapitel 3.1.2) und PDBQT (Kapitel 3.1.3) vorgestellt.

3.1.1 Das SMILES Dateiformat

Eine weit verbreitete Methode zur Beschreibung von chemischen Verbindungen ist die Kodierung der Molekülinformation in einer linearen Notation. Das 1986 von Weininger veröffentlichte lineare Notierungssystem SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) ist das bekannteste System zur Kodierung der zweidimensionalen Molekülstruktur in einer linearen Zeichenabfolge [62].
Die Terminologie des SMILES Konzeptes definiert die 2D-Strukturformel als Molekülstruktur. Hierbei wird die chemische Verbindung durch eine lineare Abfolge von Buchstaben, Zahlen und Symbolen dargestellt. Wasserstoffatome müssen in der SMILES Notation nicht explizit mit angegeben werden. Nachfolgend sollen die grundlegenden Regeln dieser Notation kurz wiedergegeben werden:

1. Die Kodierung der Atome erfolgt analog zur IUPAC-Konvention.
2. Weitere Spezifikationen wie Ladung (Symbol + für positive, Symbol – für negative Ladung) und Wasserstoffe erfolgt immer innerhalb eckiger Klammern. Durch einen Multiplikator vor oder nach dem Ladungssymbol können entsprechende Ladungszustände beschrieben werden.
3. Die Symbole -, =, # und : werden zur Darstellung einer Einfach-, Zweifach-, Dreifach- und aromatischen Bindung zwischen zwei Atomen verwendet, wobei Einfachbindungen und aromatische Verknüpfungen nicht explizit angegeben werden müssen. Dies geschieht bereits automatisch durch die aufeinanderfolgenden Atomsymbole.
4. Eine Seitenkette bzw. Verzweigung des Moleküls erfolgt entsprechend obiger Konvention, wird aber durch runde Klammern als solche gekennzeichnet.
5. Bei der Kodierung von zyklischen Systemen (Ringe) wird eine formale Spaltung einer Einfachbindung oder aromatischen Verknüpfung vorgenommen, um den Ring somit in eine lineare Abfolge von Atomen zu bringen. Start- und endständiges Atom eines Ringes wird mit einer Zahl zur eindeutigen Identifizierung des gerade betrachteten zyklischen Systems versehen. Die Kodierung aller Atome folgt den obigen Überlegungen.
6. Bei aromatischen Systemen werden bei den Atomsymbolen Kleinbuchstaben verwendet, ansonsten Großbuchstaben.

7. Die Kennzeichnung bzw. Trennung bei nicht-kovalenten Bindungen erfolgt durch ein Punktzeichen.

Die SMILES Sprache beinhaltet keine Regeln zur hierarchischen Kodierungsreihenfolge der Atome. Dies hat zur Folge, dass eine Molekülstruktur meist durch eine Vielzahl an unterschiedlichen SMILES-Zeichenketten dargestellt werden kann. Eine Kanonisierung kann durch den CANGEN Algorithmus erfolgen [63], ist jedoch nicht Bestandteil einer Betrachtung in der vorliegenden Arbeit.

Der Einsatz der SMILES Notation ist bei dem im Folgenden beschriebenen Programm GARLig aufgrund der einfachen Handhabung von Zeichenketten gewählt worden. Die Manipulation eines in SMILES kodierten Moleküls ist deutlich einfacher als in anderen Dateiformaten. Im Anschluss können durch 3D-Strukturgeneratoren SMILES Dateien leicht in weitere Molekülformate übersetzt werden.

3.1.2 Das Sybyl mol2 Dateiformat

Das Sybyl mol2-Dateiformat ist ein in der Computer-Chemie weitverbreitetes ASCII-Format. Die Textdateien beinhalten Informationen wie Atomnamen, Atomtypen, 3D-Koordinaten, Partialladungen und Bindungstypen. Optional können Informationen wie etwa molekulare Deskriptoren angegeben werden. Sybyl mol2-Dateien zeichnen sich durch eine Klassifikation der Atome eines Moleküls in 17 Atomtypen aus (Tabelle 6), welche den Namen des tatsächlichen Atomsymbols tragen, denen durch ein Punktsymbol getrennt ein Suffix mit einer weiterführenden Typisierung des Elementes folgt. Die Verwendung der Atom-Typisierung und eines Valenzmodells gestatten den Verzicht auf eine explizite Angabe von Protonen.

Bei der Typisierung von Aminosäure-Atomen eines Proteins kann es sowohl ein- als auch mehrdeutige Zuordnungen geben. Da Protonen in der Röntgenkristallographie

Tabelle 6: Die Benennung der verwendeten Atomtypen folgt der Sybyl-Notation (SYBYL).

Bedeutung	**Atomtyp**
sp^3-hybridisierter Kohlenstoff	C.3
sp^2- und sp^1-hybridisierter Kohlenstoff	C.2 (C.1)
Kohlenstoff in aromatischen Ringen	C.ar
Kohlenstoff in Amidino- und Guanidinogruppen	C.cat
sp^3-hybridisierter Stickstoff	N.3 (N.4)
Stickstoff in aromatischen Ringen und sp^2-hybridisierter Stickstoff	N.ar (N.2)
Stickstoff in Amidbindungen	N.am
Stickstoff in Amidino- und Guanidinogruppen	N.pl3
sp^3-hybridisierter Sauerstoff	O.3
sp^2-hybridisierter Sauerstoff	O.2
Sauerstoff in Carboxylgruppen	O.co2
sp^3- und sp^2-hybridisierter Schwefel	S.3 (S.2)
sp^3-hybridisierter Phosphor	P.3
Fluor	F
Chlor	Cl
Brom	Br
Calcium, Zink, Eisen, Nickel	Met

mit Ausnahme von Strukturen mit einer Auflösung < 1 Å nicht zu erkennen sind, liegen Informationen über Protonierungszustände von Protein und Ligand selten vor. Um diese Problematik zu umgehen, werden Annahmen hinsichtlich Protonierungszuständen und Ladungen vorab getroffen, wie etwa, dass Asparaginsäure oder Glutaminsäure i.d.R. die geladene Form annehmen und der terminale Sauerstoff den Atomtyp O.co2 zugewiesen bekommen soll. Für Liganden aus einer PDB-Datei ist die Zuweisung von Atomtypen immer schwierig, denn in den Heteroatom-Einträgen einer solchen Datei sind keine Typ-Informationen für Ligandatome gegeben, sodass sich Algorithmen lediglich auf Atomabstände und

Konnektivitäten stützen, um so mit einer festen Regelbasis die Zuweisung vorzunehmen. Zusätzliche Informationen könnten hier durch eine Bestimmung von pKa-Werten seitens des Liganden erfasst werden [64].

3.1.3 Das PDBQT Dateiformat

In der neuesten Version des Docking-Programms AutoDock 4 [65] haben nun sowohl der Ligand als auch das Protein das gleiche Dateiformat. Strukturell ist dieses Format an die PDB-Spezifikation gebunden (wwpdb.org/documentation/format23/). Eine PDBQT-Datei ist eine ASCII-Datei und beinhaltet Gasteiger PEOE-Partialladungen [66] nebst den in AutoDock 4 definierten 15 Atomtypen. Bei der expliziten Angabe von Wasserstoffen müssen lediglich polare H-Atome berücksichtigt werden. Wie auch in der vorherigen Version AutoDock 3 [67] wird bei Liganden ein Torsionsbaum erzeugt, welcher die niedermolekulare Verbindung in rigide und flexible Bereiche unterteilt. Bei einem Torsionsbaum gibt es immer eine Wurzel und keine, ein oder mehrere Zweige, wobei jeder Zweig eine rotierbare Bindung repräsentiert:

- Mit einem ROOT Eintrag beginnt der rigide Bereich des Liganden, von dem keine, ein oder mehrere rotierbare Bindungen ausgehen können
- Der ROOT Block endet mit der ENDROOT-Kennzeichnung
- Atome, die durch frei drehbare Bindungen bewegt werden können, befinden sich in einer BRANCH/ENDBRANCH-Umgebung
- Die letzte Zeile einer PDBQT-Datei enthält noch eine Integerzahl, welcher die Gesamtzahl der Torsionsfreiheitsgrade des Liganden beziffert.

3.2 *Corina – ein Ansatz zur Erzeugung 3-dimensionaler Geometrien niedermolekularer Verbindungen*

Werden bei Chemie-informatorischen Analysen räumliche Betrachtungen, wie etwa die Bindungsgeometrie eines Liganden in der Bindetasche eines Rezeptors, miteinbezogen, sind im Regelfall dreidimensionale Strukturdaten erforderlich. Notwendige 3D-Molekülkonformationen können dabei mit dem Strukturgenerator CORINA (**Co**o**r**d**ina**tes) berechnet werden [68]. Als Grundlage zur Erzeugung einer 3D-Struktur dienen dem Programm standardisierte Daten wie Bindungslängen, Bindungswinkel und Ringgeometrien sowie ein Satz von Regeln, der Erfahrungswerte aus Kraftfeldrechnungen, kristallographischen Daten und geometrischen Überlegungen beinhaltet. Durch die große Menge an implementierten Regeln ist das Programm prinzipiell nicht auf eine maximale Anzahl an Atomen beschränkt und kann die strukturelle Vielfalt der organischen Chemie sowie einen Teil der metallorganischen Komplexe behandeln.

Zur Berechnung der dreidimensionalen Struktur verwendet CORINA eine Konnektivitätstabelle der entsprechenden Verbindung. Die Eingabe der Bindungsverhältnisse erfolgt über chemische Austauschformate wie beispielsweise Sybyl mol2 oder SDF.

Die Generierung erfolgt durch eine Reihe von Einzelschritten: nachdem im ersten Schritt alle Bindungslängen und Bindungswinkel basierend auf dem Atomtyp und der Hybridisierung des betrachteten Atoms mit standardisierten Werten belegt wurden, erfolgt zur weiteren Berechnung die Aufspaltung des Moleküls in zyklische und azyklische Teilsysteme. Die cyclischen Systeme werden je nach Größe und Eigenschaften unterschiedlich behandelt. Während kleinere Ringsysteme (drei bis acht Atome) aufgrund ihres eingeschränkten konformativen Raumes durch vordefinierte Torsionstabellen beschrieben werden, wird zur Ermittlung großer Ringsysteme auf regelbasierte Methoden zurückgegriffen. Anschließend erfolgt eine geometrische Verfeinerung der Ringe durch ein Pseudo-Kraftfeld, welches mehr auf geometrischen Betrachtungen als auf physikalischen Funktionen basiert. Azyklische Molekülteile werden ebenfalls anhand einer Torsionstabelle analysiert, welche mit der CSD (Cambridge Structural Database) entnommenem kristallographischen Wissen unter Einbeziehung von etwa 900 Regeln entwickelt wurde. Sollten sich

hierbei mehrere räumliche Möglichkeiten des betrachteten Teilsystems ergeben, werden Geometrien bevorzugt, bei denen repulsive Wechselwirkungen am unwahrscheinlichsten sind. In einem abschließenden Schritt werden zyklische und azyklische Fragmente kombiniert, wobei das System hinsichtlich möglicher Atomüberlagerungen oder zu kurzer Atomabstände überprüft wird. Mögliche Konflikte werden dabei durch eine eingeschränkte Konformationsanalyse gelöst, um etwa unpassende weitreichende Wechselwirkungen zu beseitigen. Zielfunktion dieser Analyse ist eine Kombination aus dem 12-6-Lennard-Jones Potential für nichtbindende Wechselwirkungen und dem bereits beschriebenen Torsionsenergieterm.

3.3 Literaturbekannte Ansätze zur Vorhersage von Ligand-Geometrien in Protein-Bindetaschen

Dieses Kapitel gibt einen Überblick über gängige Verfahren zur Modellierung von Ligand-Geometrien in Protein-Bindetaschen.

Molekulares Docking dient der Bestimmung einer relevanten Geometrie niedermolekularer Verbindungen in der Bindetasche eines Rezeptors, wobei eine Maximierung der Interaktionen beider Moleküle angestrebt wird. Docking-Programme versuchen dabei, zwei bekannte Probleme zu lösen: die Erzeugung einer relevanten, experimentell beobachteten Geometrie und eine korrekte Abschätzung der Bindungsaffinität der kleinen Moleküle, wobei auch eine korrekte Reihung der Affinitäten unterschiedlicher Kleinmoleküle für ein Protein angestrebt werden soll. Docking-Methoden der ersten Generation betrachteten sowohl den Rezeptor als auch die niedermolekularen Verbindungen als rigide und bei der Simulation wurden nur sechs Freiheitsgrade (Translation und Rotation) zugelassen. Mit der Verbesserung der Rechnerkapazitäten und der Entwicklung neuer Algorithmen und Kraftfelder wurden auch konformative Freiheitsgrade auf beiden Seiten zugelassen.

Betrachtet man die grundlegende Funktionsweise der Algorithmen, so können zwei Klassen an Dockingtechniken genannt werden. Bei der ersten Gruppe wird der Ligand meist entlang seiner frei drehbaren Bindungen in mehrere Molekül-Fragmente zerlegt. Dann wird das Kernfragment, der so genannte Anker, in die Bindetasche des Rezeptors gedockt und anschließend in einer inkrementellen Prozedur werden alle weiteren Fragmente angefügt. Als etablierte Vertreter können die Programme DOCK und FlexX genannt werden. Diese Verfahren sind bekannt für eine schnelle Berechnung vieler Ligand-Geometrien, jedoch ist hier der initiale Platzierungsschritt des Ankers entscheidend für die eingeschlagene Strategie und damit für die Qualität der Docking-Ergebnisse. Die zweite Gruppe von Algorithmen versucht, den konformativen Raum der Liganden mittels heuristischer Methoden abzusuchen. Hierbei werden etwa Monte Carlo Suchstrategien oder Genetische Algorithmen eingesetzt, wie es bei den in dieser Arbeit eingesetzten Programmen GOLD und AutoDock der Fall ist.

Die in dieser Arbeit verwendeten Docking-Programme GOLD und AutoDock werden in den zwei folgenden Kapiteln in ihrer Funktionsweise näher erläutert.

3.3.1 AutoDock

Mit dem Dockingprogramm AutoDock werden Vorhersagen zur Geometrie flexibler Liganden in Bindetaschen makromolekularer Rezeptoren getroffen, wobei ein Genetischer Algorithmus für diese Berechnungen verwendet wird. Weiterhin ist in der neuesten Version AutoDock 4 eine partielle Flexibilität des Proteins gestattet, jedoch wurde hier ein Limit von 11 Aminosäure-Seitenketten aufgrund der hohen Rechenintensität gesetzt. Da in dieser Arbeit sämtliche Studien mit rigiden Proteinkörpern durchgeführt wurden, hat Proteinflexibilität an dieser Stelle keine weitere Bedeutung.

Zu Beginn einer Docking Simulation erstellt der Genetische Algorithmus eine Zufallspopulation von Liganden benutzerdefinierter Anzahl. Translationswerte bekommen gleichverteilte Zufallswerte aus dem Intervall der Minimum- und Maximum-Werte des den Suchraum begrenzenden Gitters und Rotationswerte werden zufällig zwischen -180° und 180° gewählt. Anschließend wird der Genotyp in den Phänotyp übersetzt. Dies erlaubt eine Energieevaluierung mithilfe des in AutoDock implementierten Programms AutoGrid. Dieses erstellt ein kubisches Gitter mit Gitterpunkten im Abstand von standardmäßig 0.375 Å um die Bindetasche des Zielmoleküls herum. An jedem Punkt des Gitters ist für alle möglichen in AutoDock verwendeten Atomtypen die Wechselwirkungsenergie eines Sondenatoms mit Atomen der Bindetasche verzeichnet. Dies erlaubt eine Berechnung der Fitness in konstanter Zeit. Die Selektion von Ligandkonformationen basiert dann auf der vorher evaluierten Fitness. Durch Cross-Over-Methoden und Mutationen werden neue Individuen generiert und die parentale Generation entfernt. Der letzte Schritt ist nötig, um die Populationsgröße konstant zu halten.

Bei AutoDock wird der GA in Verbindung mit einer adaptiven lokalen Suchstrategie als Lamarck‘scher Genetischer Algorithmus (LGA) verwendet. Mit dieser Hybridmethode wird die Suche nach lokalen Minima effizienter bewältigt. Die Schrittweite bei der lokalen Suche wird durch die Historie der zuletzt evaluierten

Energien bestimmt. Ein Anstieg in den Energien erwirkt eine Verdopplung der Schrittgröße, im umgekehrten Fall halbiert sie sich.

Am Ende einer Berechnung werden Fitness, Zustandsvariablen, Koordinaten der gedockten Konformation und die Abschätzung der freien Bindungskonstanten ausgegeben.

Die in AutoDock verwendete Funktion AutoDock Score gehört zur Klasse der Kraftfeld-basierten Bewertungsfunktionen, welche anhand von 30 Protein-Ligand-Komplexen kalibriert wurde und Terme wie van der Waals- und elektrostatische Wechselwirkungen, Desolvatationsbeiträge, Änderung in den rotatorischen Freiheitsgraden, Konformationsenergien und H-Brücken verwendet.

3.3.2 GOLD

Das Docking-Programm GOLD (Genetic Optimization for Ligand Docking) verwendet ebenfalls einen Genetischen Algorithmus zur Exploration des konformativen Raumes eines untersuchten Liganden und gestattet, diesen während der Berechnung als flexibel zu betrachten. Die Prozedur startet mit einer Anzahl zufällig erzeugter Ligandkonformationen, die anhand zweier ausgewählter Bewertungsfunktionen beurteilt werden. Die Platzierung des Liganden in der Bindetasche erfolgt anhand so genannter ‘fitting points‘, wobei die Optimierung des Genetischen Algorithmus eine Maximierung komplementärer polarer und hydrophober Übereinstimmungspunkte zwischen Protein und Ligand anstrebt. Hierbei können auch multiple Orientierungen polarer Wasserstoffe seitens des Proteins berücksichtigt werden [69].

GOLD gestattet die Verwendung zahlreicher Zusatzbedingungen wie beispielsweise eine Festlegung auf eine während der Docking-Simulation auszubildende Wechselwirkung. Desweiteren ist in der neuesten Version GOLD 4.0 eine partielle Flexibilität des Protein-Grundgerüstes sowie dessen Seitenketten gestattet, wobei hier eine Limitierung auf 10 Aminosäurereste aufgrund der Rechenintensität gesetzt wurde. Wie in Kapitel 3.3.1 bereits erwähnt, sind Docking Studien mit Einbezug der Proteinflexibilität nicht Teil dieser Arbeit.

Die in GOLD implementierte Bewertungsfunktion GOLDScore gehört ebenfalls zur Klasse der Kraftfeld-basierten Bewertungsfunktionen (siehe Kapitel 2.1.1) und beinhaltet Terme für intermolekulare Wasserstoffbrücken, ein 4-8 Dispersionspotential, ein 6-12 intramolekulares Potential für die interne Ligandenenergie und intramolekulare Wasserstoffbrücken.

Die ebenfalls implementierte Bewertungsfunktion ChemScore ist eine empirische Bewertungsfunktion (siehe Kapitel 2.1.2), welche Wasserstoffbrücken-, Metall- und lipophile Wechselwirkungen sowie den Verlust konformativer Entropie in Termen beschreibt, wobei die gewonnenen Koeffizienten aus einem Datensatz von 82 Protein-Ligand Komplexen abgeleitet wurden.

3.4 Literaturbekannte Ansätze zur Lösung von Optimierungsproblemen

Inhaltlich gesehen hat eine Optimierung die Auffindung eines optimalen Parametersatzes einer – meist komplexen – mathematischen Funktion zum Ziel. Formal gesehen kann eine Optimierung als die Lösung des Problems

$$f^* = f(x^*) = \min f(x) \qquad \textbf{(10)}$$

betrachtet werden, wobei f die zu optimierende Funktion darstellt, welche mit f^* den besten Wert am Optimum x^* annimmt. Zur Vereinfachung wird angenommen, dass das Optimum dem Minimum entspricht. Analog lassen sich alle im Folgenden erwähnten Methoden auch auf Maximierungsprobleme anwenden. Überträgt man nun ein Optimierungsproblem auf die in den folgenden Kapiteln behandelte Problematik der Seitenkettendekoration wirkstoffartiger Grundgerüste, so kann f als eine Protein-Ligand Bewertungsfunktion, x als eine Zusammenstellung von möglichen Seitenketten und x^* als die optimale Dekoration des Grundgerüsts angesehen werden, welche dem globalen Minimum entspricht. Die nachfolgenden Kapitel geben einen Überblick zum Einsatz lokaler und globaler Suchmethoden bei der Auffindung des globalen Minimums einer Zielfunktion.

3.4.1 Lokale Suchverfahren

Lokale Suchmethoden sind iterative Verfahren, die eine in ihrem Kurvenverlauf unbekannte, multivariate Zielfunktion lokal abtasten und schrittweise benachbarte Lösungen erzeugen. Die Iteration endet, wenn keine Verbesserung des Funktionswertes erzielt werden kann; dann ist der Algorithmus in ein lokales- oder das globale Minimum konvergiert. Lokale Suchverfahren werden mathematisch gesehen in Gradienten-basierte und Gradienten-freie Verfahren eingeteilt.

Bekannte Vertreter der ersten Gruppe sind Methoden des steilsten Abstiegs, welche oftmals für nichtlineare, multivariate Optimierungsprobleme verwendet werden. Der benötigte Gradient ist hierbei die erste Ableitung an einem Punkt der Energiefunktion.

Der bekannteste Vertreter der Gradienten-freien Verfahren ist der Downhill-Simplex Algorithmus [70]. Der Algorithmus spannt einen *N+1*-dimensionalen Simplex im *N*-dimensionalen Parameterraum auf. Jeder den Simplex aufspannende Punkt entspricht dabei einem durch die Energiefunktion benötigten Parametersatz und zu jedem Punkt kann ein Funktionswert berechnet werden. Unter den *N+1* Punkten wird nun der schlechteste Wert durch einen neu erzeugten Punkt ersetzt, mit der Hoffnung, einen besseren Datenpunkt zu finden. Diese Iteration wird bis zur Konvergenz fortgeführt. Der Algorithmus gilt im Vergleich zu den Gradienten-basierten Verfahren als rechenaufwändig, da mehr Funktionsauswertungen durchgeführt werden müssen. Das Simplex-Verfahren gilt aber zeitgleich als robuster, denn es bietet sich auch für jene Fälle an, bei denen die Gradienten-Bestimmung aufwändig oder unmöglich ist.

Die hier beschriebenen Gruppen lokaler Suchverfahren haben beide zum Ziel, eine im Ganzen unbekannte, meist multivariate Zielfunktion zur Reduktion der Auswertungszeit in lokaler Umgebung eines Datenpunktes abzutasten. Dabei kann es passieren, ein besseres lokales Minimum oder das globale Minimum durch einen zu großen Schritt oder durch zu wenig erzeugte Datenpunkte zu übersehen.

3.4.2 Globale Suchverfahren

Den lokalen Suchmethoden stehen globale Suchverfahren unterstützend beiseite. Die einfachste globale Suche wäre ein rein zufälliges Absuchen der Energiehyperfläche. Die Kopplung einer lokalen Suche an solch einen Zufallsgenerator nennt man Multistart, dessen Ergebnis immer zu einem lokalen Minimum führt [71]. Diese Form der Suche ist jedoch nicht hinreichend intelligent, da per Zufall gleiche Datenpunkte mehrfach erzeugt werden können und kein bisheriges Wissen über günstige und ungünstige Bereiche der Energieoberfläche generiert oder genutzt wird. Hier können Techniken Abhilfe schaffen, die eine Form eines Gedächtnisses aufweisen, um den Suchraum zu beschränken. Als Beispiel seien hier Monte-Carlo-Simulationen erwähnt. Ausgehend von einer zufälligen Erzeugung von Datenpunkten auf der Energiehyperfläche werden zufällige Veränderungen herbeigeführt, die nur bei einer Verbesserung des Funktionswertes akzeptiert werden [72]. Um ein Entkommen aus lokalen Minima zu ermöglichen, können Algorithmen der „Simulierten Abkühlung“ verwendet werden, welche mit einer iterativ immer kleiner werdenden Wahrscheinlichkeit zu Anfang der Prozedur auch schlechtere Funktionswerte als neue Datenpunkte zulassen [73]. Lernfähige Algorithmen, die sich an der biologischen Evolution orientieren und bisher errechnete gute Funktionswerte im Speicher behalten sind u.a. Differentielle Evolution (DE) [74], Partikel-Schwarm-Optimierer (PSO) [75] und Populationsbasiertes Inkrementelles Lernen (PBIL) [76]. In dieser Arbeit sollen vor allem Evolutionäre Strategien (ES) [77] und Genetische Algorithmen (GA) [78] als bekannteste Vertreter dieser Algorithmenklasse erwähnt werden. Sie erzeugen Datenpunkte – bei dieser Klasse von Algorithmen Individuen genannt - wobei gute Individuen Paarungen und Mutationen eingehen können, um daraus gegebenenfalls Punkte niedrigerer Energie zu erzeugen. Viele dieser Methoden speichern zusätzlich noch die besten *n* Lösungen.

Evolutionäre Strategien gehören zu den reell-wertigen Optimierern. Ein Individuum besitzt ein Genom, einen Vektor aus reellen Werten, die einen Datenpunkt auf der Energiehyperfläche und somit dessen Funktionswert erzeugen. Es gibt eine Vielzahl an ES-Varianten. Die einfachste Variante ist die (1+1)-ES, welche zu einem Elternteil entsprechend einer DNS-Replikation eine Kopie erzeugt, wobei diese, mit einer Wahrscheinlichkeit behaftet, mutiert werden kann. Mit Mutation sei eine

zugrundeliegende mathematische Funktion gemeint, die die einzelnen reellwertigen Elemente des Genoms abändert, um daraus einen zur Vorgängergeneration ähnlichen Nachkommen zu erzeugen. Nun kommt die zugrundeliegende Bewertungsfunktion zum Einsatz, welche den Individuen einen Funktionswert zuweist. Analog zur Evolutionstheorie „survival of the fittest“ überlebt jenes Individuum mit niedrigerem Funktionswert. Der Vorgang beginnt von vorne, bis die Generationszyklen erschöpft sind.

Die *(μ+λ)*-ES folgt der Strategie, dass *μ* Eltern *λ* mutierte Nachkommen erzeugen und aus diesen die *μ* besten wiederum überleben. Da jedes Mal sowohl aus der Eltern- als auch der Nachkommen-Generation die *μ* besten Individuen aus dem Pool ausgewählt werden, können sich die Funktionswerte über die Generationen hinweg nicht verschlechtern.

Beim Selektionsdruck handelt es sich um einen Parameter, der die Auswahl an Individuen, die in einer Generation anteilsmäßig zu der Gesamtgröße der erzeugten Nachkommenschaft überleben sollen, reguliert. Der Selektionsdruck wird durch den Quotienten $s = (\mu / \lambda)$ errechnet und liegt zwischen Null und Eins. Bei null liegend ($\lambda > \mu$) handelt es sich um einen hohen, bei Eins liegend $(\lambda < \mu)$ um einen niedrigen Selektionsdruck. Weiterführend können mittels Selektionsdruck auch Populationswellen simuliert werden. Hierbei sei etwa eine beliebige Veränderung von *μ* bei gleichbleibendem *λ*-Wert gemeint; ein aus der Natur als Nischenbildung bekanntes Phänomen. Desweiteren sind durch den Einsatz mathematischer Funktionen auch periodische Veränderungen des Quotienten *s* denkbar.

Aufgrund der Vielzahl an Optimierungsmethoden und deren Parameterisierungsmöglichkeiten soll auf das No-Free-Lunch Theorem verwiesen werden, welches den mathematischen Beweis liefert, dass alle Optimierungsalgorithmen bei einer Problemklasse (bspw. reellwertige Optimierungen) im Mittel die gleiche Lösung liefern [79]. Hiermit sollte der Einsatz von GAs gerechtfertigt sein, sodass nun eine Überleitung zu ihnen erfolgen kann. Genetische Algorithmen und die Kernalgorithmik des dort vorgestellten Programms werden detailliert ab Kapitel 3.5ff behandelt.

3.5 GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm

In Combinatorial Chemistry, molecules are assembled by linking suitable reagents taken in a combinatorial fashion from a large fragment space of starting materials. Often the number of possible combinations greatly exceeds the amount feasible to handle for in depth *in silico* analysis and even more for synthetic realization. Therefore, powerful tools to efficiently search in large solution spaces are required and can be provided by genetic algorithms which mimic Darwinian evolution. GARLig (**G**enetic **A**lgorithm using **R**eagents to compose **Lig**ands) has been developed to perform subset selections in large fragment spaces which satisfy target-specific 3D-scoring criteria. GARLig uses different scoring schemes such as AutoDock4 Score, GOLDScore and DrugScore^{CSD} as fitness functions. It has been optimized with respect to its genetic parameters and validated with respect to several targets of pharmaceutical interest. A large tripeptidic library of 20^3 members has been used to profile amino acid frequencies in putative substrates for trypsin, thrombin, factor Xa, and plasmin. A peptidomimetic scaffold assembled from a 25^3 entries large building block was used to test the performance of the evolutionary algorithm to suggest potent inhibitors of the enzyme cathepsin D. In a final case study, our program has been validated on a combinatorial drug-like library comprising 33750 members designed as putative inhibitors of thrombin.

These case studies demonstrate that GARLig finds experimentally confirmed potent leads by processing a significantly smaller subset of the fully enumerated combinatorial library. Furthermore, the profiles of amino acids computed by the genetic algorithm resemble observed amino acid frequencies found by screening peptide libraries in substrate cleavage assays. These results lead to the conclusion that GARLig provides an efficient and fast converging search through large compound spaces. It can therefore also be used in a prospective manner to detect the most promising candidates from large combinatorial libraries in de novo design projects.

3.5.1 Introduction

A major goal in computer-aided drug design is the automated generation of suitable ligands binding to a target protein under consideration. When these techniques depart from a given protein binding site they are summarized as *de novo* design approaches. The rapidly increasing amount of novel structurally characterized proteins, identified as putative targets for drug therapy, demands faster and more efficient approaches to suggest the most promising drug-like candidates which can easily be synthesized by medicinal chemists. Although a vast amount of software tools is available to design in a target-specific fashion individual ligands or medium-sized compound libraries using combinatorial principles, there is definite need for efficient tools to virtually screen large combinatorial libraries that cover fairly comprehensively a particular part of chemical space.

Several computer-aided ligand design methods have been reported [80]. The most popular *de novo* ligand design programs from the early 90ies are CAVEAT [81], SPROUT [82] and LUDI [11]. Since then, many new algorithms have been developed, particularly in the field of combinatorial docking. CombiDOCK [83] extends the well known program DOCK [7] by linking scaffolds and fragments combinatorially. Boehm *et al.* extended LUDI toward combinatorial applications and reported the discovery of nanomolar thrombin inhibitors [84]. FlexX^{C} [85] is an extension of the FlexX program series using the incremental built-up procedure in a combinatorial fashion. FlexNovo [86] uses a sequential growth strategy to link chemical fragments taken from a large space of starting materials. Here, the build-up procedure is based on a set of synthesis rules, physicochemical property filters and the FlexX scoring function. KNOBLE [87] designs novel small molecules by linking molecular fragments to a given core skeleton using at all levels simple and feasible chemistry. Potential candidates of fragments are retrieved from subpockets of proteins exhibiting similar pharmacophoric patterns, as identified by the Cavebase approach [88]. SQUIRREL [89] is a shape-based alignment method which decomposes small molecules into building blocks and compares them to a predefined query structure. The alignment is performed by means of a subgraph matching routine and the similarity is calculated using a fuzzy pharmacophore function.

Reliable and discriminative subset selection strategies have been proposed using iterative and deterministic strategies [90, 91] to avoid combinatorial explosion. Le Bailly de Tilleghem *et al.* suggested a probabilistic exchange of fragments where the overall procedure continues in an iterative manner [92]. The probability for a fragment to be exchanged depends on a fitness score which is optimized during the design process.

Besides the deterministic procedures mentioned above, heuristic optimization algorithms were considered in many ligand design approaches. One of the well-known algorithms to escape local minima on rugged energy landscapes is Simulated Annealing (SA). SAGE [93], HARPick [94] and Focus-2D [95, 96] use an SA strategy to virtually assemble molecular fragments to complete ligands. The similarity to a given reference molecule serves as objective function. PICCOLO performs an SA-driven subset selection of compounds with a multiobjective fitness function [97].

Many methods construct small molecules or specific libraries using evolution-inspired algorithms. TOPAS [98], Flux [99] and a new method developed by Schuller *et al.* [100] are well-known representatives. A disadvantage of these approaches is that structural information about a template ligand must be available.

Genetic Algorithms (GA) also belong to the class of evolutionary algorithms and have been widely applied in the field of *de novo* design, e.g. for the construction of small molecules, their subsequent structure optimization or the design of compound libraries [100-106]. However, there are only a few methods like SYNOPSIS, ENPDA, ADAPT and a multiobjective graph evolution method recently developed by Pattichis *et al.* [107-111] that use docking scores as fitness criteria in the selection and optimization of putative drug candidate molecules.

In this contribution, we propose GARLig, a self-adaptive genetic algorithm which has been tailored to meet the demands of library design. The underlying intention is to combine molecular fragments in order to assemble substrates or drug-like inhibitors as candidate molecules that are handled as populations in a genetic algorithm. This process is iterated over multiple cycles where successive generations of candidate molecules with an improving averaged fitness score are generated. A schematic overview presenting the applied GA is given in Figure 14. In more detail, GARLig performs side chain optimizations at a pre-defined molecular scaffold.

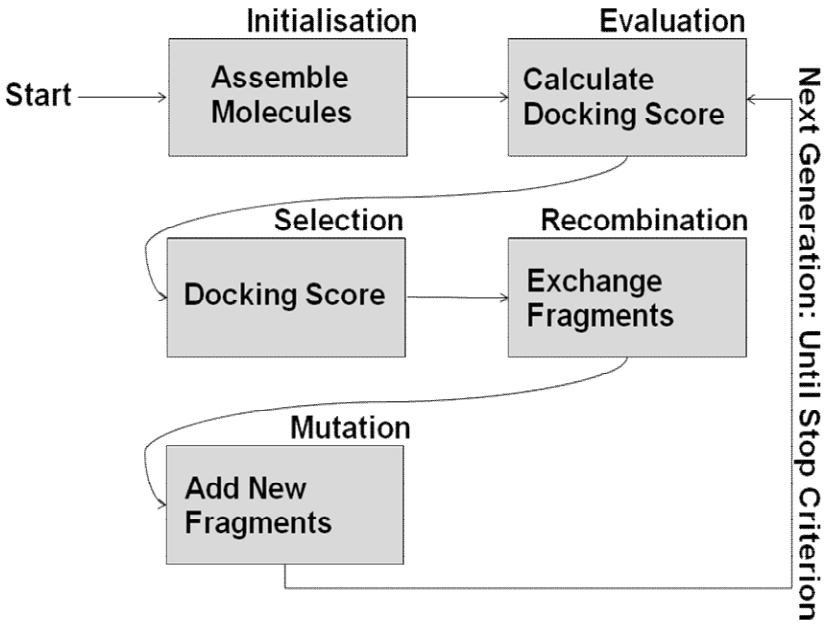


Figure 14: Schematic overview of our genetic algorithm showing the different stages of small molecule generation and evaluation.

The decorations to be attached are represented in terms of SMILES line notation. Special flags in the SMILES string denote attachment points for predefined chemical linking reactions. After assembly of the initial ligand population, their drug-likeness can be estimated according to Lipinski's Rule-of-Five [42]. Subsequently, 3D coordinates are generated using the program CORINA [112]. GARLig is then interfaced to two popular docking engines, AutoDock4 [65] and GOLD3.2 [38]. The whole workflow can be parameterized via one GA configuration file. Essential for the performance of GARLig is a reliable scoring scheme which considers, apart from the implemented scoring schemes in AutoDock4 Score and GOLDScore, the scoring function DrugScore^{CSD} [6]. These ranking systems are used to calculate the fitness value during the evolutionary process. Using such a docking score as fitness criterion keeps the approach independent of a priori knowledge about the possible binding of a ligand to the target protein under investigation. Apart from standard mutation- and crossover operators such as the Roulette Wheel Selection or Tournament Selection [113], an operator named Simulated Binary Crossover [114] has been implemented providing the genetic algorithm with a self-adaptive feature known from evolutionary strategies.

To understand how GARLig performs under the regime of different objective functions, a parameterization analysis has been performed using the Sequential

Parameter Optimization Toolbox (SPOT) [115], implemented in MATLAB (2007a, the MathWorks, Natick, MA). Therefore, an interface between GARLig and MATLAB has been realized. This interface supported the crucial parameterization step of the genetic algorithm for different scoring schemes and with respect to different target proteins.

The validation studies have been performed using several proteins of particular pharmaceutical relevance: trypsin, thrombin, factor Xa, plasmin and cathepsin D. The obtained results were compared to similar protocols using random search and Monte Carlo sampling algorithms. In the case of the serine proteases, a large tripeptidic library of 20^3 entries has been selected to identify those sequences which are the most likely proteinogenic substrates of different enzymes by the evolutionary learning process. For the enzyme cathepsin D, the most promising inhibitors have been selected from a peptidomimetic library comprising 25^3 theoretical members [116]. For this example, the performance of GARLig can be compared to the ADAPT program which uses DOCK Score as a fitness function. Furthermore, experimental data for some members of this library have been reported and allow to trace the relevance of our computational approach. As a final validation scenario, a large drug-like library of 33750 entries has been analyzed with respect to thrombin inhibition. Experimental reference data are available for this study as well. Thus, for all examples, enrichments with respect to experimentally known substrates or inhibitors can be compared with the most promising candidates found in the final generation of our computational approach.

3.5.2 Theory

Genetic Algorithms belong to the class of stochastic, population-based search algorithms which mimic Darwinian evolution. The procedure starts with an initial population of individuals proposed to be a solution for a given problem. Here, this population is a set of small molecules. The goal of the approach is to find suitable compounds in their most likely bioactive conformation adopted in the binding site of the target protein under investigation. An iterative process starts to detect better fitting individuals. They are more likely to dominate and undergo subsequent mutations and crossovers. The iterative GA continues until a termination criterion is exceeded. In the following, algorithmic details of GARLig will be specified.

Compound representation. GARLig is an rcGA that uses a real-value- instead of a binary-coded chromosome [117]. Thus, it is closely related to the algorithmic class of evolutionary strategies (ES). The following advantages are expected for these algorithms:

- No special function is needed to transform the chromosome from genotype to phenotype.
- For GAs, the energy landscape must be known a priori to find suitable encoding and decoding functions. However, rcGAs need less a priori knowledge about the search domain.
- Binary GAs have a known problem referred to as Hamming Cliffs [113]. Small changes in the binary code can take a dramatically large effect on the phenotype of an individual. This is usually undesired, especially at the end of the optimization process.
- A real-value encoding allows for small movements on the energy landscape of the given problem. A step-by-step approximation towards the global optimum is more likely for rcGAs.

Figure 15 shows a schematic overview of the main steps performed by GARLig. The user must provide the initial core scaffold and the reagents in SMILES line notation. They must be provided in a canonical way that conforms to the branching functionality of the SMILES notation. After assembly of the initial population, a 3D

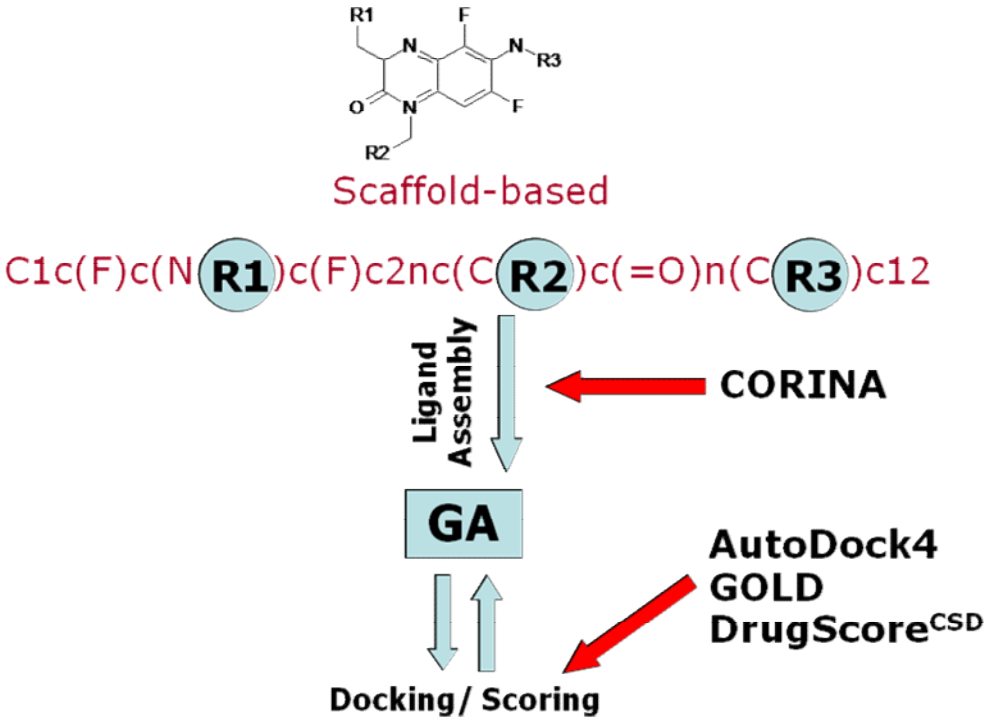


Figure 15: Workflow of GARLig. Scaffold and reagents are provided by the user in SMILES line notation. Fitness evaluation is performed via calculation of a docking score. The evaluated compounds are fed back into the core of the workflow, which is a genetic algorithm and mimics Darwinian evolution by applying mutation- and crossover operators. Red arrows denote where external tools support our workflow.

conformer is generated for each individual of the population in a format appropriate for the subsequently applied docking programs.

Fitness evaluation. GARLig provides an interface to the well-established docking programs GOLD3.2 and AutoDock4. The fitness value to be optimized during a GARLig run is a docking score, which gives an estimate of the expected binding affinity of a ligand to the receptor site under consideration. Important parameters needed for the docking calculation can be set directly via the GARLig configuration file. During the GA steps, docking geometries of the assembled molecules can be evaluated by the functions AutoDock4 Score and GOLDScore. Furthermore, a rescoring scheme can be applied, e.g. the function DrugScore^{CSD}, developed in our group [6].

Breeding. After the initial docking cycle, a new generation of small molecules is produced via the genetic operations crossover and mutation. Furthermore, GARLig takes advantage of the elitism functionality also known as the “survival of the fittest” principle. It allows direct copying of the fittest individual into the next generation

without undergoing further recombinatorial events. Two well-known selection strategies named “Tournament Selection” and “Roulette Wheel Selection” have been implemented into the program [113]. Both methods are based on the assumption that better fitting molecules are more likely to undergo crossover events which are characterized by breaking and reforming bonds in the selected individuals. Therefore, uniform single-point- and two-point crossover operators have been implemented to perform reagent exchange between two selected molecule partners [78].

Furthermore, an operator called “Simulated Binary Crossover” (SBX) [114] has been implemented and is intended to provide self-adaption functionality in the evolutionary process. First, the reagents have to be sorted with respect to their chemical similarity which is enumerated based on molecular descriptors such as molecular weight and number of H-bond donors/acceptors. These descriptors were selected because many other ones can be hardly used considering reagent-sized molecules.

The crossover operator SBX works as follows:

```
Foreach Selected Pair of Individuals:
    Foreach Residue Position:
        μm = pm(rk), μn = pn(rk)
        σ = abs(δ(μm,μn))
        cm = random.gauss(μm,σ), cn = random.gauss(μn,σ),
```

where the *abs*-function returns an absolute value of its argument, *random.gauss* returns a random number, μ is the value of a parent’s residue r at position k, δ denotes the distance (chemical similarity) between μ_m and μ_n and σ is the standard deviation. This procedure delivers more diverse compounds, if chemically diverse parents have been selected for mating. However, if parents were selected composed

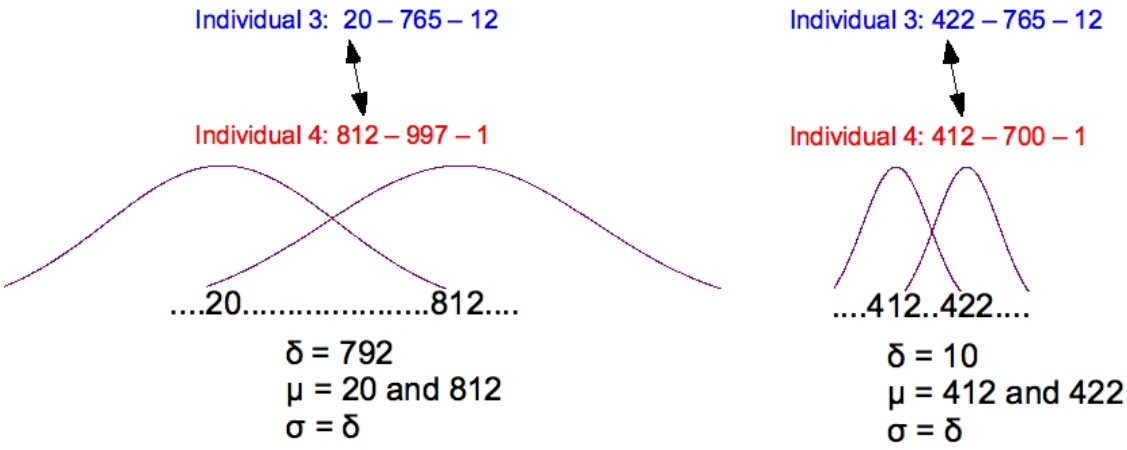


Figure 16: Depiction of the Simulated Binary Crossover (SBX) parameter. Two library decorations are selected for deriving a new decoration depending on the distance δ of the selected decorations. The fragments are ordered according to their chemical similarity, i.e. decoration 20 and 812 seem to be dissimilar. A new decoration can be chosen randomly according to two Gaussian probability distributions, where the mean are the integers of the selected decorations and standard deviation equals δ. On the left side, the genetic algorithm did not converge yet as a large number of fragments are possible for being selected. On the right side, the algorithm seems to converge to a chemical subspace. New decorations will be more similar than in the first case. In both examples, contracting crossover has the same probability as expanding crossover to avoid a prematurely converging GA.

of closely related or chemically similar substituents it will converge to a chemically similar subspace with a higher probability (Figure 16). Uniform mutation events can occur at each gene position of all chromosomes after crossing over whereby the frequency of switching genes is determined via a probability parameter.

Parameter optimization. The described algorithm has a few adjustable parameters which have to be optimized since the obtained result of the genetic algorithm will strongly depend on the parameter selection. Many methods can be applied to optimize external parameters. In the present case, a strategy for multiple parameter optimizations is required. We have chosen the recently developed **s**equential **p**arameter **o**ptimization (SPO) method [115]. SPO is a semi-automatic method that tries to keep the computational costs for determining an improved parameterization low. It requires a set of predefined intervals, each specifying the allowed values for a parameter (e.g. [1,50] for the population size) called the *region of interest* (ROI).

After defining the ROI, a set of parameters is generated to be used in the genetic algorithm. This first set comprises $q = 0.5 \cdot (k^2 + 3 \cdot k + 2)$ parameterizations, where k is the number of parameters being optimized. In the beginning of this procedure, *latin hypercube sampling* (LHS) is performed to get design points that are scattered

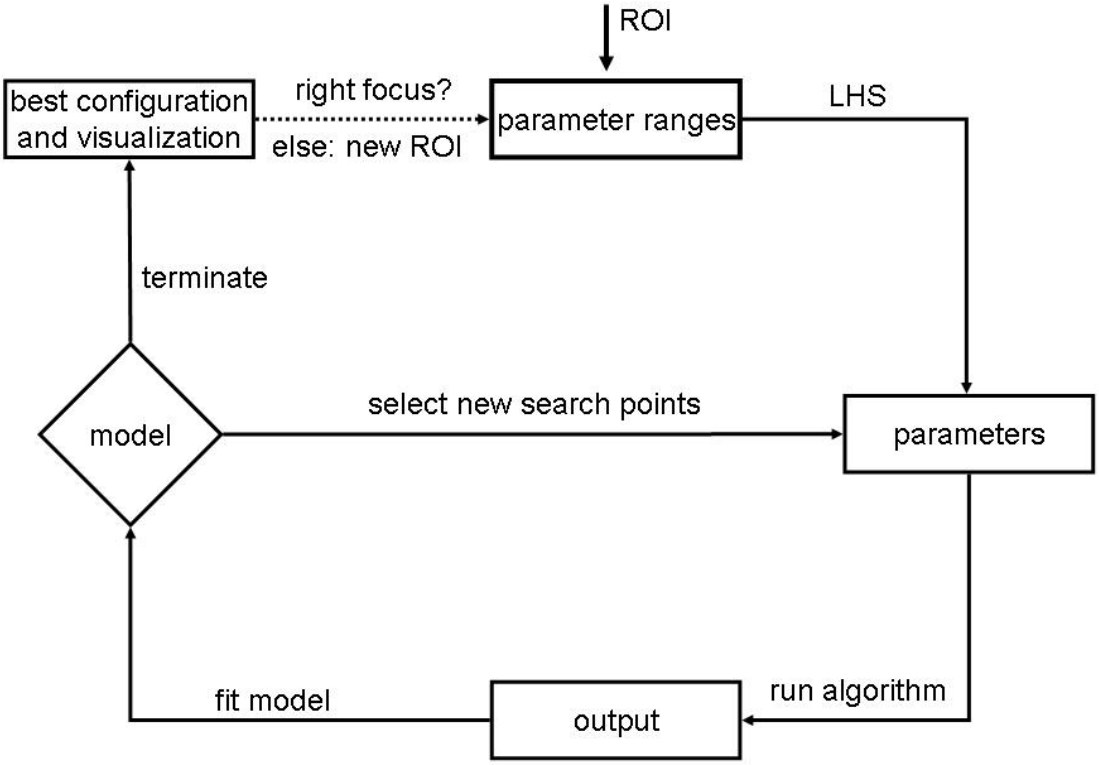


Figure 17: The parameter optimization performed by SPOT demands for a region of interest (ROI) given by the user. Here, borders must be specified for the parameters being optimized. Subsequently, *k* parameterizations are generated and improved in a recursive procedure. At the end of the procedure, the best parameterization for a scoring function under investigation is suggested.

uniformly over the whole ROI. Then, the SPO-loop illustrated in Figure 17 is executed. In each iteration the chosen parameterizations are evaluated by running the GA with this set of values obtaining the fitness scores *y*. Subsequently, SPO allows building a regression model using the derived parameterizations and the corresponding fitness values. The obtained model is further used to generate additional parameterizations *x* for the following purposes:

1. improving the regression model
2. finding better parameterizations.

The second (and most important) purpose is fulfilled, once the improvement (11) is maximized for *x*:

$$I(x) = \begin{cases} f(x) - f^* & if \quad f^* < f(x) \\ 0 & else \end{cases}, \qquad \textbf{(11)}$$

where *f** denotes the best value found so far. However, the exact value *I(x)* will not be known a priori, so that the SPO has to use the expected improvement based on the model *Y*.

Standard random search. This method is independent of generation cycles and randomly assembles a user-defined number of molecules and evaluates them via docking.

Monte Carlo sampling algorithm. A stochastic search without using crossover- or mutation events has been applied for comparison. The algorithm runs through a predefined number of generations and randomly assembles molecules. The best evaluated individual from each generation is kept and will only be replaced once a better fitting molecule is generated.

3.5.3 Programs, Datasets and Materials

Validation of GARLig using a tripeptide library with 20^3 entries for trypsin, thrombin, factor Xa and plasmin. GARLig was used to suggest the preferred amino acid profiles of tripeptides as potential substrates for the cleavage reaction in serine proteases. These sequences result from the final generation in our GA. The obtained profiles can be compared to experimental screening data collected in an enzymatic assay that records the cleavage of peptidic substrates labeled by a fluorescence probe [118]. The idea behind this simulation is that efficient cleavage of a peptide substrate requires selective and potent binding of the peptide sequence to be cleaved. Two main objectives were addressed in this validation:

1. Is the GA able to identify only those members of a fully enumerated substrate library that are known to be cleaved by the serine protease?

2. Is a docking score sufficient to discriminate between substrates and non-substrates?

The results obtained by GARLig were compared with a Monte Carlo Sampling and a standard random search. All three methods were intended to validate only 7.5% of a 20^3 entries large tripeptide library (8000 possible substrates, only 600 fitness evaluations via our GA). The validation has been performed using the crystal structures of trypsin, thrombin, factor Xa and plasmin (PDB-codes 1k1p, 1ype, 2w26 and 1bui) and a tripeptide scaffold linking all 20 proteinogenic amino acid residues at each position P1-P3. To investigate the sampling properties of the GA and the reliability of the applied scoring functions with respect to the targets, a parameterization study has been carried out exemplarily on trypsin using the MATLAB implementation of SPOT to reduce run time, GA fitness evaluations and to improve the docking scores in the procedure. Here, five parameters of the GA were selected to be optimized where the region of interest (ROI) was defined as follows:

- population size: 50-150 individuals per generation
- mutation probability ranging from 0 to 1
- crossover probability ranging from 0 to 1
- crossover type: one-point- and two-point crossover, SBX
- selection type: Roulette Wheel- and Tournament Selection

In the following, $q = 0.5 \cdot (5^2 + 3 \cdot 5 + 2) = 21$ parameterization scenarios were applied to determine the initial regression model, which was improved in four sequential steps each composed of three further parameterizations. The parameters suggested by the SPOT sampling procedure complemented the remaining GARLig- and docking parameters in the input file of the GA. At the end of the SPOT analysis, the best performing parameters (Bst) with respect to the chosen scoring function were adopted.

Before the docking calculations could be started, the assembled tripeptides had to be converted from SMILES line notation into a 3D representation in mol2-format using CORINA or pdbqt-format using AutoDockTools [119]. Docking geometries were generated using AutoDock4 and GOLD as docking engine, and DrugScore^{CSD} can additionally be applied to rescore the solutions generated by GOLD. With respect to docking into the serine proteases a set of “fast” docking parameters suggested by CCDC [120] has been applied to all GOLD docking runs. For each assembled tripeptide, 10 geometries were computed and for all compounds, standard protonation states were assumed, i.e. carboxylate groups were considered as deprotonated, and aliphatic amines and amidino-/guanidino groups were considered protonated. The protonation state at pH 8.0 of the protein was predicted using MOE [39].

Similarly, parameters more suitable for high throughput docking were used for AutoDock4. Hence, 10 geometries were computed for each ligand, and the total number of energy evaluations was set to 150000 using a population size of 150 individuals and 27000 generations in AutoDock’s Lamarckian GA.

In order to allow for an evaluation of larger compound libraries, the Condor Queuing System [121] was used to distribute the individual docking runs on a 50 nodes compute-cluster simultaneously.

Validation of GARLig on cathepsin D using a peptidomimetic library of 25^3 inhibitors. In a second validation scenario, GARLig was tested on a peptidomimetic library comprising 15625 entries designed to bind to the aspartyl protease cathepsin D (PDB-code 1lyb). This data set has already been evaluated using the program ADAPT. GARLig has been applied in a similar way, thus enabling direct comparison of the implemented GAs as both programs use docking scores as objective functions. Since for some of the entries experimentally determined affinity data have been published, it is possible to compute an enrichment of active compounds in the final generation of the optimization procedure. Again, the results of GARLig are compared to a Monte Carlo Sampling and a random search. As scoring functions perform differently well on individual target proteins, the results of GARLig might possibly depend on the selected scoring function. Accordingly, a SPOT parameterization study has been carried. To allow for a direct comparison with ADAPT, we parameterized GARLig comparable to the protocol used in ADAPT. The region of interest was defined as follows:

- population size: 10-50 individuals per generation
- mutation probability ranging from 0 to 1
- crossover probability ranging from 0 to 1
- crossover type: one-point- and two-point crossover, SBX
- selection type: Roulette Wheel- and Tournament Selection

In total, only 3.2% of the conceivable product space (only 600 compounds) was evaluated for each scoring function under investigation. With respect to the docking settings, AutoDock4 was parameterized with a higher number of energy evaluations (10^6) to improve the accuracy of the docking calculations. For GOLD standard settings were applied as no special parameterization has been suggested for the application to aspartic proteases. The assignment of protonation states was handled as mentioned above.

Validation of GARLig on thrombin using a library of 33750 sulfonic acid esters. In a third validation scenario, GARLig was tested on a compound space of 33750 drug-like inhibitors with a sulfonic acid ester scaffold with respect to thrombin binding [122] (Figure 18). Since experimental binding data have been published for twelve library entries [122], an enrichment of active compounds in the final generation of the

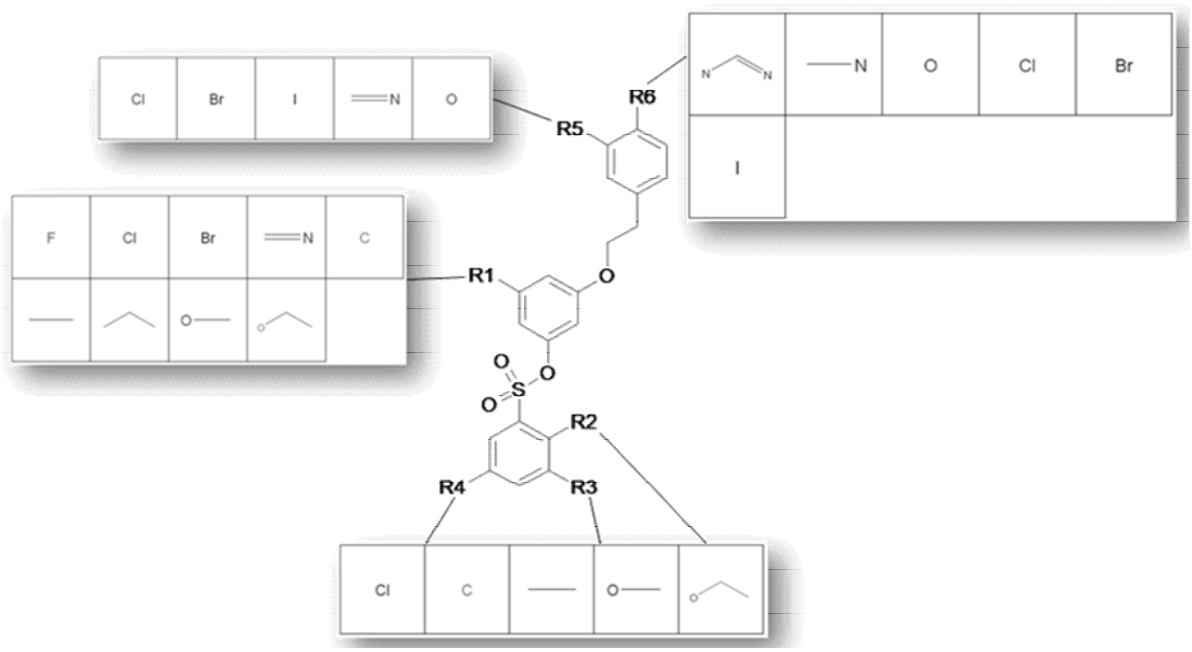


Figure 18: Scaffold of sulfonic acid ester derivatives as thrombin inhibitors. Different chemical groups have been attached to the six corresponding substitution sites.

optimization procedure could be computed for this study as well. All GA parameters optimized for the docking of the tripeptide substrates to thrombin were also applied in this study. Again, only 3.2% of the fully enumeratable product space (only 1080 compounds) were evaluated. In the present scenario, only GOLDScore was considered as objective function because it performed best at retrieving the tripeptide substrates for the serine proteases described above. GOLD's "fast" docking setup was used and the assignment of protonation states was handled as previously mentioned.

3.5.4 Discussion

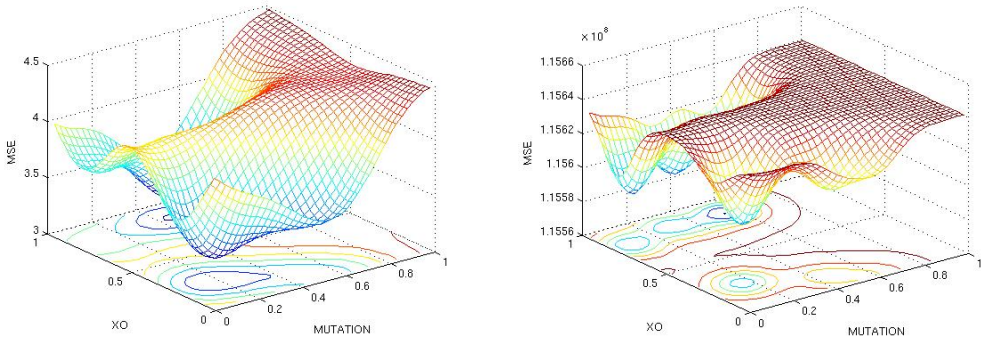
GARLig applied to a tripeptide substrate library towards trypsin, thrombin, factor Xa and plasmin. Figure 19 shows the results obtained by SPOT, which has been used to study the influence of different fitness functions and parameters on the GA. A set of surfaces is depicted showing the mean square error produced by the regression model expressed in terms of a combination of two optimized parameters. Closer inspection of these surfaces suggests that the minima of the mean square errors are located at similar positions with respect to the scoring functions DrugScore^{CSD} and GOLDScore. This overall agreement is achieved for the minima in the crossover type – crossover probability, selection type – crossover type and mutation probability – population size surfaces. The agreement in the regression models suggests to parameterize the GA similarly for the different scoring functions. The only exception was a GARLig run using AutoDock4 Score as an objective function. Here, the parameters being optimized seem to be mutually independent preventing any optimization. Probably, this behavior finds an explanation in the reduced number of energy evaluations considered in the docking setup.

Table 7 lists the best parameterizations for each scoring scheme. It is remarkable that similar GARLig parameterizations can be used for GOLDScore and DrugScore^{CSD}. Interestingly, mutation- and crossover probability parameters must be set to ≤ 20% and ≥ 60% for all applied scoring functions. These probability thresholds are consistent with values found in GA literature [113].

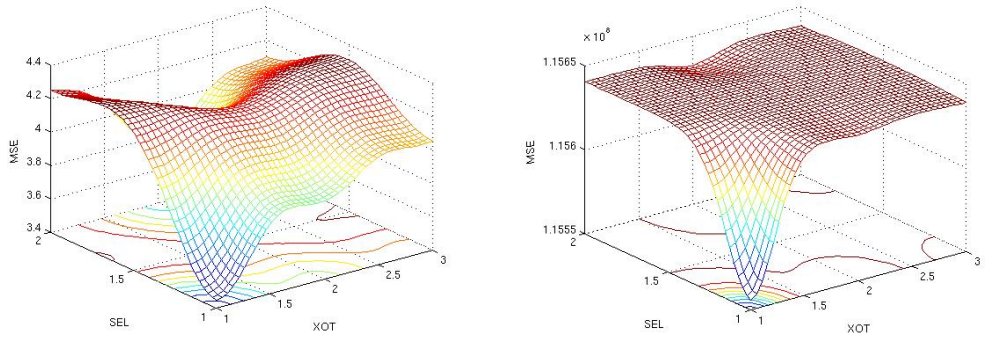
Figure 20 shows the convergence of the GARLig runs using different scoring functions as fitness criteria. In each case, the best parameterization setup, suggested by SPOT, was applied. The results of the GA were then compared to a standard random search and a Monte Carlo Sampling algorithm. As expected, GARLig outperforms the two other algorithms particularly resulting in a much faster convergence of better fitness values.

Table 8 shows the top scored tripeptide sequences of substrates for trypsin in the final generation of the GA with respect to different scoring schemes. For each scoring function, amino acids are suggested among the top scored four solutions of the entire library that are experimentally known to occur at the different positions P1-P3.

DrugScore^{CSD} **GOLDScore**



DrugScore^{CSD} **GOLDScore**



DrugScore^{CSD} **GOLDScore**

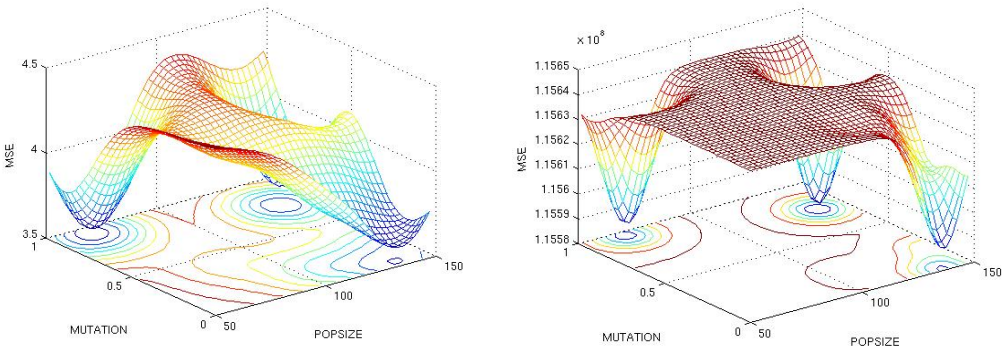


Figure 19: Energy Landscapes as a result of the SPOT parameter determination. Results are shown for the two scoring functions GOLDScore and DrugScore^{CSD}. The plots show the mean square error (MSE) of the regression model computed by SPOT as a function of different variables such as mutation probability (MUTATION), population size (POPSIZE), selection type (SEL), crossover type (XOT) and crossover probability (XO). Convincing similarity can be detected considering the landscapes of the different scoring functions as both of them have the same error minima of the regression model. This leads to the conclusion that a similar parameterization can be chosen among different fitness functions applied to the GA.

Table 7: Best parameterization for each scoring scheme.

Scoring Function	Best Score	[1]POPSIZE	[2]MUTATION	[3]XO	[4]XOT	[5]SEL
GOLDScore	48.98	**60**	**0.05**	**0.71**	**1**	**2**
DrugScore^{CSD}	-147788	**60**	**0.05**	**0.71**	**1**	**2**
AutoDock4 Score	-7.09	68	0.06	0.77	2	2

GOLDScore and DrugScore^{CSD} can be equally parameterized in the serine proteases study. The table shows the variables [1]population size (POPSIZE), [2]mutation probability (MUTATION), [3]crossover probability (XO), [4]crossover type (XOT) and [5]selection type (SEL) for each scoring function in use. A GARLig run can be started with a mutation probability ≤ 20% and a crossover probability ≥ 59% in all cases.

Table 8: Results of a GARLig run using the best GA parameterization for different scoring schemes.

Scoring Function	Substrate	Scoring Rank Final Generation
DrugScore^{CSD}	P-H-R	2
GOLDScore	K-H-R	1
AutoDock4 Score	Q-A-R	4

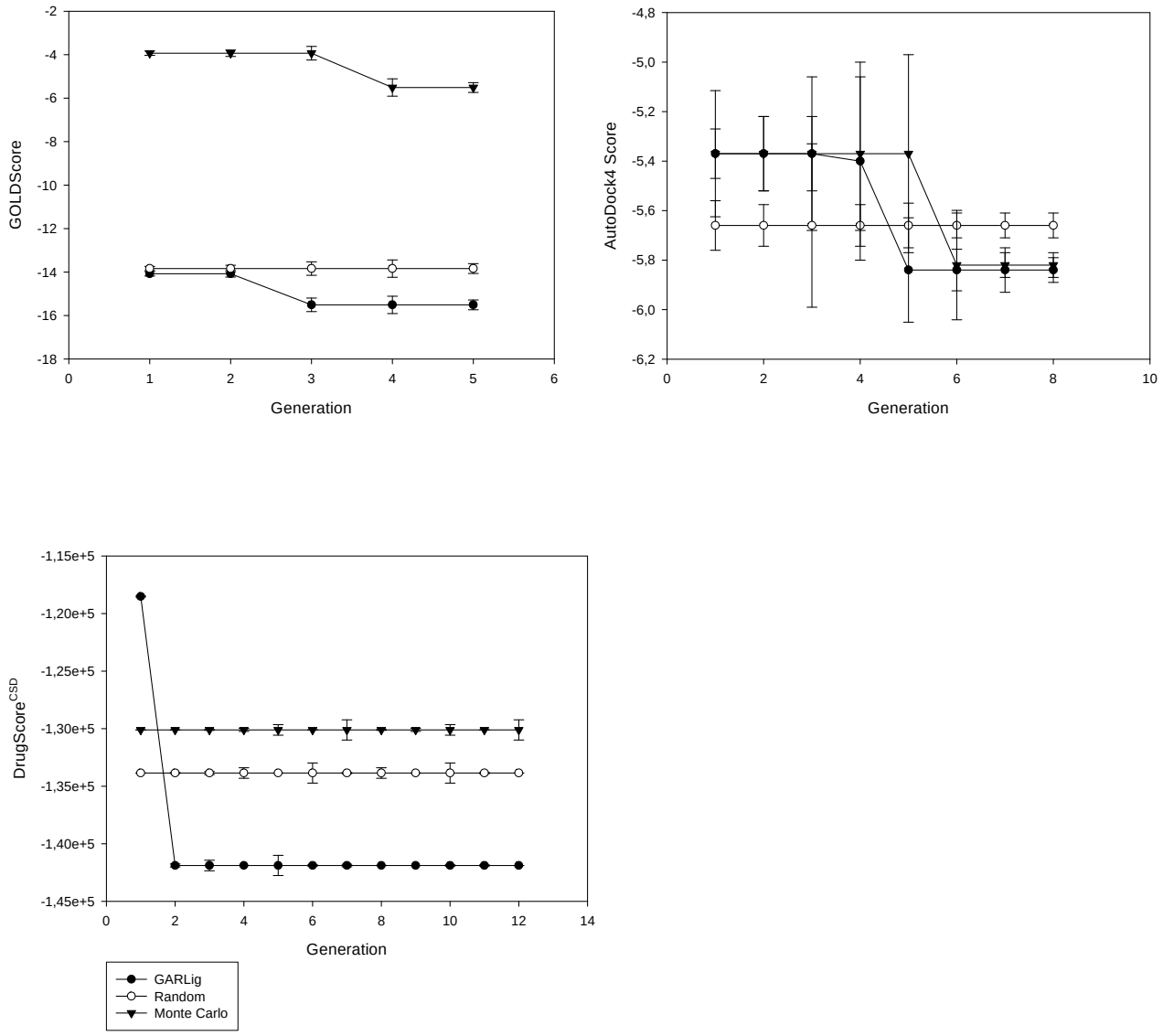


Figure 20: GARLig has been run three times with the best parameters determined by SPOT. The GA results are compared to a Monte Carlo Sampling and a standard random search. The error bars show the standard deviation of the fitness values computed in the three runs and the points show the fitness value as a function of the generation cycle.

This result could be obtained by a docking parameterized as “faster and less accurate”. Obviously, it can still identify sequences which are known trypsin substrates. Since GOLDScore was the only scoring function able to place a proteinogenic trypsin substrate (K-H-R) on rank 1 [118], DrugScore^{CSD} and AutoDock4 Score were not further investigated in case of the remaining serine proteases.

Figure 21 shows that GARLig suggests amino acids at the different positions P1, P2 and P3 in its final generation that are actually known to occur in substrate sequences of the different proteases. For the serine proteases studied here, amino acid profiles are in good agreement with the results obtained experimentally [118]. The position P1 shows high frequencies of lysine and arginine residues in the trypsin case. Additionally, tyrosine is proposed by our calculations to be a potential S1 anchor group. At position P2, aspartate, valine and glutamine are not identified in our simulation, but frequently occuring residues like tyrosine and proline are captured. Considering P3, there is a fair agreement between experimental and computational results except for proline, which is highly frequent in our calculations but does not occur in the cleavage experiments. The factor Xa case study reveals a good computational versus experimental profile agreement for the positions P1 and P2. Arginine, glutamine and phenylalanine were not captured by the algorithm, but more frequent amino acids like tryptophan and tyrosine are recognized by our algorithm. In the thrombin case study, a good agreement between experimental and computational results for the positions P1 and P2 is obtained. As the F-P-R sequence is known to be a proteinogenic substrate and phenylalanine was not recognized at P3 in our results, we decided to perform the calculations for thrombin again but this time, the P3 amino acid was docked into the protein in D-configuration. The histograms of the substrates with the P3 amino acid in D-configuration are more similar to the experimental results. Now, the DF-P-R substrate was found on the first scoring rank. A cleavage of this substrate was proven earlier [123]. Substrates with a D-phenylalanine are able to interact with Trp215, Ile174 and Leu99 in the P3-subpocket of thrombin whereas this residue remains solvent-exposed applied in its L-configuration [124]. For this reason, the D-P3 amino acid profile is clearly more similar to the profile determined experimentally. The results collected from the plasmin study are overall convincing. The most frequent P1 amino acids such as lysine and arginine are recognized by our approach. P2 amino acids like phenylalanine, tryptophan and tyrosine and P3 residues like glutamine are in good agreement with the experimental cleavage preferences. Table 9 finally shows the substrate sequences found on rank 1 among all serine proteases in the final generation of our GA.

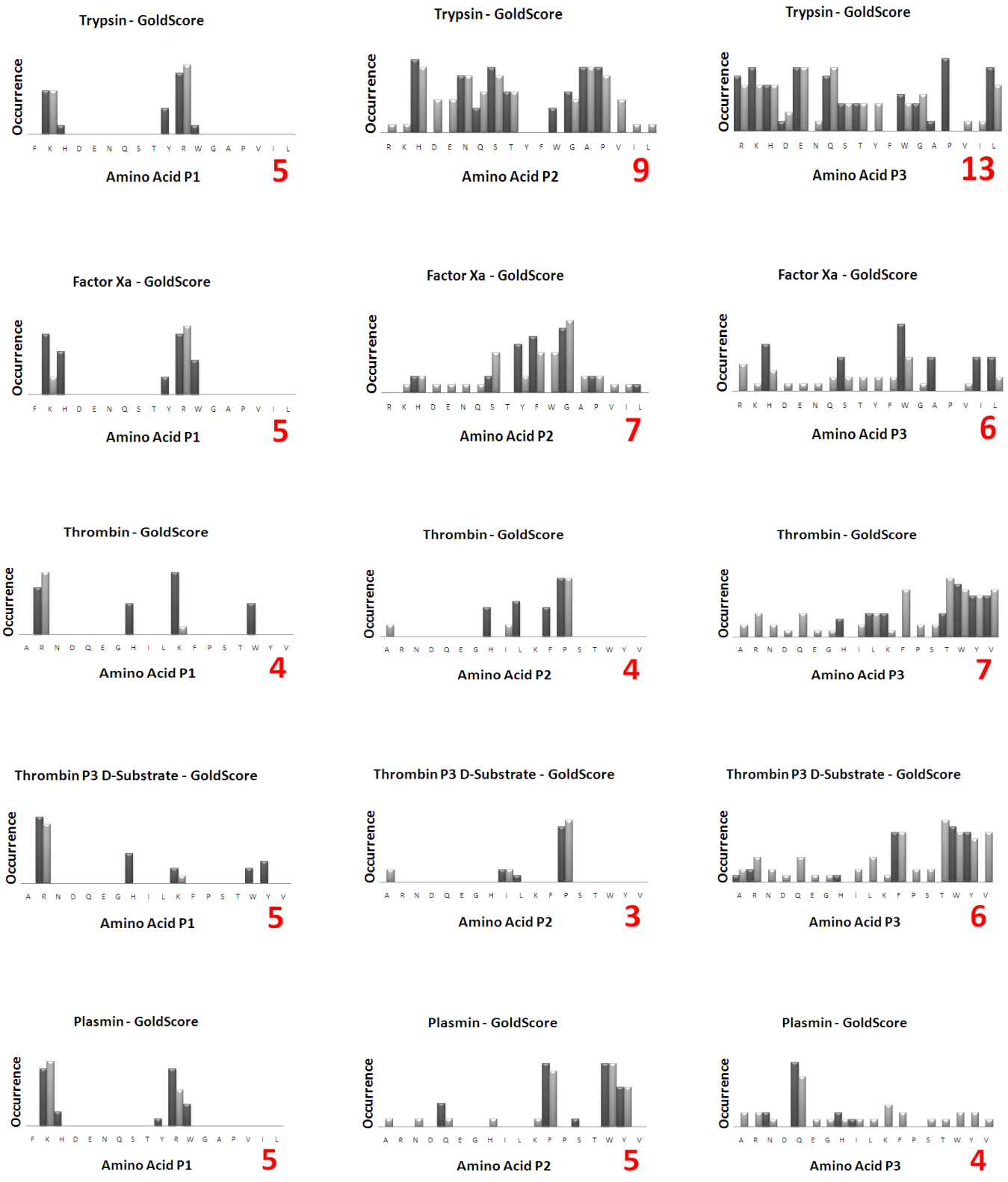


Figure 21: Results in the tripeptide substrate study on the serine proteases. For each protein target, the preferred profile of experimentally observed proteinogenic amino acids is recorded across the positions P1-P3 [118]. Experimentally obtained amino acid frequencies (light blue) are overlaid onto the computed ones (grey). Red numbers denote the number of amino acids which remained at the positions P1-P3 in the last generation of the GA.

Table 9: Results of a GARLig run using GOLDScore as a fitness function.

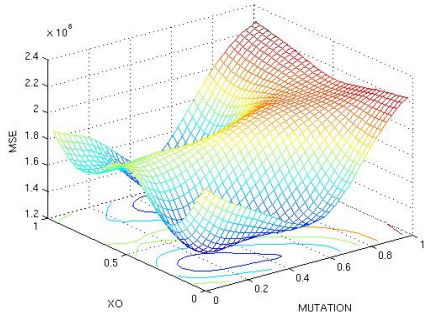
Target Protein	Substrate
Thrombin	DF-P-R
Thrombin	W-L-K
Factor Xa	W-G-K
Plasmin	H-F-W

The P1 position of the calculated tripeptide substrates shows preferred accumulation of lysine and arginine among all serine proteases. These residues are known to be cleaved preferentially in trypsin-like proteases which exhibit an aspartic acid residue in the S1 pocket. Furthermore, the algorithm suggests histidine, tyrosine and tryptophan at P1 which were not reported to occur in the best substrates [118]. However, several crystal structures have been reported which show that such residues can be accommodated in the S1 pocket of these proteases [125-128]. Not all experimentally observed residues are identified by our algorithm. Therefore, our approach seems to be more suitable to prioritize a fraction of a fully enumerated library for experimental evaluation. Figure 21 clearly shows that our approach is reliable to predict preferred residues at the respective positions in nearly all cases among the different serine proteases and significantly reduces the library size in the last GA cycle.

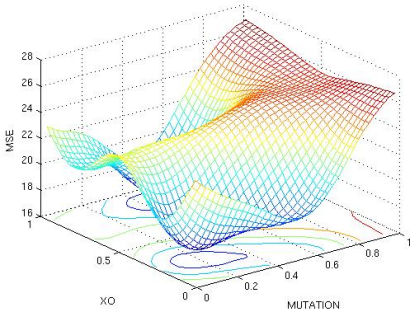
Cathepsin D library. The parameterization analysis (Figure 22) suggests again an overall agreement with respect to the minima of the estimated errors among all scoring functions. This time, using AutoDock4 Score as the objective function, the parameters being optimized seem to be dependent on each other. This might be due to the increased number of energy evaluations in the docking setup of AutoDock4.

Table 10 shows that comparable parameters can be used in the GA in this case. Among the different scoring schemes, GARLig performed better using the Tournament Selection and the self-adaptive Simulated Binary Crossover parameter.

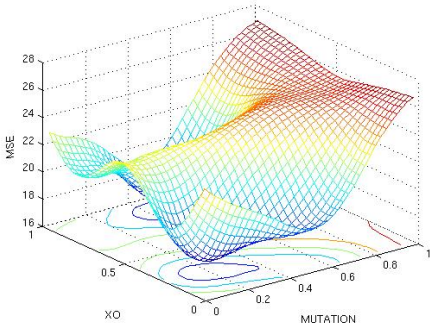
DrugScore^{CSD}



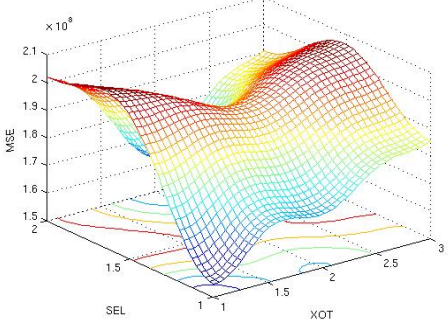
AutoDock4 Score



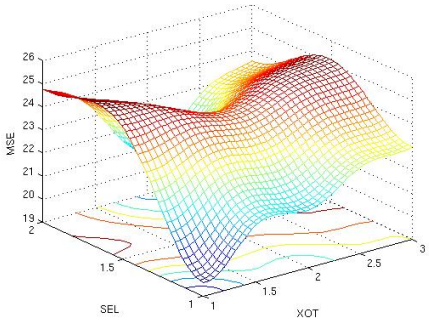
GOLDScore



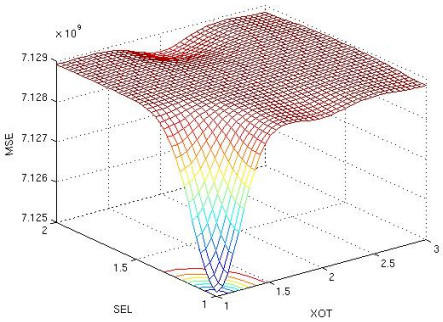
DrugScore^{CSD}



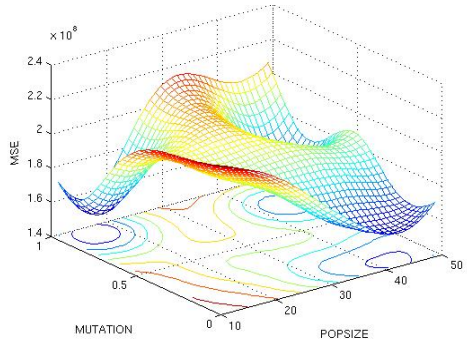
AutoDock4 Score



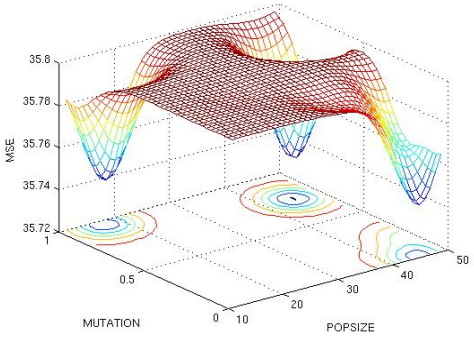
GOLDScore



DrugScore^{CSD}



AutoDock4 Score



GOLDScore

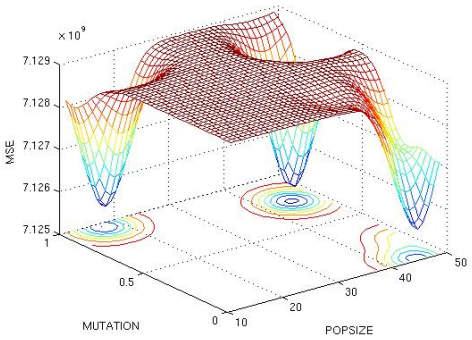


Figure 22: Energy Landscapes as a result of the SPOT parameter determination in the cathepsin D study. The plots show the mean square error (MSE) of the regression model computed by SPOT as a function of the mutation probability (MUTATION), population size (POPSIZE), selection type (SEL), crossover type (XOT) and crossover probability (XO). Again, a convincing agreement can be seen considering the landscapes of the different scoring functions as all of them have the same error minima of the regression model. This leads to the conclusion that a similar parameterization can be chosen among different fitness functions applied to the GA.

Table 10: Best parameterization for each scoring scheme.

Scoring Function	Best Score	[1]POPSIZE	[2]MUTATION	[3]XO	[4]XOT	[5]SEL
GOLDScore	82.14	**29**	**0.09**	**0.98**	**3**	**1**
DrugScore^{CSD}	-201447	**29**	**0.09**	**0.98**	**3**	**1**
AutoDock4 Score	-12.22	35	0.11	0.63	2	1

In the cathepsin D study, GARLig can be parameterized equally for GOLDScore and DrugScore^{CSD}. The table shows the variables [1]population size (POPSIZE), [2]mutation probability (MUTATION), [3]crossover probability (XO), [4]crossover type (XOT) and [5]selection type (SELTYPE). All GARLig runs can be started with a mutation probability ≤ 16% and a crossover probability ≥ 52%. Furthermore, there is an agreement in using Tournament Selection and the self-adaptive Simulated Binary Crossover parameter.

Either GOLDScore or DrugScore^{CSD} can be applied with the same set of parameters and mutation- and crossover probabilities must be generally set to ≤ 16% and ≥ 52%. Figure 23 shows the results of the different GARLig runs compared to a standard random search and a Monte Carlo Sampling.

Although experimental binding data are only reported for 9 highly potent entries out of the total of 15625 compounds (Figure 24a), the GA was able to identify some of these hits in the last generation. In all scenarios, the chemical variation suggested in the last cycle is dramatically reduced compared to the initial chemical space (Table 11). Our best performing scenario was the GOLDScore run, which comprised 2 known binders and 10 additional entries of the total combinatorial library of 15625 entries. Figure 24b shows the fragment set which remained in the last generation and Figure 24c depicts interaction diagrams of the compounds placed on rank 1 and 2. The top scoring compound is known to inhibit cathepsin D at 5.8 nM whereas, unfortunately, affinity data is not published for the second compound. However, comparing the key interactions performed by these two compounds suggests that also the second best scored compound should be a high affinity binder.

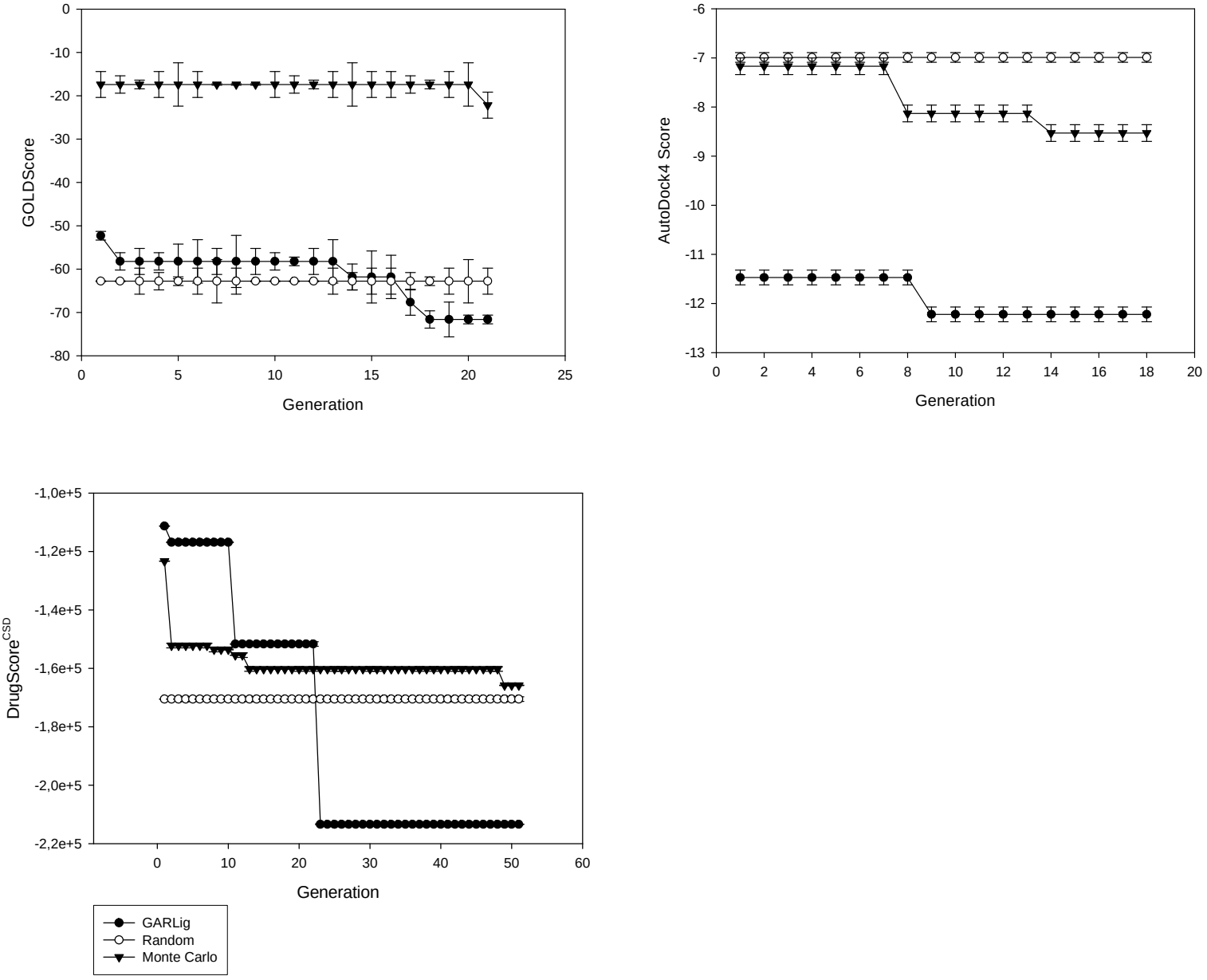
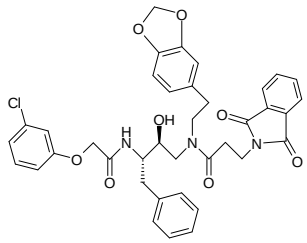
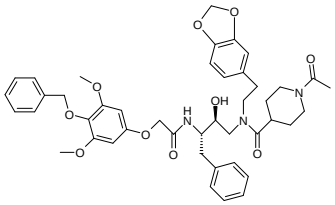
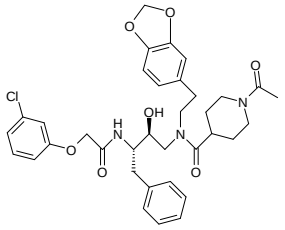
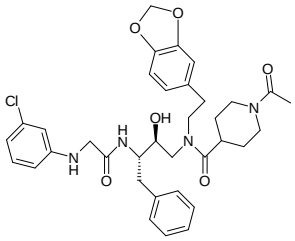
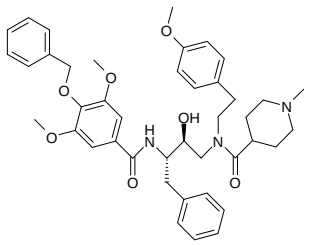
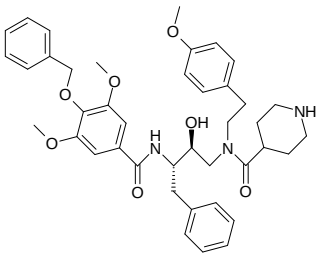
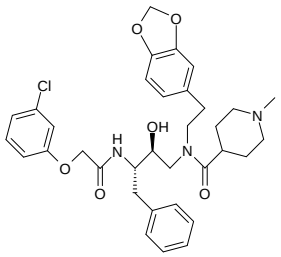
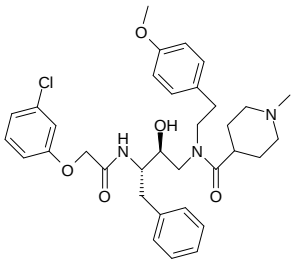
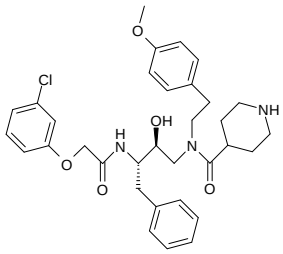


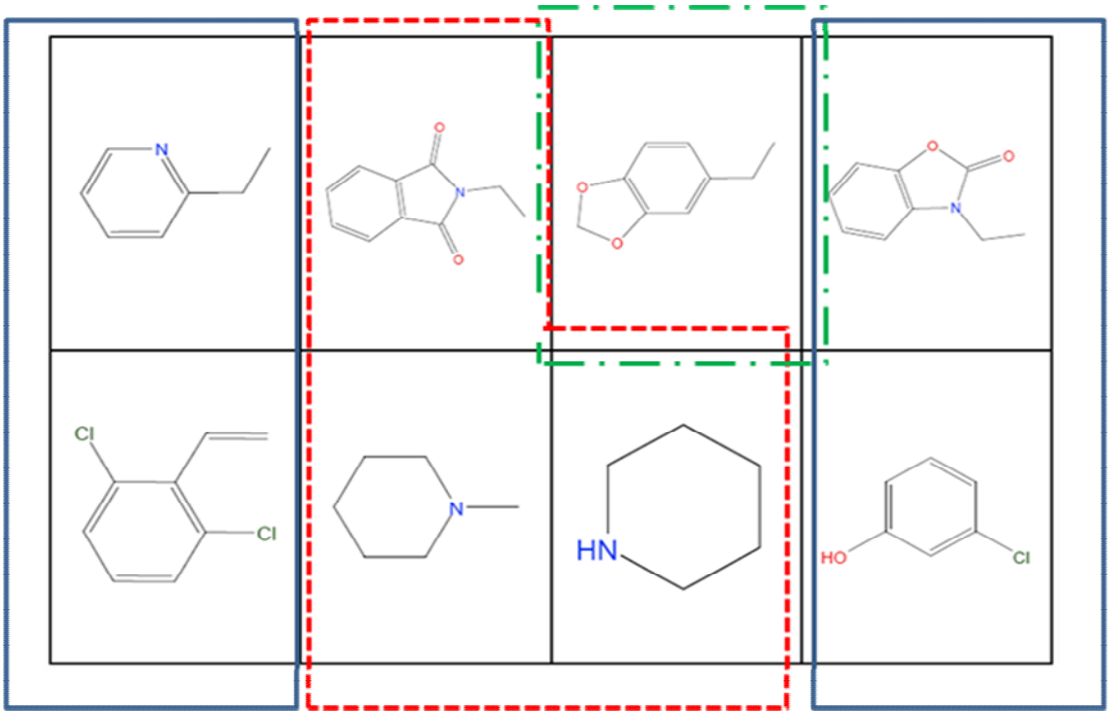
Figure 23: GARLig has been run three times with the best parameters determined by SPOT. The GA results are compared to a Monte Carlo Sampling and a standard random search. The error bars show the standard deviation of the fitness values computed in the three runs and the points show the fitness value as a function of the generation cycle.

(a)

Entry No.	Chemical Structure	Cathepsin D K_i
1		15 nM
2		4.3 nM
3		1.9 nM
4		1.3 nM
5		58 nM
6		71 nM

7		5.8 nM
8		9.8 nM
9		63 nM

(b)



(c)

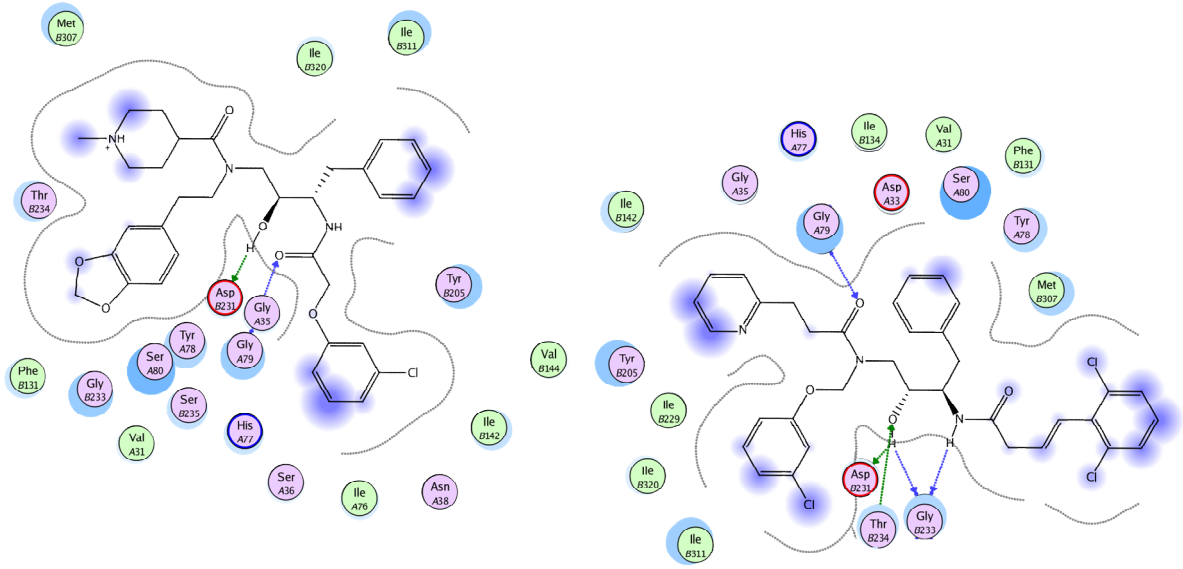


Figure 24: (a) Experimental binding data and chemical structures of 9 highly potent members of the cathepsin D library. **(b)** The fragment collection which remained in the last generation of the GARLig run. The fragments surrounded by the solid lines were only found at position R3, encapsulated by dashed lines at R2 and highlighted by dashed and dotted line at R1. The remaining chemical space shown here (3 x 1 x 4) contains two known cathepsin D binders (compound 1 and 7). **(c)** The top scoring compound proposed by GARLig (left) is a cathepsin D binder known to inhibit at 5.8 nM (compound 7). The right picture shows the second ranked compound. The key interaction between the hydroxyl group of this inhibitor and the Asp 231 residue of the protein and rather hydrophobic interactions (blue spheres) performed by the side chains appear in both results.

The GA runs using DrugScore^{CSD} also converged to a small subset of 12 library entries, however, not including one of the experimentally confirmed binders. As mentioned unfortunately, only for 9 high affinity binders experimental data are reported, thus there might be a couple of reasonable binders among the suggested candidates. Using AutoDock4 Score as an objective function, 4 of the experimentally characterized binders are listed among 64 entries suggested in the final GA generation.

All of our GA runs converge much faster and create significantly smaller libraries in the final generation compared to the ADAPT program. In the cathepsin D experiment we used the same peptidomimetic library and an identical number of GA evaluations. The GOLDScore run converged within 17 generations to the final 12 entries comprising two known highly potent binders. The AutoDock4 run converged within 10 generations to 64 entries comprising four of the known binders. To select the same

Table 11: Results of a GARLig run using the best parameterization for the different scoring schemes.

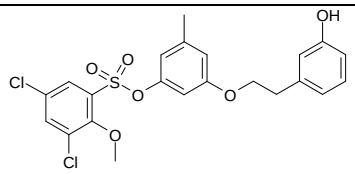
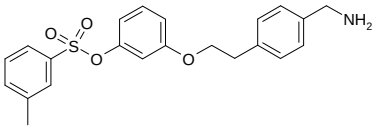
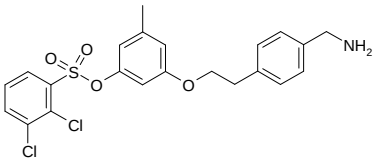
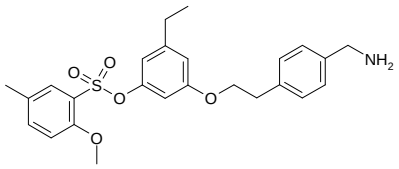
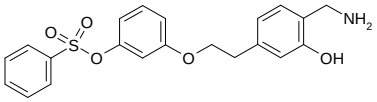
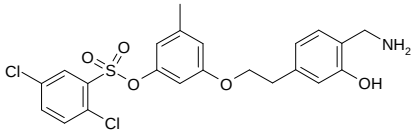
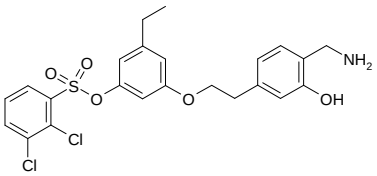
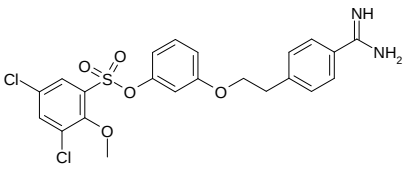
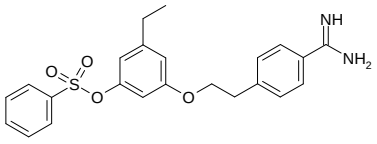
Scoring Function	Library Size Last Generation (R1-R3)	Known Binders in Library
GOLDScore	**3 x 1 x 4**	2
DrugScore^{CSD}	**3 x 2 x 2**	0
AutoDock4 Score	**4 x 4 x 4**	4

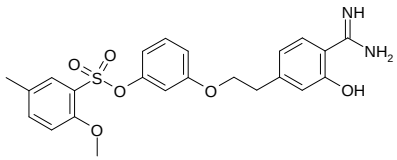
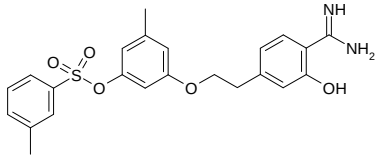
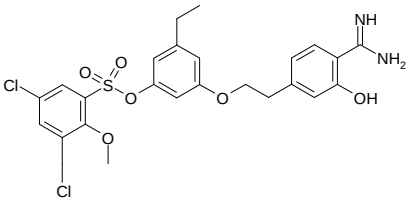
amount of known hits, ADAPT converges only after 50 generations with a subset of 392 compounds (8 x 7 x 7).

Library of sulfonic acid ester inhibitors for serine proteases. For the last example, the results of our GA are less convincing and clearly point to the limitations of such an approach. The initial library of 33750 entries could only be reduced to about a third, still comprising 11250 compounds. Nevertheless, the 12 binders reported as potent inhibitors of thrombin could be detected amongst them (Figure 25a).

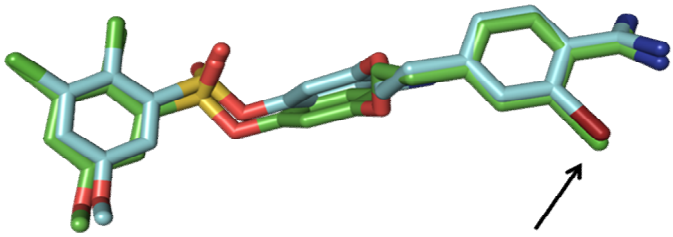
Seeking for an explanation for the limited power to reduce the total size of the initial library by our GA, we detected that all generated docking solutions obtained rather similar GOLDscores. Thus, a sufficient discrimination cannot be expected. On the one hand, this could indicate the relevance and reliability of the docking solutions and a pretty target-tailored choice of the building blocks for the thrombin library. On the other hand, GA parameters such as, e.g. Tournament Selection run into decision problems when docking scores are not sufficiently discriminating among a generation of individual compounds. The per-atom contribution to the docking score, e.g. of a bromine or chlorine atom placed at closely related positions differ only slightly, much too little to sufficiently guide the GA with respect to scoring differences. Therefore, the selection between both the derivatives by the GA will remain arbitrary and rather inefficient (Figure 25b).

(a)

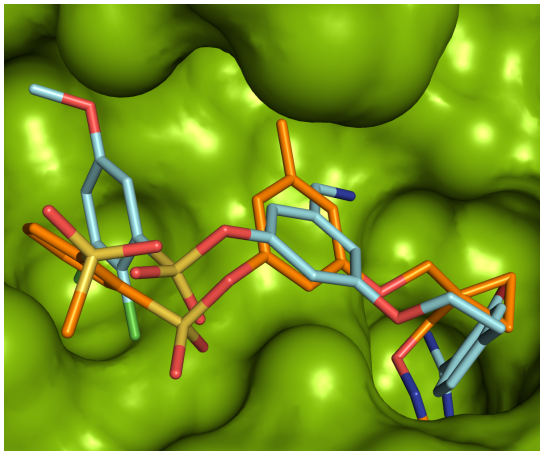
Entry No.	Chemical Structure	Thrombin pK_i
1		4.53
2		4.51
3		6.19
4		5.78
5		5.25
6		6.31
7		5.23
8		7.53
9		7.88

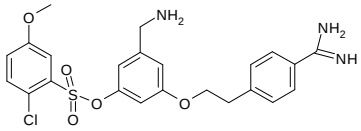
10		6.52
11		7.15
12		6.82

(b)

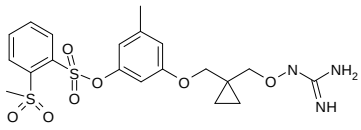


(c)





GARLig best hit



Crystal structure

Figure 25: (a) Experimental binding data and chemical structures of 12 highly potent thrombin inhibitors experimentally found in the sulfonic acid ester library. **(b)** Superposition of the GOLD docking solutions of two compounds generated in a GARLig run. The arrow points to a position where the substituents differ among the different inhibitors. The green compound (GOLDScore: 66.69) contains a chlorine atom whereas the cyan compound (GOLDScore: 66.47) contains a bromine atom. The high similarity of the docking geometries leads to decision problems in the selection step of the genetic algorithm. **(c)** Superposition of the best scored library member found by GARLig (blue) with a similar sulfonic acid ester inhibitor cocrystallized with thrombin (orange, PDB-code 1t4u).

Consequently, both alternative decorations in the library will be progressed to the last generation of the GA producing a library with only slightly reduced chemical diversity and therefore a huge number of equally scored library entries. In conclusion, many chemical groups selected as putative substituents at the central scaffold are appropriate as potential interaction partners with the target protein. Figure 25c shows the best ranked compound generated by GARLig superimposed with the crystal structure of a related sulfonic acid ester derivative (PDB-code 1t4u). The docking geometry of the best-ranked library candidate (no experimental inhibition data available) is in fair agreement with the experimentally determined binding mode of the related compound.

3.5.5 Conclusions

GARLig, a genetic algorithm to support the design of combinatorial libraries with self-adaptive features has been introduced. It can employ AutoDock4 Score, GOLDScore, and DrugScore^{CSD} as fitness functions. The use of docking scores as a fitness criterion can be regarded controversially with respect to library design. As a major advantage our calculations can be initiated without requiring a priori information about ligands previously described to bind to the target structure under consideration. As major disadvantage the computational complexity of the multiple docking step has to be regarded in the context of the known limitations of our currently applied scoring functions.

The program has been validated on several proteases of pharmaceutical relevance: trypsin, thrombin, factor Xa, plasmin and cathepsin D. In all validation cases, parameter optimization using the tool SPOT was performed prior to the actual GARLig runs. Interestingly, similar parameters were found to be the optima, independent of the applied fitness function and the considered biological target. Using these parameters, GARLig was able to predict profiles of preferred amino acids to be found in putative substrates cleaved by the different serine proteases. They show convincing similarity to experimentally determined substrate profiles. Our GA suggested reasonable substrate sequences or generates known inhibitors for serine proteases or the aspartic protease cathepsin D on high scoring ranks by reducing the size of the fully enumerated library to only 7.5% and 3.2% of all possible entries. Compared to ADAPT, GARLig shows much faster convergence and better enrichments of known binders in the final generation. We think this faster convergence leading to a smaller chemical space can be explained by the combination of a good GA parameterization and GARLig's self-adaptive feature. Once a parameterization study has been performed on a new protein, our method can be used to dramatically reduce the combinatorial chemical space by evaluating lower fractions of the given data than in case of the ADAPT program.

For library of thrombin inhibitors, all 12 known binders were comprised in the final library, however, this library was only reduced to about a third of the initial chemical space. As most of the substituents selected as primary building blocks are known to

interact with thrombin at their respective positions and the obtained docking scores do not discriminate sufficiently enough, it cannot be expected that a dramatic reduction of the chemical space is achieved for this example.

In all cases, GARLig performs better compared to a random search and a simple Monte Carlo Sampling.

Further validation studies would be required evaluating larger data sets. Unfortunately, only a small number of libraries are available in public domain, for which a significant fraction of its members have been characterized in terms of structure and binding affinity.

4 Literaturverzeichnis

1. Bajorath, J., *Understanding chemoinformatics: a unifying approach.* Drug Discov Today, 2004. **9**(1): p. 13-4.
2. Gohlke, H., M. Hendlich, and G. Klebe, *Knowledge-based scoring function to predict protein-ligand interactions.* J Mol Biol, 2000. **295**(2): p. 337-56.
3. Dixon, J.S., *Evaluation of the CASP2 docking section.* Proteins, 1997. **Suppl 1**: p. 198-204.
4. Warren, G.L., C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, and M.S. Head, *A critical assessment of docking programs and scoring functions.* J Med Chem, 2006. **49**(20): p. 5912-31.
5. Deng, Z., C. Chuaqui, and J. Singh, *Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions.* J Med Chem, 2004. **47**(2): p. 337-44.
6. Velec, H.F., H. Gohlke, and G. Klebe, *DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction.* J Med Chem, 2005. **48**(20): p. 6296-303.
7. Ewing, T.J., S. Makino, A.G. Skillman, and I.D. Kuntz, *DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases.* J Comput Aided Mol Des, 2001. **15**(5): p. 411-28.
8. Morris, G.M., D.S. Goodsell, R. Huey, and A.J. Olson, *Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4.* J Comput Aided Mol Des, 1996. **10**(4): p. 293-304.
9. Weiner, S.J., Kollman, P. A., Nguyen, D. T., Case, D. A., *An all-atom force field for simulations of proteins and nucleic acids.* J Comput Chem, 1986. **7**.
10. Bohm, H.J., *LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads.* J Comput Aided Mol Des, 1992. **6**(6): p. 593-606.
11. Bohm, H.J., *The computer program LUDI: a new method for the de novo design of enzyme inhibitors.* J Comput Aided Mol Des, 1992. **6**(1): p. 61-78.

12. Rarey, M., B. Kramer, T. Lengauer, and G. Klebe, *A fast flexible docking method using an incremental construction algorithm.* J Mol Biol, 1996. **261**(3): p. 470-89.
13. Eldridge, M.D., C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee, *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* J Comput Aided Mol Des, 1997. **11**(5): p. 425-45.
14. Kirtay, C.K., Mitchell, J. B. O., Lumley, J. A., *Knowledge Based Potentials: the Reverse Boltzmann Methodology, Virtual Screening and Molecular Weight Dependence.* QSAR & Combinatorial Science, 2005. **24**: p. 11.
15. Sippl, M.J., *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.* J Mol Biol, 1990. **213**(4): p. 859-83.
16. Muegge, I., *PMF scoring revisited.* J Med Chem, 2006. **49**(20): p. 5895-902.
17. DeWitte, R.S., Shakhnovich, E. I., *SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence.* J Am Chem Soc, 1996. **118**: p. 11.
18. Lyne, P.D., *Structure-based virtual screening: an overview.* Drug Discov Today, 2002. **7**(20): p. 1047-55.
19. Gohlke, H. and G. Klebe, *Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors.* Angew Chem Int Ed Engl, 2002. **41**(15): p. 2644-76.
20. Ivanov, A.S., A. Liul'kin Iu, V.S. Skvortsov, and A.B. Rumiantsev, *The rational computer-aided design of new drugs: a review of the methods.* Vestn Ross Akad Med Nauk, 1995(12): p. 51-6.
21. Rarey, M. and J.S. Dixon, *Feature trees: a new molecular similarity measure based on tree matching.* J Comput Aided Mol Des, 1998. **12**(5): p. 471-90.
22. Marialke, J., S. Tietze, and J. Apostolakis, *Similarity Based Docking.* J Chem Inf Model, 2007.
23. Verdonk, M.L., J.C. Cole, and R. Taylor, *SuperStar: a knowledge-based approach for identifying interaction sites in proteins.* J Mol Biol, 1999. **289**(4): p. 1093-108.
24. Fradera, X., R.M. Knegtel, and J. Mestres, *Similarity-driven flexible ligand docking.* Proteins, 2000. **40**(4): p. 623-36.

25. Hindle, S.A., M. Rarey, C. Buning, and T. Lengaue, *Flexible docking under pharmacophore type constraints.* J Comput Aided Mol Des, 2002. **16**(2): p. 129-49.
26. Cross, S.S., *Improved FlexX docking using FlexS-determined base fragment placement.* J Chem Inf Model, 2005. **45**(4): p. 993-1001.
27. Breu, S., Gohlke, *Consensus Adaption of Fields for Molecular Comparison (AFMoC) Models Incorporate Ligand and Receptor Conformational Variability into Tailor-made Scoring Functions.* J Chem Inf Model, 2007.
28. Radestock, S., M. Bohm, and H. Gohlke, *Improving binding mode predictions by docking into protein-specifically adapted potential fields.* J Med Chem, 2005. **48**(17): p. 5466-79.
29. Pfeffer, P. and H. Gohlke, *DrugScore(RNA)-Knowledge-Based Scoring Function To Predict RNA-Ligand Interactions.* J Chem Inf Model, 2007. **47**(5): p. 1868-76.
30. Chuaqui, C., Z. Deng, and J. Singh, *Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening.* J Med Chem, 2005. **48**(1): p. 121-33.
31. Deng, Z., C. Chuaqui, and J. Singh, *Knowledge-based design of target-focused libraries using protein-ligand interaction constraints.* J Med Chem, 2006. **49**(2): p. 490-500.
32. Mpamhanga, C.P., B. Chen, I.M. McLay, and P. Willett, *Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions.* J Chem Inf Model, 2006. **46**(2): p. 686-98.
33. Renner, S., S. Derksen, S. Radestock, and F. Morchen, *Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints.* J Chem Inf Model, 2008. **48**(2): p. 319-32.
34. Kelly, M.D. and R.L. Mancera, *Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design.* J Chem Inf Comput Sci, 2004. **44**(6): p. 1942-51.
35. Marcou, G. and D. Rognan, *Optimizing fragment and scaffold docking by use of molecular interaction fingerprints.* J Chem Inf Model, 2007. **47**(1): p. 195-207.

36. Garrett M. Morris, D.S.G., Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, Arthur J. Olson, *Automated Docking using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function.* Journal of Computational Chemistry, 1999. **19**(14): p. 1639-1662.
37. Wang, R., X. Fang, Y. Lu, and S. Wang, *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.* J Med Chem, 2004. **47**(12): p. 2977-80.
38. Verdonk, M.L., J.C. Cole, M.J. Hartshorn, C.W. Murray, and R.D. Taylor, *Improved protein-ligand docking using GOLD.* Proteins, 2003. **52**(4): p. 609-23.
39. CCG, C.C.G.I., *MOE (Molecular Operating Environment).* 2005.
40. Gerber, P.R. and K. Muller, *MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry.* J Comput Aided Mol Des, 1995. **9**(3): p. 251-68.
41. Durant, J.L., B.A. Leland, D.R. Henry, and J.G. Nourse, *Reoptimization of MDL keys for use in drug discovery.* J Chem Inf Comput Sci, 2002. **42**(6): p. 1273-80.
42. Lipinski, C.A., F. Lombardo, B.W. Dominy, and P.J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.
43. Verdonk, M.L., V. Berdini, M.J. Hartshorn, W.T. Mooij, C.W. Murray, R.D. Taylor, and P. Watson, *Virtual screening using protein-ligand docking: avoiding artificial enrichment.* J Chem Inf Comput Sci, 2004. **44**(3): p. 793-806.
44. Reuter, K., Ficner, R., *Sequence analysis and overexpression of the Zymomonas mobilis tgt gene encoding tRNA-guanine transglycosylase: purification and biochemical characterization of the enzyme*. J Bacteriol, 1995. **177**(18): p. 5284-288.
45. Romier, C., Reuter, K., Suck, D., Ficner, R., *Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange*. Embo J, 1996. **15**(11): p. 2850-857
46. Zhao, Y. and G. Karypis, *Data clustering in life sciences.* Mol Biotechnol, 2005. **31**(1): p. 55-80.

47. Downs, F., *Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data.* J Chem Inf Comput Sci, 1994. **34**: p. 1094-1102.

48. Hopkins, A.L., Groom, C.R. and Alex, A., *Ligand efficiency: a useful metric for lead selection.* Drug Discov Today, 2004. **9**(10): p. 430-1.

49. Holmes, M. A.; Matthews, B. W. Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis. *Biochemistry* **1981,** 20, 6912-20.

50. Monzingo, A. F.; Matthews, B. W. Binding of N-carboxymethyl dipeptide inhibitors to thermolysin determined by X-ray crystallography: a novel class of transition-state analogues for zinc peptidases. *Biochemistry* **1984,** 23, 5724-9.

51. Sturzebecher, J., Sturzebecher, U. et al. (1989). "Synthetic inhibitors of bovine factor Xa and thrombin comparison of their anticoagulant efficiency." Thromb Res **54**(3): 245-52.

52. Dixon, M., The determination of enzyme inhibitor constants. Biochemical Journal, 1953. 55(1): p. 2.

53. Weimer, S., K. Oertel, and H.L. Fuchsbauer, *A quenched fluorescent dipeptide for assaying dispase- and thermolysin-like proteases.* Anal Biochem, 2006. **352**(1): p. 110-9.

54. Leatherbarrow, R.J., *GraFit Version 4*. 1998, Erithacus Software Limited: Staines, UK.

55. Grädler, U., Gerber, H.D., Goodenough-Lashua, D.M., Garcia, G.A., Ficner, R., Reuter, K., Stubbs, M.T., Klebe, G., *A New Target for Shigellosis: Rational Design and Crystallographic Studies of Inhibitors of tRNA-guanine Transglycosylase*. J Mol Biol 2001. **306**(3): p. 455-67

56. Meyer, E., Donati, N., Guillot, M., Schweizer, B., Diederich, F., Stengl, B., Brenk, R., Reuter, K., Klebe, G., *Synthesis, Biological Evaluation, and Crystallographic Studies of Extended Guanine-Based (lin-Benzoguanine) Inhibitors for tRNA-Guanine Transglycosylase (TGT)*. Helvetica Chimica Acta 2006. **89**(4): p. 573-97

57. Holmes, M. A.; Matthews, B. W. Structure of thermolysin refined at 1.6 A resolution. *J Mol Biol* **1982,** 160, 623-39.

58. Z. Otwinowski, W. Minor. In *Methods Enzymol.*, Carter, C. W., Jr., Ed. Academic Press: 1997; Vol. 276, pp 307-326.

59. Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **1998,** 54, 905-21.

60. G. M. Sheldrick, T. R. Schneider. In *Methods Enzymol.*, W. C. J. Charles, M. S. Robert, Ed. Academic Press: 1997; Vol. 277, pp 319-343.

61. Emsley, P.; Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **2004,** 60, 2126-32.

62. Weininger, D., *SMILES, a chemical language and information system.* Chem Des Autom News, 1986. **1**: p. 13.

63. Weininger, D., Weininger, A., Weininger, J. L., *SMILES. 2. Algorithm for Generation of Unique SMILES Notation.* J Chem Inf Comput Sci, 1989. **30**: p. 3.

64. Czodrowski, P., I. Dramburg, C.A. Sotriffer, and G. Klebe, *Development, validation, and application of adapted PEOE charges to estimate pKa values of functional groups in protein-ligand complexes.* Proteins, 2006. **65**(2): p. 424-37.

65. Huey, R., G.M. Morris, A.J. Olson, and D.S. Goodsell, *A semiempirical free energy force field with charge-based desolvation.* J Comput Chem, 2007. **28**(6): p. 1145-52.

66. Gasteiger, J., Marsili, M., *Iterative Partial Equalization of Orbital Electronegativity - A rapid Access to Atom Charges.* Tetrahedron, 1980. **36**: p. 9.

67. Morris, G., D. Goodsell, R. Halliday, R. Huey, W. Hart, R. Belew, and A. Olson, *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1999. **19**(14): p. 1639-1662.

68. Sadowski, J., Gasteiger, J., *From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders.* Chem Reviews, 1993. **93**: p. 14.

69. Kairys, V., M.X. Fernandes, and M.K. Gilson, *Screening drug-like compounds by docking to homology models: a systematic study.* J Chem Inf Model, 2006. **46**(1): p. 365-79.
70. Nelder, J.A., Mead, R., *A Simplex Method for Function Minimization.* Comp J, 1965. **7**: p. 5.
71. Leary, R.H., *Global Optimization on Funneling Landscapes.* Journal of Global Optimization, 2000. **18**: p. 16.
72. Metropolis, N. and S. Ulam, *The Monte Carlo method.* J Am Stat Assoc, 1949. **44**(247): p. 335-41.
73. Kirkpatrick, S., C.D. Gelatt, Jr., and M.P. Vecchi, *Optimization by Simulated Annealing.* Science, 1983. **220**(4598): p. 671-680.
74. Price, S., Lampinen, *Differential Evolution, A Practical Approach to Global Optimization*. 2005: Springer Verlag.
75. Eberhart, R., Kennedy, J., *Particle Swarm Optimization.* IEEE Press, 1995: p. 6.
76. Baluja, S., Caruana, R. *PBIL - Population Based Incremental Learning*. in *MACHINE LEARNING-INTERNATIONAL WORKSHOP*. 1995.
77. Rechenberg, I., *Evolution Strategy in Computational Intelligence.* IEEE Press, 1994.
78. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*. 1989: Addison-Wesley.
79. Wolpert, D.H., Macready, W. G., *No free lunch theorems for optimization.* IEEE Transactions on Evolutionary Computation, 1997. **1**: p. 15.
80. Schneider, G. and U. Fechner, *Computer-based de novo design of drug-like molecules.* Nat Rev Drug Discov, 2005. **4**(8): p. 649-63.
81. Lauri, G. and P. Bartlett, *CAVEAT: A program to facilitate the design of organic molecules.* Journal of Computer-Aided Molecular Design, 1994. **8**(1): p. 51-66.
82. Law, J.M.S., D.Y.K. Fung, Z. Zsoldos, A. Simon, Z. Szabo, I.G. Csizmadia, and A.P. Johnson, *Validation of the SPROUT de novo design program.* THEO CHEM, 2003. **8463**.
83. Sun, Y., T.J.A. Ewing, A.G. Skillman, and I.D. Kuntz, *CombiDOCK: Structure-based combinatorial docking and library design.* Journal of Computer-Aided Molecular Design, 1998. **12**(6): p. 597-604.

84. Bohm, H.-J., D. Banner, and L. Weber, *Combinatorial Docking and combinatorial chemistry: Design of potent non-peptide thrombin inhibitors.* J Comput Aided Mol Des, 1999(13): p. 51-56.
85. Gastreich, M., M. Lilienthal, H. Briem, and H. Claussen, *Ultrafast de novo docking combining pharmacophores and combinatorics.* J Comput Aided Mol Des, 2007.
86. Degen, J. and M. Rarey, *FlexNovo: structure-based searching in large fragment spaces.* ChemMedChem, 2006. **1**(8): p. 854-68.
87. Gerlach, C., M. Munzel, B. Baum, H.D. Gerber, T. Craan, W.E. Diederich, and G. Klebe, *KNOBLE: a knowledge-based approach for the design and synthesis of readily accessible small-molecule chemical probes to test protein binding.* Angew Chem Int Ed Engl, 2007. **46**(47): p. 9105-9.
88. Kuhn, D., N. Weskamp, S. Schmitt, E. Hullermeier, and G. Klebe, *From the similarity analysis of protein cavities to the functional classification of protein families using cavbase.* J Mol Biol, 2006. **359**(4): p. 1023-44.
89. Proschak, E., K. Sander, H. Zettl, Y. Tanrikulu, O. Rau, P. Schneider, M. Schubert-Zsilavecz, H. Stark, and G. Schneider, *From molecular shape to potent bioactive agents II: fragment-based de novo design.* ChemMedChem, 2009. **4**(1): p. 45-8.
90. Truchon, J.F. and C.I. Bayly, *GLARE: A New Approach for Filtering Large Reagent Lists in Combinatorial Library Design Using Product Properties.* J. Chem. Inf. Model., 2006.
91. Agrafiotis, D.K. and V.S. Lobanov, *Ultrafast algorithm for designing focused combinational arrays.* J Chem Inf Comput Sci, 2000. **40**(4): p. 1030-8.
92. Le Bailly de Tilleghem, C., B. Beck, B. Boulanger, and B. Govaerts, *A fast exchange algorithm for designing focused libraries in lead optimization.* J Chem Inf Model, 2005. **45**(3): p. 758-767.
93. Zheng, W., S.J. Cho, C.L. Waller, and A. Tropsha, *Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: a novel computational tool for universal library design and database mining.* J Chem Inf Comput Sci, 1999. **39**(4): p. 738-46.
94. Good, A.C. and R.A. Lewis, *New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick.* J. Med. Chem., 1997. **40**(24): p. 3926-3936.

95. Tropsha, *Rational Principles of Compound Selection for Combinatorial Library Design.* Combinatorial Chemistry & High Throughput Screening, 2002. **5**(2): p. 111-123.
96. Zheng, W., S.J. Cho, and A. Tropsha, *Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries.* J. Chem. Inf. Model., 1998. **38**(2): p. 251-258.
97. Zheng, W., S.T. Hung, J.T. Saunders, and G.L. Seibel, *PICCOLO: a tool for combinatorial library design via multicriterion optimization.* Pac Symp Biocomput, 2000: p. 588-599.
98. Schneider, G., M.-L. Lee, M. Stahl, and P. Schneider, *De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks.* Journal of Computer-Aided Molecular Design, 2000. **14**(5): p. 487-494.
99. Fechner, U. and G. Schneider, *Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design.* J. Chem. Inf. Model., 2006. **46**(2): p. 699-707.
100. Schuller, A. and G. Schneider, *Identification of hits and lead structure candidates with limited resources by adaptive optimization.* J Chem Inf Model, 2008. **48**(7): p. 1473-91.
101. Singh, J., M.A. Ator, E.P. Jaeger, M.P. Allen, D.A. Whipple, J.E. Soloweij, S. Chowdhary, and A.M. Treasurywala, *Application of Genetic Algorithms to Combinatorial Synthesis: A Computational Approach to Lead Identification and Lead Optimization.* J. Am. Chem. Soc., 1996. **118**(7): p. 1669-1676.
102. Brown, R.D. and Y.C. Martin, *Designing Combinatorial Library Mixtures Using a Genetic Algorithm.* J. Med. Chem., 1997. **40**(15): p. 2304-2313.
103. Sheridan, R.P., S.G. SanFeliciano, and S.K. Kearsley, *Designing targeted libraries with genetic algorithms.* J Mol Graph Model, 2000. **18**(4-5).
104. Westhead, D.R., D.E. Clark, D. Frenkel, J. Li, C.W. Murray, B. Robson, and B. Waszkowycz, *PRO-LIGAND: an approach to de novo molecular design. 3. A genetic algorithm for structure refinement.* J Comput Aided Mol Des, 1995. **9**(2): p. 139-48.
105. Gillet, V.J., W. Khatib, P. Willett, P.J. Fleming, and D.V. Green, *Combinatorial library design using a multiobjective genetic algorithm.* J Chem Inf Comput Sci, 2002. **42**(2): p. 375-85.

106. Douguet, D., E. Thoreau, and G.Â.r. Grassy, *A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm.* Journal of Computer-Aided Molecular Design, 2000. **14**(5): p. 449-466.
107. Dey, F. and A. Caflisch, *Fragment-based de novo ligand design by multiobjective evolutionary optimization.* J Chem Inf Model, 2008. **48**(3): p. 679-90.
108. Nicolaou, C.A., J. Apostolakis, and C.S. Pattichis, *De Novo Drug Design Using Multiobjective Evolutionary Graphs.* J Chem Inf Model, 2009.
109. Vinkers, H.M., M.R. de Jonge, F.F. Daeyaert, J. Heeres, L.M. Koymans, J.H. van Lenthe, P.J. Lewi, H. Timmerman, K. Van Aken, and P.A. Janssen, *SYNOPSIS: SYNthesize and OPtimize System in Silico.* J Med Chem, 2003. **46**(13): p. 2765-73.
110. Belda, I., S. Madurga, X. Llora, M. Martinell, T. Tarrago, M. Piqueras, E. Nicolas, and E. Giralt, *ENPDA: an evolutionary structure-based de novo peptide design algorithm.* Journal of Computer-Aided Molecular Design, 2005. **19**(8): p. 585-601.
111. Pegg, S.C., J.J. Haresco, and I.D. Kuntz, *A genetic algorithm for structure-based de novo design.* J Comput Aided Mol Des, 2001. **15**(10): p. 911-933.
112. Sadowski, J., Schwab, C. H., Gasteiger, J., *CORINA, 3D Structure Generator*: Erlangen, Germany.
113. Back, T., *Evolutionary Algorithms in Theory and Practise*. 1996: Oxford University Press.
114. Deb, K. and H.G. Beyer, *Self-adaptive genetic algorithms with simulated binary crossover.* Evol Comput, 2001. **9**(2): p. 197-221.
115. Bartz-Beielstein, T., *Experimental Research in Evolutionary Computation - The New Experimentalism*. Natural Computing Series. 2006, Heidelberg: Springer Verlag.
116. Kick, E.K., D.C. Roe, A.G. Skillman, G. Liu, T.J. Ewing, Y. Sun, I.D. Kuntz, and J.A. Ellman, *Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D.* Chem Biol, 1997. **4**(4): p. 297-307.
117. Hans-Georg Beyer, H.-P.S., *Evolution Strategies - A comprehensive introduction.* Natural Computing, 2002. **1**(1): p. 3-52.

118. Harris, J.L., B.J. Backes, F. Leonetti, S. Mahrus, J.A. Ellman, and C.S. Craik, *Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries.* Proc Natl Acad Sci U S A, 2000. **97**(14): p. 7754-9.

119. MGLTools v1.5.2, http://mgltools.scripps.edu/.

120. CCDC, http://ccdc.cam.ac.uk/products/life_sciences/gold/case_studies/gold_validation_virtual_screening.

121. The Condor Queuing System, http://www.cs.wisc.edu/condor/.

122. Linusson, A., J. Gottfries, T. Olsson, E. Ornskov, S. Folestad, B. Norden, and S. Wold, *Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors.* J Med Chem, 2001. **44**(21): p. 3424-39.

123. Bode, W., I. Mayr, U. Baumann, R. Huber, S.R. Stone, and J. Hofsteenge, *The refined 1.9 A crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment.* EMBO J, 1989. **8**(11): p. 3467-75.

124. Mathews, II, K.P. Padmanabhan, V. Ganesh, A. Tulinsky, M. Ishii, J. Chen, C.W. Turck, S.R. Coughlin, and J.W. Fenton, 2nd, *Crystallographic structures of thrombin complexed with thrombin receptor peptides: existence of expected and novel binding modes.* Biochemistry, 1994. **33**(11): p. 3266-79.

125. Sall, D.J., J.A. Bastian, S.L. Briggs, J.A. Buben, N.Y. Chirgadze, D.K. Clawson, M.L. Denney, D.D. Giera, D.S. Gifford-Moore, R.W. Harper, K.L. Hauser, V.J. Klimkowski, T.J. Kohn, H.S. Lin, J.R. McCowan, A.D. Palkowitz, G.F. Smith, K. Takeuchi, K.J. Thrasher, J.M. Tinsley, B.G. Utterback, S.C. Yan, and M. Zhang, *Dibasic benzo[b]thiophene derivatives as a novel class of active site-directed thrombin inhibitors. 1. Determination of the serine protease selectivity, structure-activity relationships, and binding orientation.* J Med Chem, 1997. **40**(22): p. 3489-93.

126. Malikayil, J.A., J.P. Burkhart, H.A. Schreuder, R.J. Broersma, Jr., C. Tardif, L.W. Kutcher, 3rd, S. Mehdi, G.L. Schatzman, B. Neises, and N.P. Peet, *Molecular design and characterization of an alpha-thrombin inhibitor containing a novel P1 moiety.* Biochemistry, 1997. **36**(5): p. 1034-40.

127. Riester, D., F. Wirsching, G. Salinas, M. Keller, M. Gebinoga, S. Kamphausen, C. Merkwirth, R. Goetz, M. Wiesenfeldt, J. Sturzebecher, W. Bode, R. Friedrich, M. Thurk, and A. Schwienhorst, *Thrombin inhibitors identified by computer-assisted multiparameter design.* Proc Natl Acad Sci U S A, 2005. **102**(24): p. 8597-602.

128. van de Locht, A., D. Lamba, M. Bauer, R. Huber, T. Friedrich, B. Kroger, W. Hoffken, and W. Bode, *Two heads are better than one: crystal structure of the insect derived double domain Kazal inhibitor rhodniin in complex with thrombin.* EMBO J, 1995. **14**(21): p. 5149-57.

5 Zusammenfassung

In dieser Arbeit sind die zwei neuen Computer-Methoden DrugScore Fingerprint (DrugScore^{FP}) und GARLig in ihrer Theorie und Funktionsweise vorgestellt und validiert worden.

DrugScore^{FP} ist ein neuartiger Ansatz zur Bewertung von computergenerierten Bindemodi potentieller Liganden für eine bestimmte Zielstruktur. Das Programm basiert auf der etablierten Bewertungsfunktion DrugScore^{CSD} und unterscheidet sich darin, dass anhand bereits bekannter Kristallstrukturen für den zu untersuchenden Rezeptor ein Referenzvektor generiert wird, der zu jedem Bindetaschenatom Potentialwerte für alle möglichen Interaktionen enthält. Für jeden neuen, computergenerierten Bindungsmodus eines Liganden lässt sich ein entsprechender Vektor generieren. Dessen Distanz zum Referenzvektor ist ein Maß dafür, wie ähnlich generierte Bindungsmodi zu bereits bekannten sind. Eine experimentelle Validierung der durch DrugScore^{FP} als ähnlich vorhergesagten Liganden ergab für die in unserem Arbeitskreis untersuchten Proteinstrukturen Trypsin, Thermolysin und tRNA-Guanin Transglykosylase (TGT) sechs Inhibitoren fragmentärer Größe und eine Thermolysin Kristallstruktur in Komplex mit einem der gefundenen Fragmente.

Das in dieser Arbeit entwickelte Programm GARLig ist eine auf einem Genetischen Algorithmus basierende Methode, um chemische Seitenkettenmodifikationen niedermolekularer Verbindungen hinsichtlich eines untersuchten Rezeptors effizient durchzuführen. Zielsetzung ist hier die Zusammenstellung einer Verbindungsbibliothek, welche eine benutzerdefiniert große Untermenge aller möglichen chemischen Modifikationen Ligand-ähnlicher Grundgerüste darstellt. Als zentrales Qualitätskriterium einzelner Vertreter der Verbindungsbibliothek dienen durch Docking erzeugte Ligand-Geometrien und deren Bewertungen durch Protein-Ligand-Bewertungsfunktionen. In mehreren Validierungsszenarien an den Proteinen Trypsin, Thrombin, Faktor Xa, Plasmin und Cathepsin D konnte gezeigt werden, dass eine effiziente Zusammenstellung Rezeptor-spezifischer Substrat- oder Ligand-Bibliotheken lediglich eine Durchsuchung von weniger als 8% der vorgegebenen Suchräume erfordert und GARLig dennoch im Stande ist, bekannte Inhibitoren in der Zielbibliothek anzureichern.

6 Summary

This thesis describes the theory, development, application and validation of two new chemoinformatic tools to find new drugs named DrugScore^{FP} and GARLig.

DrugScore^{FP} is an extension of the well-known scoring function DrugScore^{CSD}. The method can be used for rescoring docking solutions by adding structural information from a user-defined set of protein-ligand complexes resulting in a tailor-made protein-specific scoring function. The new method demonstrates significant improvements in finding near-native poses in comparison to 12 established scoring functions. Using the in-house investigated protein structures trypsin, thermolysin and tRNA-guanin transglycosylase (TGT), we identified six fragment-sized molecules which were found to inhibit these targets and one thermolysin crystal structure in complex with one of the predicted fragments.

GARLig is a library design tool based on docking and a self-adaptive genetic algorithm for structure-based sidechain-optimization of small molecule skeletons. Multiple scoring functions such as AutoDock4 Score, GOLDScore and DrugScore^{CSD} can be applied to the search procedure as possible decision criteria for potential library candidates. The program has been validated on trypsin, thrombin, factor Xa, plasmin and cathepsin D. GARLig was able to find natural substrates and known binders by validating less than 8% of large combinatorial libraries, meanwhile outperforming other search strategies such as a random search and a Monte Carlo Sampling.

7 Erklärung

Ich versichere, dass ich meine Dissertation

„Optimiertes Design kombinatorischer Verbindungsbibliotheken durch Genetische Algorithmen und deren Bewertung anhand wissensbasierter Protein-Ligand Bindungsprofile“

selbständig, ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich gekennzeichneten Quellen und Hilfen bedient habe. Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen Prüfungszwecken gedient.

Marburg, den

(Patrick Pfeffer)

8 Veröffentlichungen, Vorträge und Posterbeiträge

Aufsätze:

P. Pfeffer, G. Neudert, T. Ritschel, L. Englert, B. Baum, and G. Klebe

DrugScore^{FP} – Profiling Protein-Ligand Interactions using Fingerprint Simplicity paired with Knowledge-Based Potential Fields

Manuskript in Vorbereitung

P. Pfeffer, T. Fober, E. Hüllermeier, and G. Klebe

GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm

Eingereicht

Vorträge:

DrugScore Fingerprint – Profiling Protein-Ligand Interactions

German Conference on Cheminformatics (GDCH-Tagung)

Tagungszentrum, Goslar / Deutschland

DrugScore^{FP} – Using Fingerprint Simplicity Paired with Knowledge-based Potential Fields

236th American Chemical Society (ACS) National Meeting

Convention Center, Philadelphia / USA

Posterbeiträge:

GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm

Molecular Graphics and Modelling Society

Friedrich-Alexander Universität, Erlangen / Deutschland

GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm

German Conference on Cheminformatics (GDCH-Tagung)

Tagungszentrum, Goslar / Deutschland

Gewinner des Posterpreises für die Arbeit “GARLig“

GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm

Molecular Graphics and Modelling Society

The Brunai Gallery, London / UK

9 Danksagung

Mein herzlicher Dank gilt:

- in besonderem Maße Herrn Prof. Dr. Gerhard Klebe für die Möglichkeit, unter seiner Anleitung die vorliegende Arbeit anzufertigen. Ihm gelang stets die korrekte Gewichtung aus inspirierender Führung und akademischen Freiheitsgraden

- Besonders Gerd Neudert („Gerd“) für eine sehr gute Zusammenarbeit bei dem DrugScore^{FP} Projekt

- Thomas Fober für eine sehr gute Zusammenarbeit bei dem GARLig Projekt

- Dr. Alexander Hillebrecht („Hille“) für seine Fähigkeit, Menschen jederzeit zum Lachen bringen zu können

- Tina Ritschel („Dino“), Lisa Englert („Liehssa“) und Bernhard Baum („Bernie-Boy“) für die fruchtbare Zusammenarbeit bei dem DrugScore^{FP} Projekt

- Andreas Spitzmüller („Action-Andy“) für manches gemütliche Bier nach der Arbeit

- Martin Sippel für seinen guten Charakter

- Jörg Leonhardt für eine gute Zeit bei JLP

- Meinen Eltern für alles, was sich durch Schriftzeichen nicht darstellen lässt

10 Lebenslauf

Patrick Pfeffer

6, rue des Canettes

75006 Paris

Frankreich

Mobil: +49(0)176 / 62228238

E-Mail: pfefferp@gmail.com

Patrick Pfeffer

Persönliche Informationen	Familienstand: ledig
	Nationalität: deutsch
	Geburtsdatum: 18.11.1980 in Frankfurt am Main
Schulausbildung	
1987-1991	Viktoria Grundschule, Kronberg im Taunus
1991-2000	Altkönig-Schule, Kronberg im Taunus
	Abschluss: Abitur
Studium	
10.2001 – 12.2005	Johann Wolfgang Goethe Universität, Frankfurt
	Fachrichtung: Bioinformatik (Abschluss: Diplom Bioinformatiker)
	Thema der Diplomarbeit:
	§ Entwicklung einer statistischen Bewertungsfunktion zur Vorhersage von RNA-Ligand-Wechselwirkungen
03.2007 – 06.2008	Fernuniversität, Hagen
	Parallelstudium: Betriebswirtschaftslehre (Abschluss: Betriebswirt)
03.2006 – heute	Philipps Universität, Marburg
	Promotion in der Pharm. Chemie (Abschluss: Dr. rer. nat.)
	Thema der Dissertation:
	Optimiertes Design kombinatorischer Verbindungsbibliotheken durch Genetische Algorithmen und deren Bewertung anhand wissensbasierter Protein-Ligand Bindungsprofile