Leitstrukturoptimierung mit Hilfe von Matched Molecular Pairs im Kontext der Rezeptorumgebung

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich

Biochemie, Chemie und Pharmazie

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Julia Weber

aus Rüsselsheim

Frankfurt am Main (2015)

(D30)

vom Fachbereich Biochemie, Chemie und Pharmazie der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:	Prof. Dr. Michael Karas
Gutachter:	Jun.-Prof. Dr. Ewgenij Proschak
	Prof. Dr. Stefan Knapp
Datum der Disputation:	

Meinen Eltern

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abkürzung	Bedeutung
3D	Dreidimensional
2D	Zweidimensional
ADMET	Absorption, Verteilung, Metabolismus, Exkretion, Toxizität
CDK	Engl. Chemistry Development Kit
CDK2	Cyclin-abhängige Kinase 2
COX-2	Cyclooxygenase-2
Da	Dalton
EGFR	Engl. Epidermal Growth Factor Receptor
EPAS1	Engl. Endothelial PAS domain-containing protein 1
H-Brücke	Wasserstoffbrücke
hERG	Engl. human Ether-a-go-go Related Gene
HTS	Engl. High-Throughput Screening
IC_{50}	Mittlere inhibitorische Konzentration
K_d	Dissoziationskonstante
K_i	Inhibitionskonstante
KNIME	Engl. Konstanz Information Miner
LogP	Octanol/Wasser-Verteilungskoeffizient
LORD	Engl. Lead Optimization by Rational Design
LORD_FP	Engl. LORD Fingerprint
MCS	Engl. Maximum Common Substructure
MMP	Engl. Matched Molecular Pair
MOE	Engl. Molecular Operating Environment
MW	Molekulargewicht

NMR	Engl. Nuclear Magnetic Resonance
P38α	Mitogen-aktivierte Proteinkinase 14
PDB	Engl. Protein Data Bank
QSAR	Quantitative Struktur-Aktivitätsbeziehung
QSPR	Quantitative Struktur-Eigenschaftsbeziehung
SAR	Struktur-Aktivitätsbeziehung
SDF	Engl. Structure Data Format
SMARTS	Engl. Smiles Arbitrary Target Specification
SMILES	Engl. Simplified Molecular Input Line Entry System
VAMMPIRE	Engl. Virtually Aligned Matched Molecular Pairs Including Receptor Environment

1 Einleitung

1.1 Strukturbasierte Wirkstoffentwicklung

Unter strukturbasierter Wirkstoffentwicklung versteht man die Berücksichtigung der dreidimensionalen (3D) Struktur eines Zielmoleküls (im Folgenden als Target bezeichnet) im Entwicklungsprozess eines Wirkstoffes. Bei einem Wirkstoff-Target handelt es sich in den meisten Fällen um Proteine, Enzyme, Rezeptoren, Ionenkanäle oder Transportproteine, wobei auch Nukleinsäuren und Ribosomen adressiert werden können.[1–3] Ein geeignetes Target für die strukturbasierte Wirkstoffentwicklung ist ein Makromolekül, welches in Verbindung mit einem humanen Krankheitsbild gebracht wird und dessen Funktion durch den Einsatz eines kleinen Moleküls, dem Liganden, moduliert werden kann.[4,5]

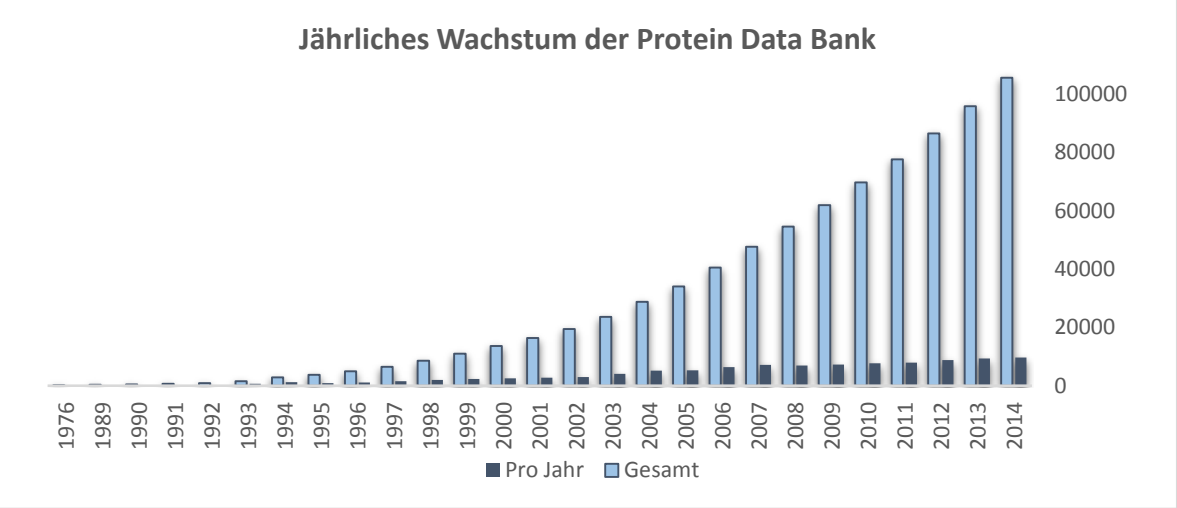


Abbildung 1. Jährliches Wachstum der Protein Data Bank. Anzahl der in ihrer Struktur aufgeklärten Makromoleküle gesamt (hellblau) und pro Jahr (dunkelblau). Statistik von http://www.rcsb.org/ (Stand: Dezember 2014).

In den vergangenen 40 Jahren war ein exponentieller Anstieg publizierter 3D-Strukturen zu beobachten, wodurch auch die strukturbasierte Wirkstoffentwicklung immer mehr an Bedeutung gewann.[6–8] In **Abbildung 1** ist dieses Wachstum am Beispiel der Protein Data Bank

(PDB)[9] dargestellt, welche eine Sammlung dieser, meist durch Röntgenstrukturanalyse oder Kernspinresonanzspektroskopie (Abk. NMR-Spektroskopie, engl. nuclear magnetic resonance) aufgeklärten, makromolekularen Komplexe enthält. Ein Grund für dieses stetige Wachstum ist unter anderem der Erfolg bei der rationalen Entwicklung und Optimierung von Wirkstoffen durch Fragmentbasierte Ansätze[10–15] und der dadurch entstandene Fokus auf die Entwicklung von Proteinexpressionssystemen und den Aufbau von effizienten Kristallisations-Einrichtungen.[16–23]

Sowohl bei der Identifizierung potentieller Leitstrukturen als auch für die anschließende Leitstrukturoptimierung ist es von großem Vorteil die Struktur des Targets zu kennen, da sie Aufschluss über Größe und Form der Bindetasche, aber auch über potentielle Interaktionen des Targets mit dem Wirkstoff gibt. Die Identifizierung und Validierung eines solchen krankheitsrelevanten Targets sowie die Aufklärung des tatsächlichen Einflusses des Targets am untersuchten Krankheitsbild steht am Anfang der Entwicklung eines jeden Wirkstoffkandidaten.[24–26] Anschließend gilt es aus einer ausgewählten Moleküldatenbank erste Liganden zu identifizieren, welche das Target konzentrationsabhängig modulieren. Klassischerweise werden diese sogenannten „Hits“ (engl. für Treffer) in biologischen oder biophysikalischen Hochdurchsatz-Testverfahren (Abk. HTS, engl. high throughput screening) generiert.[26–28] Nicht selten werden jedoch computergestützte Ansätze vorangestellt, um den chemischen Raum einzugrenzen und lediglich eine Auswahl von vielversprechenden Kandidaten *in vitro* zu testen.[29,30]

Das bekannteste Verfahren für strukturbasiertes „virtuelles Screening“ ist das molekulare Docking (kurz Docking, dt. Einpassen), bei dem eine direkte Modellierung der Interaktion von Ligand und Target stattfindet. Dabei werden entweder Konformationen des zu modellierenden Liganden erzeugt und direkt in die in die Bindetasche platziert,[31,32] oder der vorab fragmentierte Ligand schrittweise innerhalb der Bindetasche rekonstruiert.[33] Vor der Bewertung des entstandenen Protein-Ligand-Komplexes durch kraftfeldbasierte-, empirische- oder wissensbasierte Bewertungsfunktionen erfolgt gegebenenfalls eine Energieminimierung des platzierten Liganden innerhalb der Bindetasche.[7,34,35] Für die darauf folgende Testung wird klassischerweise eine definierte Anzahl Liganden mit höchster Bewertung ausgewählt, weshalb die Qualität der Bewertungsfunktion einen limitierenden Faktor beim Docking darstellt. Li et al. (2014)[36,37] konnten in einer aktuellen Studie an ko-

kristallisierten Liganden mit experimentell bestimmter Bindungsaffinität zeigen, dass je nach verwendeter Methode die Reproduzierbarkeit der bioaktiven Konformation in 60 - 80 % der Fälle gelingt, während die direkte Vorhersage von Bindungsaffinitäten unterschiedlicher Moleküle immer noch ein großes Problem darstellt.

Die Aufklärung der 3D-Struktur eines Targets ist nicht immer möglich. Eine große Herausforderung stellt beispielsweise die Kristallisation der nicht-löslichen Membranproteine dar.[38,39] Ist die 3D-Struktur eines Targets nicht bekannt, kommen Ligandenbasierte Modelle zum Einsatz, vorausgesetzt es ist mindestens ein Ligand bekannt, der das Target in gewünschtem Maße moduliert.[40] Auf der Basis bekannter Liganden eines Targets können anschließend Moleküldatenbanken nach ähnlichen Liganden durchsucht werden. Dazu werden die bekannten Liganden zunächst durch eine oder mehrere Moleküleigenschaften beschrieben. Diese Beschreibung eines Moleküls wird auch als Molekül-Deskriptor bezeichnet und kann sehr allgemein gehalten (z. B. durch die Molekülmasse oder den Octanol/Wasser-Verteilungskoeffizienten), aber auch sehr spezifisch (z. B. durch die paarweise räumliche Anordnung von Atomen oder Atomtypen) sein. Die mathematische Darstellung einer Reihe molekularer Eigenschaften eines Moleküls in Form eines Vektors ermöglicht den effizienten Vergleich verschiedener Moleküle und die Quantifizierung ihrer Ähnlichkeit anhand der Distanz innerhalb des gewählten Vektorraumes.[41–45]

Ist eine Vielzahl unterschiedlicher Liganden für ein Target bekannt, können gezielt Vorhersagemodelle zur Abschätzung der Aktivität eines unbekannten Liganden generiert werden. Bei der Erstellung von quantitativen Struktur-Aktivitätsbeziehungen (Abk. QSAR, engl. quantitative structure-activity relationship)[46,47] werden Regressionsverfahren genutzt, um eine Korrelation zwischen strukturellen Merkmalen und der Aktivität eines Moleküls herzustellen. Dazu wird die gewünschte Molekül-Eigenschaft als Funktion von strukturellen Deskriptoren dargestellt, die das Molekül repräsentieren. Auf der Basis eines Trainingsdatensatzes, welcher experimentelle Daten zu einer Auswahl von Referenzmolekülen enthält, wird schließlich ein Vorhersagemodell generiert, um neuartige Liganden zu identifizieren.

Die im virtuellen Screening generierten Hits werden anschließend *in vitro* evaluiert, um vielversprechende Kandidaten für die Leitstrukturoptimierung herauszufiltern.[26,30,48] Eine

gängige Herangehensweise für die Optimierung von substituierten aromatischen Leitstrukturen ist beispielsweise die Synthese entlang des sogenannten „Topliss-Schemas“.[49] Dabei handelt es sich um ein organisiertes Flussdiagramm, welches eine schrittweise Synthesestrategie von Analoga eines Liganden nahelegt, um dessen biologische Aktivität zu optimieren. Dabei sind die Hydrophobizitätskonstante π, die Hammett-Konstante σ für elektronische Charakteristiken und der Taft's-Faktor für sterische Eigenschaften eines Substituenten ausschlaggebend. Je nachdem ob die Aktivität eines Analogons unverändert bleibt oder eine Steigerung oder Erniedrigung der Aktivität beobachtet wird, ist eine Anpassung der Synthesestrategie vorgesehen. Dabei werden für aromatische und aliphatische Substitutionen unterschiedliche Schemata angewendet.

Neben der Optimierung der Aktivität eines Liganden werden spätestens im Zuge der Leitstruktur-Optimierung auch die Eigenschaften Absorption, Verteilung, Metabolismus, Exkretion und Toxizität (Abk. ADMET, engl. absorption, distribution, metabolism, excretion, toxicity) berücksichtigt. Der Einsatz von „Bioisosteren“ ist eine häufig angewendete Methode um durch den Austausch von spezifischen chemischen Gruppen die biologische Aktivität eines Liganden beizubehalten oder sogar zu erhöhen und gleichzeitig, insbesondere im Hinblick auf die klinische Relevanz, dessen pharmakokinetische Eigenschaften zu verbessern.[50,51] Unter Bioisosteren versteht man Substituenten, die sich in Bezug auf elektronische und sterische Eigenschaften ähnlich sind und aus diesem Grund zu einer vergleichbaren Bioaktivität eines Moleküls führen.[52,53] Es gibt eine Vielzahl von Ansätzen zum gezielten Einsatz von bioisosterem Ersatz, wobei auch hier QSAR bzw. QSPR (engl. quantitative structure-property relationship) eine große Rolle spielen.[54–59] Die direkte Vorhersage einer nächsten, bisher noch nicht bekannten vielversprechenden Modifikation eines Moleküls, können diese QSAR- oder QSPR-Modelle allerdings nicht bewerkstelligen.

Mit den stetig wachsenden Informationen über molekulare Modifikationen (im Folgenden als Transformationen bezeichnet) und deren Effekte auf unterschiedliche molekulare Eigenschaften können solche gezielte Vorhersagen realisiert werden.[59,60] Der Einsatz von sogenannten *Matched Molecular Pairs* bietet die Möglichkeit Transformationen und deren Effekte für die Erstellung von Vorhersagemodellen zu nutzen. Matched Molecular Pairs bilden die Basis für die hier vorgestellte Arbeit und werden im Folgenden näher beschrieben.

1.2 Matched Molecular Pairs

Als Matched Molecular Pair (MMP) wird ein Paar von Molekülen bezeichnet, welches sich in genau einer molekularen Transformation unterscheidet.[61] Diese Transformation in Verbindung mit einer sich ändernden Moleküleigenschaft ist eine wertvolle Information, die genutzt werden kann, um Vorhersagemodelle für eine betrachtete Moleküleigenschaft zu erstellen. Im Vergleich zu Methoden wie QSAR und QSPR, bei denen ein globales Modell zur Vorhersage einer Eigenschaft auf der Basis einer Vielzahl bekannter Liganden generiert wird, geht es bei den MMP-basierten Methoden lediglich um die Vorhersage eines Effekts, ausgelöst durch eine molekulare Transformation, weshalb man eine bessere Vorhersagequalität erwartet.[62] Das Wissen über solche Transformationen ist in allen Phasen der Wirkstoffentwicklung nützlich, wird aber im Speziellen im Zuge der Leitstrukturoptimierung eingesetzt.[62–65] In der hier vorgestellten Arbeit wird zwischen 2D- und 3D-MMPs unterschieden. 2D-MMPs beschreiben den Effekt einer molekularen Transformation auf die Eigenschaft eines Moleküls, während 3D-MMPs zusätzlich die (bioaktive) Konformation der Moleküle oder sogar die chemische Umgebung der Transformation im Kontext des Targets berücksichtigen. Aktuelle Anwendungen von 2D- und 3D-MMPs werden im Folgenden zusammengefasst.

1.2.1 2D-Matched Molecular Pairs

Es gibt eine Vielzahl von Anwendungen die sich das Prinzip der MMPs zunutze machen.[59–75] Leach et al. (2006)[61] konnten zum Beispiel zeigen, dass pharmakologisch relevante Eigenschaften in Verbindung mit molekularen Transformationen gebracht werden können. Untersucht wurden hier zum einen die Addition verschiedener Substituenten an aromatische Ringe, zum anderen die Methylierung von Heteroatomen und der Effekt auf Löslichkeit, Plasmaproteinbindung und orale Verfügbarkeit eines Liganden (**Abbildung 2**). Eine Addition von Fluor an einen aromatischen Ring führte beispielsweise in 66 % der Fälle (711 MMPs) zu einer Verschlechterung der Löslichkeit, in 77 % der Fälle (467 MMPs) zu einer Erhöhung der Plasmaproteinbindung in Ratten und in 55 % der Fälle (551 MMPs) zu einer Verbesserung der oralen Verfügbarkeit in einem *in vivo* Ratten-Modell. Die Methylierung eines Amids hingegen führte in 79 % der Fälle (142 MMPs) zu einer Verbesserung der

Löslichkeit sowie zu einer Erniedrigung der Plasmaproteinbindung (78 %, 88 MMPs) und oralen Verfügbarkeit (67 %, 113 MMPs).

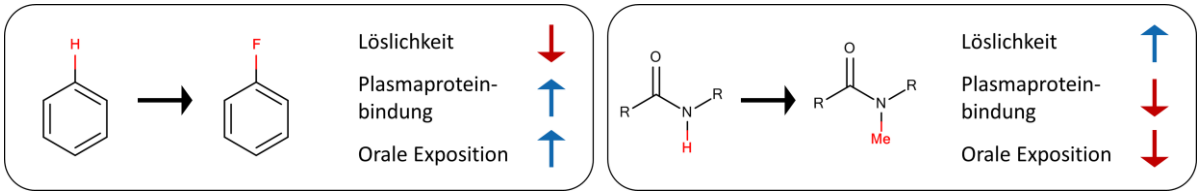


Abbildung 2. Strukturelle Änderungen und ihre Auswirkungen auf pharmakologisch relevante Eigenschaften des Moleküls (nach Leach et al. 2006).[61]

Zhang et al. (2011)[65] konnten ebenfalls mit der Aufstellung einer Fragment-Löslichkeits-Beziehung eine Datenbank für die gezielte Optimierung der Löslichkeit von Molekülen erstellten. Dazu wurden 2.794 chemische Verbindungen mit experimentell bestimmter Wasserlöslichkeit, ausgedrückt als negativer dekadischer Logarithmus des Löslichkeitsprodukts (logS), hierarchisch fragmentiert und MMPs bestimmt. Anschließend wurde der mittlere Effekt (ΔlogS) für sämtliche Substitutionen, Deletionen und Additionen von Fragmenten errechnet. Als Ergebnis wurde eine Substitutionsmatrix erhalten, mit Hilfe derer sich der Median der Auswirkung einer Substitution auf die Löslichkeit eines Moleküls ablesen lässt. In **Abbildung 3** ist die resultierende Matrix in Form einer Wärmekarte dargestellt.

Dass die Berücksichtigung der lokalen chemischen Umgebung einer Transformation (Kontext) zu einer Verbesserung der Vorhersagekraft von MMP-Modellen führen können, zeigten Papadatos et al. (2010)[67] in einer Studie am Beispiel der hERG-Inhibition. Der „Human ether-a`-go-go-related gene“-Ionenkanal (hERG-Kanal) wird mit Herzrhythmusstörungen in Verbindung gebracht und ist deshalb ein wichtiges Antitarget, dessen Inhibition häufig im Zuge der Leitstrukturoptimierung getestet wird. Aus diesem Grund sind große Datensätze mit experimentell bestimmten Aktivitätsdaten verfügbar, welche zur Erstellung von etwa 1,5 Millionen MMPs genutzt wurden. Im Vergleich zwischen der „globalen Verteilung der Transformationseffekte“ auf die Inhibition (ungeachtet des Kontextes in dem die Transformation stattfindet) und der „lokalen Verteilung der Transformationseffekte“ (repräsentiert durch die Ähnlichkeit verschiedener Molekül-Deskriptoren) konnten signifikante Unterschiede festgestellt werden. So zeigt beispielsweise die Addition einer Methoxy-Gruppe in der globalen Verteilung im Mittel keine Tendenz zu erhöhter oder erniedrigter hERG-Inhibition. Vergleicht man allerdings die Addition einer

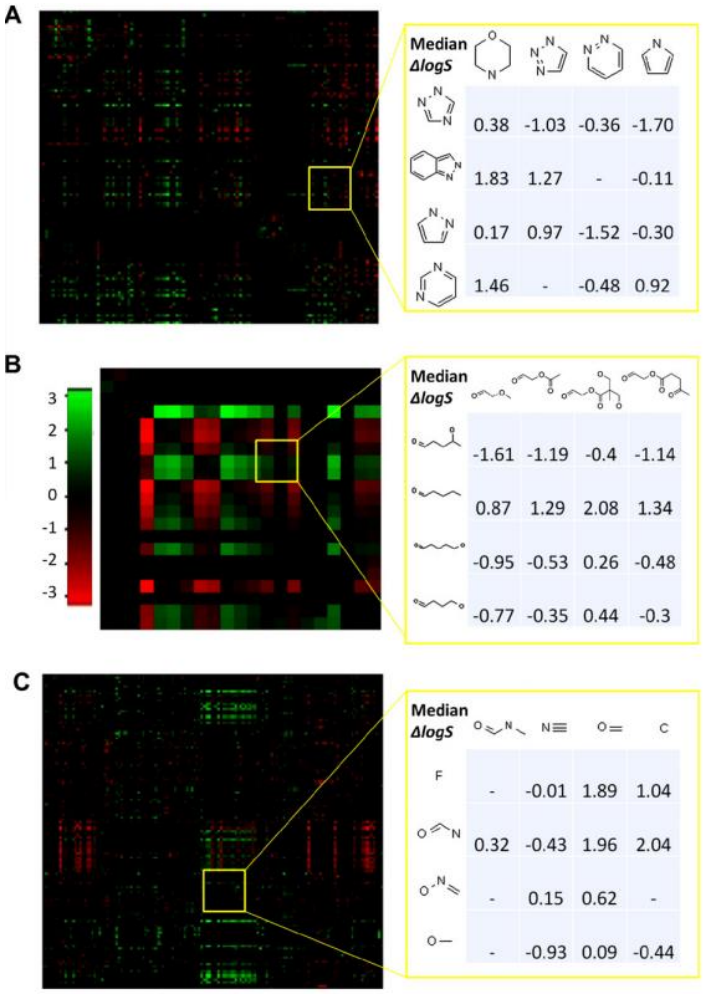


Abbildung 3. Wärmekarte die den Einfluss von chemischen Transformationen auf die Löslichkeit eines Moleküls darstellt. (A) Ring-nach-Ring Transformation; (B) Linker-nach-Linker Transformation; (C) R-nach-R Transformation. Grün steht für einen positiven ΔlogS und rot für einen negativen ΔlogS. Die Tabellen auf der rechten Seite sind Beispiele aus der Wärmekarte. Eine Transformation wird gelesen als Substitution von der ersten Spalte zur ersten Zeile. Die gelben Quadrate auf der Wärmekarte entsprechen nicht zwangsläufig den dargestellten Substitutionen (aus einer Publikation von Zhang et al. 2011[65]).

Methoxy-Gruppe an einen aliphatischen Linker mit der Addition an einen aromatischen Ring, der einen H-Brücken-Akzeptor enthält (z. B. Pyridin), so kann eine deutliche Tendenz beobachtet werden. Das Sauerstoffatom der Methoxy-Gruppe, die an einen aliphatischen Linker gebunden ist, wird stark polarisiert und erhöht die Wahrscheinlichkeit für eine H-Brücken-Bindung, während die Bindung an einen Pyridin-Ring die Polarität des Sauerstoffs erniedrigt und die Lipophilie der Methylgruppe erhöht. Im Falle des Pyridin-Rings führt die Addition der Methoxy-Gruppe überwiegend zu einer Erhöhung der hERG-Inhibition,

während ein gegenteiliger Effekt bei der Addition an einen aliphatischen Linker zu beobachten ist.

Eine Abstraktion der klassischen MMPs, die sogenannten „Fuzzy Matched Pairs“, sollen den Einfluss eines Pharmakophors auf molekulare Interaktionen aufzeigen.[63] Dazu wurden den Atomen oder funktionellen Gruppen eines Moleküls pharmakophore Eigenschaften zugewiesen, welche potentielle Interaktionen des Moleküls mit dem Target beschreiben sollen. Dazu gehören zum Beispiel Wasserstoffbrücken-Donoren und Akzeptoren sowie aromatische- und aliphatische Ringe. Als Ergebnis wird ein zusammenhängender Graph erhalten, welcher schließlich eine Pharmakophor-Abstraktion des eingegebenen Moleküls darstellt (**Abbildung 4**).

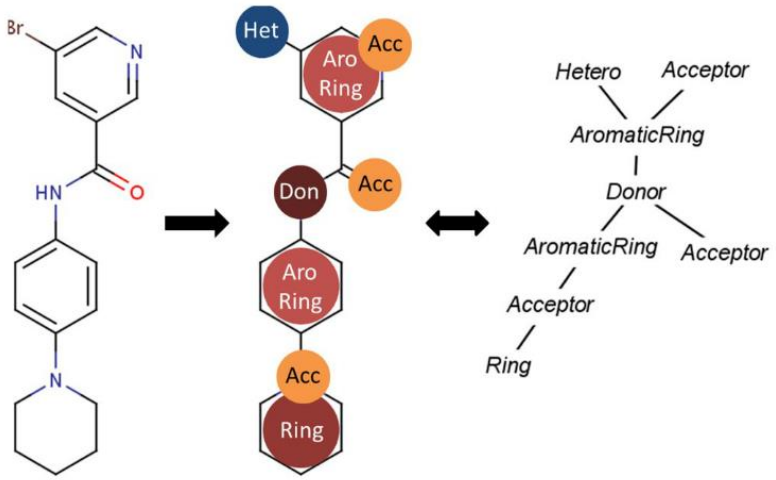


Abbildung 4. Exemplarische Typisierung eines Moleküls. Dargestellt ist die Zuweisung von Pharmakophor-Typen zu einem Molekül aus der ChEMBL Datenbank (ID: CHEMBL1333282).[76] Durch die Zuweisung der Pharmakophor-Typen entsteht eine Abstraktion des Moleküls in Form eines zusammenhängenden Graphen (aus einer Publikation von Geppert und Beck 2013[63]).

Erst nach der Typisierung des Moleküls werden MMPs bestimmt. Dies hat den Vorteil, dass sich die Auftrittshäufigkeit eines MMPs im Vergleich zu den bisher beschriebenen Methoden erhöht und die Anzahl unterschiedlicher Transformationen eingeschränkt wird. Da die Anzahl identischer Transformationen in einem Datensatz oft einen limitierenden Faktor für die statistische Analyse der MMPs darstellt, Können die *Fuzzy Matched Pairs* als alternative Methode angewendet werden.

Die Verbindung von MMP-Methoden mit nicht-linearen Regressionsverfahren wurde in einer aktuellen Publikationen von De la Vega de León & Bajorath (2014)[73] bzw. Beck &

Springer (2014)[71] vorgestellt. Dabei werden die Moleküle eines MMPs durch unterschiedliche Deskriptoren beschrieben und ein Differenzvektor gebildet, welcher indirekt die Transformation eines MMPs beschreiben soll. Auf Basis des Differenzvektors und des korrespondierenden Transformationseffekts wurden anschließend Regressionsmodelle zur Vorhersage von Transformationseffekten generiert. An heterogenen Datensätzen verschiedener Targets konnte gezeigt werden, dass die Qualität der Modelle mit klassischen QSAR-Methoden mithalten und sie teilweise sogar übertreffen kann.

1.2.2 3D-Matched Molecular Pairs

Die Hinzunahme der 3D-Konformation eines Liganden ist ein essentieller Faktor für die Vorhersage der Affinität zu einem spezifischen Target. Zum Zeitpunkt der Anfertigung dieser Arbeit waren lediglich zwei MMP-Methoden bekannt, welche die 3D-Konformation eines Liganden berücksichtigten. Posy et al. (2013)[68] zeigten am Beispiel der Mitogen-aktivierten Proteinkinase 14 (p38*α*), dass 3D-MMPs, generiert aus einer SAR bekannter p38*α* Liganden, zu neuartigen Hybrid-Liganden führen können. Dazu wurden zunächst Modelle von 4.291 P38*α* Protein-Ligand-Komplexen erstellt und übereinander gelagert. Ein Modell konnte genau dann generiert werden wenn die bioaktive Konformation des Liganden aus Kristallstrukturen anderer Kinasen bereits bekannt war. Anschließend wurden die Liganden fragmentiert und eine MMP-Datenbank für sämtliche Transformationen mit einer räumlichen Distanz von höchstens 1 Å erstellt. Durch diese Methode konnten beispielsweise Regionen innerhalb des Rezeptors identifiziert werden, in denen Transformationen besonders häufig einen starken Effekt auf die Affinität eines Liganden ausüben. Zudem kann durch die Fragmentierung eines Eingabemoleküls mit gegebener 3D-Konformation gezielt nach bevorzugten Fragmenten gesucht werden. In **Abbildung 5** ist die Erstellung der 3D-MMP-Datenbank sowie die Erstellung eines Hybridliganden exemplarisch dargestellt.

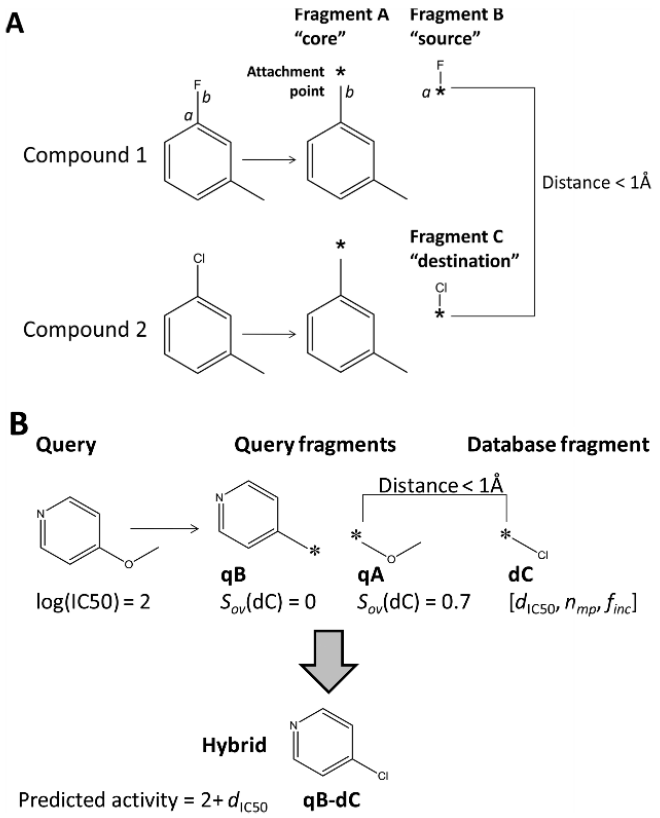


Abbildung 5. 3D-Matched Pair Workflow. (A) Erstellen der MMPs. Die Bindung zwischen Atom a und b in Molekül 1 wird entfernt um das Kernfragment A (core) und das Zielfragment B (source) zu erhalten. Die Koordinaten von Atom a und b werden als Pseudoatome (*) gekennzeichnet. Das Pseudoatom des Kernfragments dient als Verbindungspunkt (attachment point). Nach dem gleichen Prinzip wird das zweite Molekül geteilt und Fragment C erhalten. Ein 3D-MMP entsteht genau dann, wenn die Distanz der Pseudoatome kleiner als 1 Å ist. (B) Beispiel einer Datenbankanfrage. Das Eingabemolekül (query) wird fragmentiert und die Fragmente qA und qB erhalten. Fragment dC wird in der Datenbank gefunden, da die Distanz der Pseudoatome von dC und qA keiner als 1 Å ist. Das Fragment dC überlappt im Raum besser mit qA (s_{ov}=0.7) als mit qB (s_{ov}=0). Aus diesem Grund wird qA mit dC ersetzt und eine Bindung gesetzt. Somit ist ein neuer Hybrid-Ligand entstanden. Die Vorhergesagte Aktivität ist die Aktivität des Eingabemoleküls plus die Substitutions-Differenz (d_{IC50}) von Fragment dC und ergibt 2+d_{IC50} (aus einer Publikation von Posy et al. 2013[68]).

Die gezielte Vorhersage von vielversprechenden Transformationen zur Optimierung der Affinität eines Liganden zu einem spezifischen Target ist auch das Ziel der von Bradley et al. vorgestellten Software OOMMPPAA.[74] Auch hier wird die Kombination aus Aktivitätsdaten und der strukturellen Information komplexierter Liganden genutzt um Regionen für bevorzugte Interaktionen mit dem Rezeptor zu identifizieren. Dabei werden zunächst auf Basis von bereits ko-kristallisierten Liganden und Liganden mit experimentell bestimmter Bindungsaffinität MMPs gebildet. Anschließend werden für jene Liganden mit unbekannter bioaktiver Konformation unter Berücksichtigung der gemeinsamen Kernstruktur 3D-Konformationen erzeugt. Ähnlich wie bei den in Abschnitt 1.2.1 beschriebenen *Fuzzy Matched Pairs* wird die Abstraktion auf Pharmakophor-Eigenschaften der Transformationen gewählt. Aus den generierten 3D-MMPs werden nun Pharmakophor-Transformationen

extrahiert und können gemeinsam mit dem Effekt auf die Bindungsaffinität visualisiert werden.

Eine Voraussetzung für beide Methoden ist ein umfangreicher Datensatz von bereits bekannten Liganden für das untersuchte Target um eine ausreichende Anzahl MMPs für die statistische Analyse der Transformationen zu ermöglichen. Eine Weiterentwicklung dieser Methoden zu einem targetübergreifenden Modell ist deshalb das Ziel dieser Arbeit, welches im Folgenden näher erläutert wird.

1.3 Ziel der Arbeit

Alle bisher bekannten 2D- und 3D-MMP-Methoden werden targetspezifisch auf Datensätze mit einer großen Anzahl bekannter Liganden angewendet. Für Targets mit wenigen bekannten Liganden sind diese MMP-Methoden nicht anwendbar, da keine ausreichende Anzahl MMPs für eine statistische Analyse gebildet werden können. Das Ziel der im Folgenden vorgestellten Arbeit ist deshalb die Erstellung eines targetübergreifenden Modells auf Basis von MMPs im Kontext ihrer unmittelbaren Rezeptor- und Ligand-Umgebung. Dabei wird davon ausgegangen, dass eine spezifische molekulare Transformation in einer ähnlichen chemischen Umgebung einen ähnlichen Effekt auf die Bindungsaffinität eines Liganden zeigt, unabhängig davon um welchen Rezeptor oder Liganden es sich handelt.

Im ersten Schritt soll eine umfangreiche und diverse 3D-MMP-Datenbank generiert werden, welche MMPs im Kontext ihrer Targetumgebung enthält. Anschließend soll die unmittelbare Rezeptor- und Ligand-Umgebung der Transformationen bestimmt und in eine numerische Repräsentation (Deskriptor) überführt werden. Auf Basis dieses Deskriptors soll ein mathematischer Zusammenhang zwischen einem Transformationseffekt und der chemischen Umgebung einer Transformation untersucht werden. Die erstellte Datenbank von 3D-MMPs sowie das generierte Modell sollen abschließend als Webservice für die gezielte Leitstrukturoptimierung implementiert werden.

2 Material und Methoden

2.1 Analyse der ChEMBL Datenbank

In der hier vorgestellten Arbeit wurde die ChEMBL Datenbank (Abk. ChEMBLdb)[76,77] als Quelle für Moleküle und experimentell bestimmte Affinitätsdaten (K_i-, K_d-, IC_{50}-Werte) verwendet. In der genutzten Version „ChEMBL_19“ (Stand: Juli 2014) beinhaltet sie 1.638.394 chemische Verbindungen mit 12.843.338 Aktivitätswerten zu 10.579 Targets. Da es sich bei diesen Messwerten um größtenteils automatisch extrahierte Daten aus 57.156 unterschiedlichen Publikationen handelt, sollte eine Einschätzung bezüglich Vertrauenswürdigkeit und Vergleichbarkeit der unabhängig bestimmten Messwerte getroffen werden.

Jedem Messwert innerhalb der ChEMBLdb wird neben der Target-Bezeichnung auch ein Target-Typ zugewiesen. Dieser gibt zum Beispiel an, ob ein Wert in einem Assay am rekombinanten Enzym oder in einem Zellbasierten Assay bestimmt wurde. Für die Vergleichbarkeit der Messwerte aus unterschiedlichen Laboratorien ist es wichtig, dass das Target eindeutig definiert ist. Aus diesem Grund wurden Messwerte aus zellbasierten Assays oder Werte, die für homologe Proteine bestimmt wurden, nicht berücksichtigt. Auch die Relation eines Messwertes („≤“, „≥“, „~“, „%“, „=“) ist relevant, weshalb Werte mit ungenauen Relationen, wie „größer als“ oder „ungefähr“, nicht weiter berücksichtigt wurden. Bei den Messwert-Typen wurden lediglich K_i-, K_d-, und IC_{50}-Werte zugelassen, da es sich hierbei, im Gegensatz zu relativen Messwert-Typen (z. B. %-Inhibition), um größtenteils vergleichbare Größen handelt. Die Filterschritte, die für die gesamte ChEMBLdb angewandt wurden, sind in **Tabelle 1** zusammengefasst.

Tabelle 1. Filter der ChEMBL Datenbank. Aufgelistet sind die einzelnen Filterschritte und ihre Bedeutung.

ChEMBL Ausdruck	Bedeutung
target_dictionary.target_type = 'SINGLE_PROTEIN' AND target_dictionary.chembl_id != 'CHEMBL612545'	Lediglich Messwerte die eindeutig einem Target zugewiesen sind, werden akzeptiert (keine homologen Proteine, keine zellbasierten Assays). Das Target „CHEMBL612545“ ist ein „Dummy-Target“ und wird ebenfalls entfernt.
activities.standard_relation = '='	Die Relation muss einem „=“ entsprechen („<“, „>“, „≤“, „≥“, „~“ sind nicht erlaubt).
activities.standard_type = 'Kd' OR activities.standard_type = 'Ki' OR activities.standard_type = 'IC50'	Lediglich K_i-, K_d- und IC_{50}–Werte werden akzeptiert (keine %-Inhibition oder andere relative Einheiten).

Um aus allen Paaren von Messdaten eines Systems (einer Molekül-Target-Relation) jene zu filtern, die mit hoher Wahrscheinlichkeit aus verschiedenen Laboratorien stammen, entwickelten Kramer et al.[78] eine Abfolge von Filterschritten. Einige dieser Filterschritte wurden auch in dieser Arbeit verwendet, sie sind im Folgenden dargestellt (**Tabelle 2**):

Tabelle 2. Filter der ChEMBL Datenbank. Aufgelistet sind die verwendeten Filterschritte und ihre Bedeutung

Bedingungen für ein Messwertpaar	Bedeutung
Target(m1) == Target(m2) Molekül(m1) == Molekül(m2)	Die Messwerte wurden für das gleiche System (Molekül-Target-Relation) bestimmt.
Publikation(m1) != Publikation(m2)	Die Messwerte kommen aus unterschiedlichen Publikationen.
Assay(m1) != Assay(m2)	Die Messwerte sollten möglichst aus unterschiedlichen Assays stammen.
Autoren(m1) ∩ Autoren(m2) == 0	Keine Überschneidungen der Autorennamen.

m1 != m2	Die Messwerte sind nicht identisch
ABS(m1-m2) > 0.02 log Units	Die Messwerte dürfen nicht durch Auf- oder Abrundung ineinander überführbar sein.
ABS(m1-m2) != 3 log Units AND ABS(m1-m2) != 6 log Units	Messwerte dürfen nicht genau 3 oder genau 6 logarithmische Einheiten auseinander liegen, weil sie sonst als Übertragungsfehler der Einheit vermutet werden.

Alle folgenden Berechnungen wurden abhängig vom Messtypen (K_i, K_d, IC_{50}) durchgeführt und wiederum jeweils in drei Gütekriterien unterteilt. Die Gütekriterien sind als Kennzeichnung jeweils für das Target bzw. die Messdaten in der ChEMBLdb hinterlegt. Ist ein Assay-Target System mit der Bezeichnung „CONFIDENCE_SCORE == 9“ versehen, so bedeutet dies, dass der Assay eindeutig einem Protein zugewiesen wurde und nicht etwa z. B. einem homologen Protein oder Protein-Komplex. Die Bezeichnung „CURATED_BY==‘Expert‘“ bedeutet, dass die zugewiesenen Messdaten von einem Experten in der Primärreferenz überprüft wurden. Um zu analysieren ob die Gütekriterien einen Einfluss auf die Korrelation der Messwerte aus verschiedenen Laboratorien haben, wurde die statistische Auswertung separat für alle Paare, für jene Paare mit „CONFIDENCE_SCORE==9“ und für jene Paare mit „CURATED_BY==‘Expert‘“ durchgeführt. Als Maß für die Korrelation wurde der Pearson-Korrelationskoeffizient für die jeweiligen Messwertpaare berechnet. Für die visuelle Darstellung der Korrelation wurde die Software R (http://www.r-project.org/) verwendet.

2.2 VAMMPIRE Datenbank

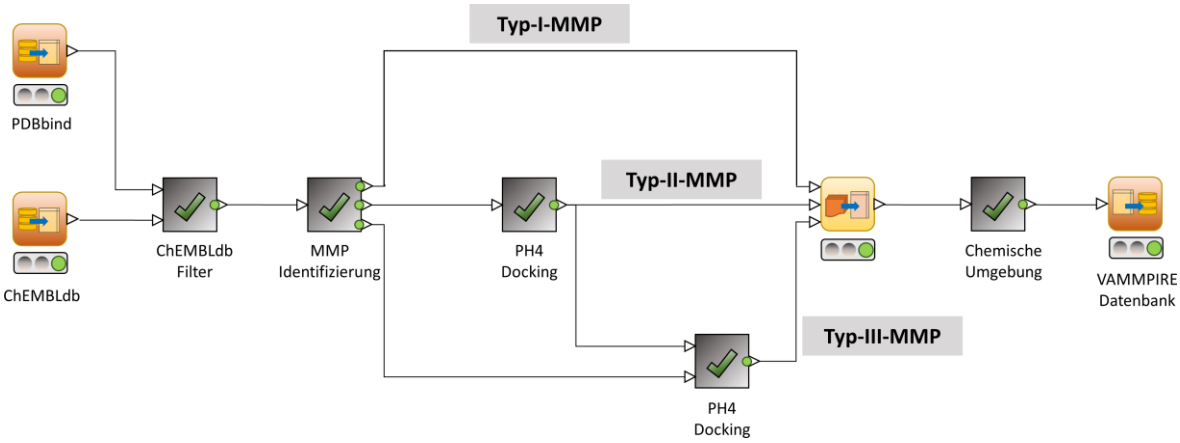


Abbildung 6. Workflow zur Erstellung der VAMMPIRE Datenbank. Die Basis bilden die Datenbanken PDBbind und ChEMBLdb. Es folgt die Identifizierung der potentiellen MMPs (pMMPs) durch Filterung der ChEMBLdb und schließlich die Identifizierung der MMPs aus den pMMPs. Im Anschluss werden die 3D-Koordinaten der Typ-II- und Typ-III-MMPs erzeugt und die chemische Umgebung der Substituenten bestimmt.

Für die Erstellung der 3D-MMP-Datenbank „VAMMPIRE“ (engl. **V**irtually **A**ligned **M**atched **M**olecular **P**airs **I**ncluding **R**eceptor **E**nvironment) wurde ein automatisierter Arbeitsablauf (Workflow) implementiert. Da die Basis der VAMMPIRE Datenbank wiederum durch zwei regelmäßig aktualisierte Datenbanken gebildet wird, ist ein solcher automatischer Ablauf von großem Vorteil. Die Implementierung des Workflows erfolgte mit Hilfe des Workflow Management Tools KNIME (Konstanz Information Miner).[79] KNIME bietet zum einen die Möglichkeit eine Reihe von vorimplementierten Funktionen (im Folgenden als Knoten bezeichnet) in beliebiger Reihenfolge (Verbindung der Knoten) auf einen Datensatz anzuwenden und zum anderen ermöglicht KNIME eigene Funktionen zu implementieren. Da die Nutzung gängiger Chemieinformatik-Software in KNIME ebenfalls möglich ist, wurde ein Großteil dieser Arbeit in Form eines Workflows Implementiert. Der Workflow, der zum Aufbau der VAMMPIRE Datenbank führt, ist in **Abbildung 6** schematisch dargestellt. Die grundlegenden Techniken zur Erstellung der VAMMPIRE Datenbank wurden bereits in der vorangegangenen Diplomarbeit erarbeitet.[80] Da die Berechnung der MMPs, sowie die Modellierung der 3D-MMPs in der hier vorgestellten Arbeit optimiert wurden, sind sie dennoch im Folgenden dargestellt.

2.2.1 Datenbankvorbereitung

Die Grundlage für die Erstellung der VAMMPIRE Datenbank bilden die beiden frei verfügbaren Datenbanken PDBbind[81,82] (Version 2014) und ChEMBLdb (Version 19). Die PDBbind beinhaltet eine Sammlung von Protein-Ligand-Komplexen, extrahiert aus der Protein Data Bank[9] (PDB) und ist wiederum unterteilt in drei Teildatenbanken: Das sogenannte *„general set“* besteht aus 10.776 Protein-Ligand-Komplexen, deren primäre Referenz manuell überprüft wurde und für die experimentell bestimmte Affinitätsdaten (K_i-, K_d-, IC_{50}-Werte) hinterlegt sind. Das *„refined set“* beinhaltet 2.959 Einträge und ist zusätzlich durch Gütekriterien, wie die Auflösung der Kristallstrukturen (mindestens 2.5 Å) und die Verlässlichkeit der Affinitätsdaten (die Werte müssen mindestens dreimal unabhängig voneinander bestimmt worden sein), gefiltert. Das *„core set“* beinhaltet 195 Einträge und dient als qualitativ hochwertiger Validierungsdatensatz mit dem Fokus auf Diversität von Proteinstrukturen und Affinitätsdaten.

Zur Vorbereitung der Kristallstrukturen wurden diese zunächst mit Hilfe der *Protonate3D* Routine protoniert. Diese Funktion wird vom Software-Paket „MOE“ (Molecular Operating Environment, Version 2014.09)[83] als KNIME-Knoten zur Verfügung gestellt. Die ChEMBLdb wurde für die Erstellung der VAMMPIRE Datenbank auf jene Targets reduziert die im PDBbind *„general set“* hinterlegt sind und beinhaltet 277.964 (teilweise redundante) Messwerte. Um eine Verbindung zwischen den Targets der PDBbind und den Targets der ChEMBLdb herzustellen musste zunächst eine gemeinsame Target-Identifikation (Target-ID) gefunden werden. Aus diesem Grund wurde im ersten Schritt eine Zuordnung von PDBcode (Target-ID der PDBbind) zu ChEMBL_ID (Target-ID der ChEMBLdb) durchgeführt. Da dies auf direktem Wege nicht möglich war, wurde der Umweg über die UniProt_ID gewählt (PDBcode → UniProt_ID → ChEMBL_ID), welche die Target-ID der Universal Protein Resource Datenbank (UniProt)[84,85] darstellt und in der ChEMBLdb hinterlegt ist. Die Zuordnung der UniProt_ID zum entsprechenden PDBcode wurde anschließend über den „ID-mapping“ Service der UniProt Webseite (http://www.uniprot.org/) bewerkstelligt.

Im Falle, dass mehr als ein Affinitätswert für ein Molekül zu einem spezifischen Target in der ChEMBLdb hinterlegt war, diente der Median im weiteren Verlauf als angenommener Messwert. Waren unterschiedliche Messtypen für eine Molekül-Target-Kombination

vorhanden, wurden bevorzugt K_i- und K_d-Werte gewählt. Für Werte, die mit mehr als einer Publikation hinterlegt wurden, wurde die früheste Quelle als Primärquelle angenommen.

Moleküle werden in der ChEMBLdb im „Structure Data Format“ (Abk. SDF) als 2D-Repräsentation hinterlegt. Mit Hilfe der w*ash* Funktion, die von MOE als KNIME-Knoten zur Verfügung gestellt wird, wurde eine einheitliche Protonierung der Moleküle bewerkstelligt und Salze oder andere Komponenten, die nicht zum Liganden gehören, entfernt. Mit Hilfe des KNIME-Knotens *Generate Coords* von *RDKit*[86] wurden initiale 3D-Koordinaten für die Moleküle erzeugt.

2.2.2 MMP-Identifizierung

Ein potentielles MMP (pMMP) wurde definiert als ein Paar von Molekülen, deren Aktivität für das gleiche Target bestimmt wurde und deren experimentell bestimmte Werte dem gleichen Messtypen (K_i, K_d oder IC_{50}) entsprechen. Für IC_{50}-Werte im Speziellen gilt, dass sie aus der gleichen Publikation stammen müssen.

Für die Identifizierung der MMPs wurde der KNIME-Knoten *Matched Pairs Detector*[56,67,87], bereitgestellt von *Erl Wood Cheminformatics* (Research IT and Computational Drug Discovery group at Erl Wood, United Kingdom), verwendet. Hier wird im ersten Schritt jedes Eingabemolekül an sämtlichen azyklischen Bindungen in jeweils zwei Teile getrennt. Dabei entsteht eine Schlüssel-Wert-Beziehung, wobei der Schlüssel im späteren Verlauf den gemeinsamen Kontext der Moleküle darstellt, während die Werte eines Schlüssels den molekularen Transformationen am Kontext entsprechen. Die Moleküle, die sich aus einem gemeinsamen Schlüssel (Kontext) und den zugehörigen Werten (Substituenten) ergeben, bilden die MMPs. Die Vorgehensweise des Algorithmus ist in **Abbildung 7** skizziert.

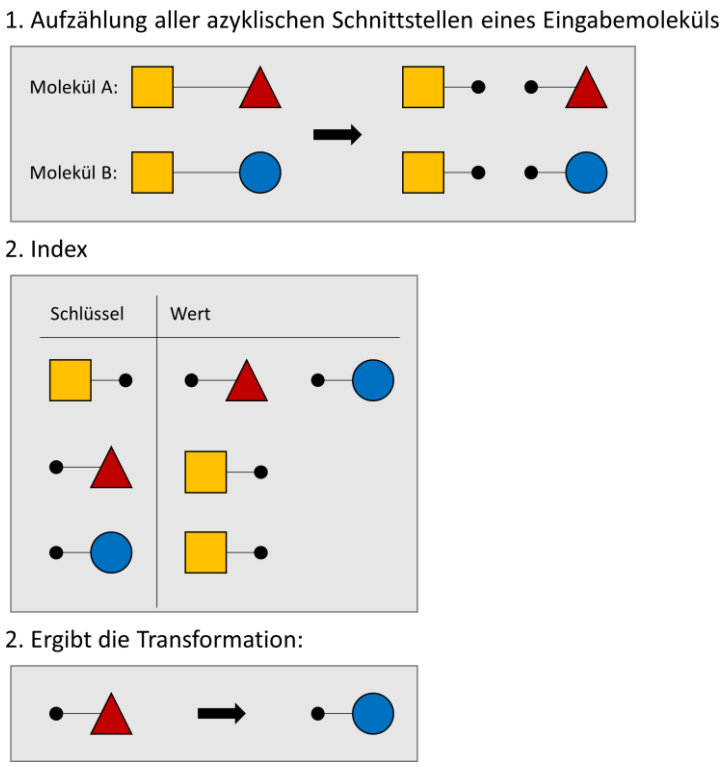


Abbildung 7. Algorithmus für die Identifizierung der Matched Molecular Pairs nach Hussain und Rea 2010.[87] 1) Die Eingabemoleküle werden an allen azyklischen Bindungen in jeweils zwei Teile getrennt. 2) Erstellung einer Schlüssel-Wert-Zuweisung, wobei der Schlüssel dem gemeinsamen Kontext der Moleküle entspricht und die Werte die Substituenten der Moleküle darstellen. 3) Alle Werte eines Schlüssels entsprechen somit den Transformationen und die entsprechenden Moleküle bilden die MMPs.

Die identifizierten MMPs wurden anschließend einem Filterschritt unterzogen. Dabei wurde festgelegt, dass der Kontext innerhalb eines MMPs mindestens doppelt so viele Nicht-Wasserstoffatome besitzen muss wie die beiden Substituenten. Dadurch wird sichergestellt, dass es sich relativ zur Größe des gesamten Moleküls um eine kleine Transformation handelt. Zusätzlich wurde die maximale Größe eines Substituenten eingeschränkt. Die Maximalgröße für Ring-Substituenten wurde auf 9 Nicht-Wasserstoffatome (z. B. 3-fach-substituiertes Phenyl) und die Maximalgröße für azyklische Substituenten auf 5 Nicht-Wasserstoffatome (z. B. Trifluormethoxy) festgelegt.

2.2.3 Transformationseffekt

Jedem identifizierten MMP wurde ein Transformationseffekt zugewiesen. Dazu wurde zunächst der negative dekadische Logarithmus der einzelnen Aktivitätswerte berechnet, welcher im Folgenden als p-Wert (pK_i, pK_d, pIC_{50}) bezeichnet wird. Der Transformationseffekt (*E*) ergibt sich aus der Differenz der p-Werte von Molekül *A* und Molekül *B* innerhalb eines MMPs und ist negativ, wenn eine Transformation (Substituent(*A*)

→ Substituent(*B*)) zu einer Verschlechterung der Affinität führt und positiv, wenn sie zu einer Verbesserung führt (Gleichung 1).

$$E = \log_{10} Aktivität(A) - \log_{10} Aktivität(B) \qquad \text{Gleichung 1}$$

Ein Transformationseffekt von 1,0 bedeutet eine Verbesserung der Affinität um eine Zehnerpotenz, während ein Transformationseffekt von -1,0 einer Verschlechterung der Affinität um eine Zehnerpotenz entspricht.

2.2.4 Modellierung der 3D-MMPs

Im Verlauf der Arbeit wurden drei verschiedene Typen von MMPs definiert. Bei MMPs vom Typ-I handelt es sich um Molekülpaare mit bekannter bioaktiver Konformation, da beide Strukturen im PDBbind *„general set"* enthalten sind. Bei MMPs vom Typ-II ist die bioaktive Konformation lediglich für eines der beiden Moleküle bekannt, welche als Basis für die Konformationsvorhersage des zweiten Moleküls dient. Diese vorhergesagte Konformation kann anschließend wieder als Basis dienen um MMPs vom Typ-III zu bilden. Eine Veranschaulichung der verschiedenen MMP-Typen ist **Abbildung 8** gegeben.

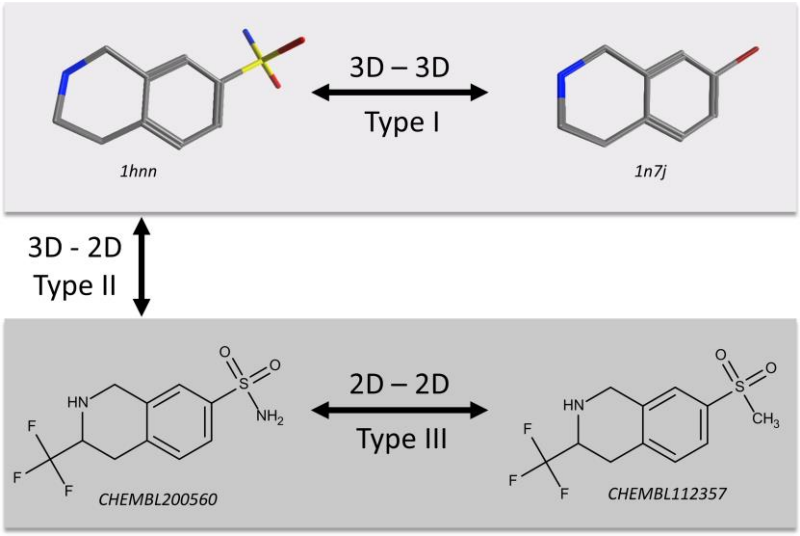


Abbildung 8. MMP-Typen. Typ-I-MMP: Beide Moleküle und deren bioaktive Konformation liegen in der PDBbind vor. Typ-II-MMP: Die bioaktive Konformation ist für eines der Moleküle bekannt und dient als Basis für die Vorhersage der Konformation des zweiten Moleküls. Typ-III-MMP: Auf Basis der vorhergesagten Konformation des Typ-II-MMPs kann auch die Konformation eines weiteren Moleküls vorhergesagt werden, obwohl es selbst kein MMP mit einem ko-kristallisierten Liganden bildet.

Aufgrund der hohen Ähnlichkeit der Moleküle innerhalb eines MMPs wurde auf einen ähnlichen Bindemodus geschlossen, der mit Hilfe von „Docking mit Pharmakophor-

Platzierung“ vorhergesagt werden sollte. Das Docking wurde mit der Software *MOE* durchgeführt, da hier die Möglichkeit besteht sogenannte „Pharmakophormodelle“ als Vorgabe für den Platzierungsalgorithmus zu verwenden. Ein Pharmakophormodell stellt eine Abstraktion potentieller Interaktionen des Liganden mit dem Rezeptor dar. Durch die Definition von Sphären um ausgewählte Atome, können Regionen definiert werden die Interaktionen repräsentieren, welche als essentiell oder bevorzugt für die Bindung des Liganden angenommen werden. Zur Modellierung der Typ-II-MMPs wurde im ersten Schritt der gemeinsame Kontext des 2D- und des 3D-Moleküls identifiziert. Der 2D-Kontext stellt den „Schlüssel“ bei der MMP-Identifizierung dar (siehe Abschnitt 2.2.2) und ist bereits bekannt. Für ein geeignetes Pharmakophormodell am gemeinsamen Kontext sind allerdings die 3D-Koordinaten essentiell, weshalb eine Zuordnung der 2D-Atomkoordinaten zu den korrespondierenden 3D-Atomkoordinaten nötig war. Diese Zuordnung wurde mit Hilfe der Funktionen des *Small Molecule Subgraph Detector (SMSD)*[88] realisiert, welche als Modul in der Java-Bibliothek *CDK* (Chemistry Development Kit) integriert sind.[89] Hier erfolgt eine eindeutige Zuordnung der Atome (engl. atom mapping) der maximalen gemeinsamen Substruktur (Abk. MCS, engl. maximum common substructure) beider Moleküle. Auf Basis der 3D-Koordinaten des gemeinsamen Kontextes wurden anschließend sogenannte „Pharmakophor-Annotationspunkte“ extrahiert. Die Zuweisung erfolgte mit Hilfe eines SVL (engl. Scientific Vector Language) Skripts, welches von *MOE* zur Verfügung gestellt wird. Diese Annotationspunkte entsprechen einer Auswahl von Pharmakophor-Eigenschaften aus dem „*Unified Scheme*“ von *MOE* welche wie folgt definiert sind (Tabelle 3):

Tabelle 3. **Verwendete Pharmakophor-Annotationen.** Dargestellt sind die verwendeten Annotationen aus dem „*Unified Scheme*“ von MOE und ihre Bedeutung.

Annotation	Bedeutung
Acc	Wasserstoffbrücken-Akzeptor
ANI	Anionisches Atom
ARO	Aromatisches Zentrum
CAT	Kationisches Atom
DON	Wasserstoffbrücken-Donor
HYD	Hydrophobes Atom
ML	Metall-Ligator

O2	Projektion eines potentiellen ligierten Metalls
PIN	Projektion entlang der π-System Ebene
PIR	Andere π-Systeme

Im nächsten Schritt wurden maximal 1000 Konformationen des zu platzierenden Liganden, mit Hilfe der *Conformation Import* Funktion des MOE-Knotens *Conformations* generiert. Mit der Methode *Pharmacophore* des MOE-Knotens *Docking Placement* wurden anschließend die Liganden auf Basis des erstellten Pharmakophormodells platziert. Die generierten Konformationen mussten nach der Platzierung das Pharmakophormodell erfüllen und durften keine Kollision mit dem Protein aufweisen. Anschließend wurde eine Energieminimierung der Konformationen in ihrer Proteinumgebung durchgeführt. Dazu wurde das *Amber12:EHT* Kraftfeld, welches innerhalb des MOE-Knotens *Pose Refinement* implementiert ist, verwendet. Das Kraftfeld *Amber12:EHT* ist eine Kombination von *Amber12*[90] und *EHT*[91], wobei die Parametrisierung des Proteins mit *Amber12* realisiert wird, während die Parametrisierung des Liganden mit der *2D erweiterten Hueckel Theorie (EHT)* erfolgt. Nach der Energieminimierung der Konformationen wurde die Wurzel des mittleren quadratischen Abstands (Abk. RMSD, engl. root mean square deviation) der MCS des generierten 3D-MMPs berechnet. Zur Berechnung der MCS wurde, wie bei der Identifizierung des 3D-Kontextes (Abschnitt 2.2.2), die *SMSD* Funktion verwendet. Um mit hoher Wahrscheinlichkeit von einem ähnlichen Bindemodus der beiden Moleküle eines MMPs ausgehen zu können, wurde ein maximaler RMSD der MCS von 1 Å festgelegt. Die Konformation mit dem kleinsten RMSD wurde im weiteren Verlauf als bioaktive Konformation angenommen. Die Modellierung der Typ-II-MMPs ist in **Abbildung 9** schematisch dargestellt.

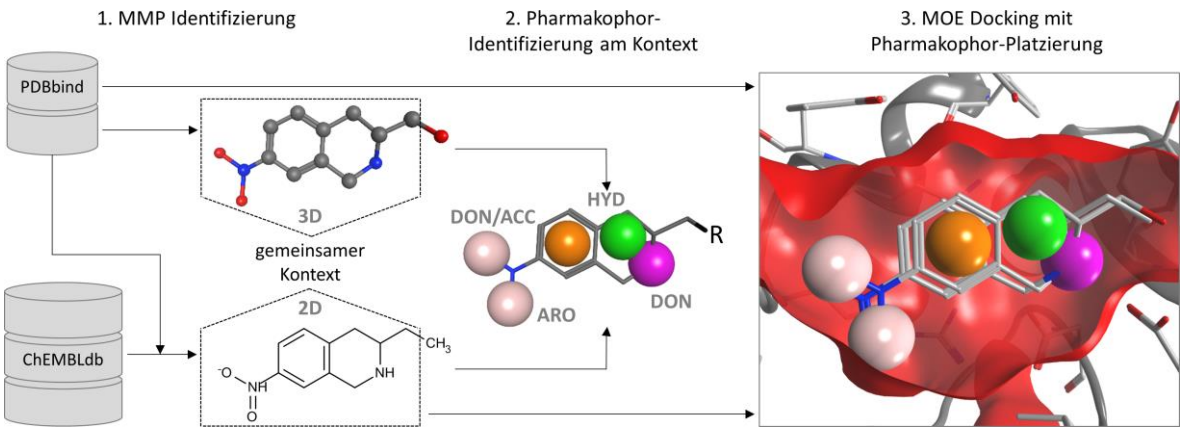


Abbildung 9. Modellierung der Typ-II-MMPs. Ein ko-kristallisierter Ligand aus der PDBbind (mit 3D-Koordinaten) bildet ein MMP mit einem Molekül aus der ChEMBLdb (mit 2D-Koordinaten). Auf Basis des gemeinsamen Kontextes beider Moleküle wird ein Pharmakophormodell erstellt, welches als Platzierungshilfe für das anschließende molekulare Docking dient.

Typ-III-MMPs konnten nun auf Basis der vorhergesagten Konformation innerhalb eines Typ-II-MMPs generiert werden, obwohl sie selbst kein MMP mit einem ko-kristallisierten Liganden bilden. Die Vorgehensweise ist analog zur Erstellung der Typ-II-MMPs.

2.2.5 Berechnung der chemischen Umgebung eines Substituenten

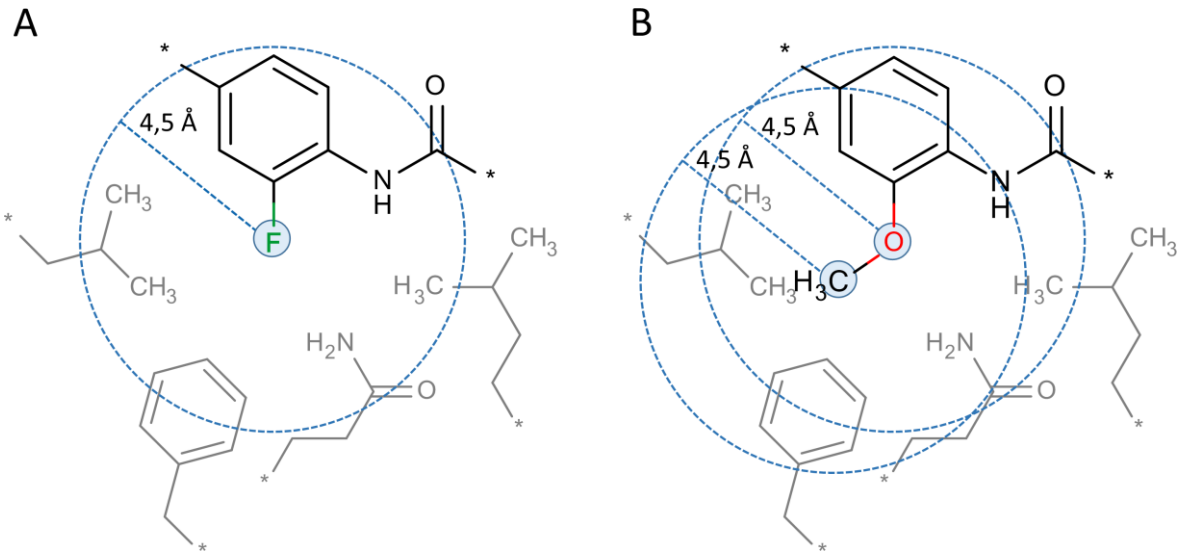


Abbildung 10. Umfang der chemischen Umgebung eines Substituenten. Zur chemischen Umgebung eines Substituenten gehören jene Rezeptoratome, die innerhalb eines Radius von 4,5 Å um ein Nicht-Wasserstoffatom des Substituenten liegen. A) Die chemische Umgebung für einen Substituenten der Größe 1 (im Beispiel das Fluoratom) ist durch einen Kreis gekennzeichnet (blau gestrichelt). B) Die eingeschlossene Umgebung für einen Substituenten der Größe 2 (im Beispiel die Methoxy-Gruppe) ist durch die Vereinigung der Umgebungen beider Atome definiert.

Die chemische Umgebung eines Substituenten wurde definiert als jene Rezeptoratome, die innerhalb eines Radius von 4,5 Å um eines der Nicht-Wasserstoffatome des Substituenten

liegen (**Abbildung 10**). Um die Koordinaten der Substituent-Atome zu bestimmen, wurde zunächst eine Substruktursuche des Kontextes im betrachteten Molekül durchgeführt. Jene Atome eines Moleküls, die nicht zur errechneten Substruktur gehörten stellten den Substituenten dar. Ein Sonderfall wurde definiert für den Fall das der Substituent selbst ein Wasserstoffatom darstellt, da alle bisherigen Berechnungen ohne explizite Wasserstoffatome durchgeführt wurden. In diesem Fall wurden zunächst Wasserstoffatome mit 3D-Koordinaten an das Molekül addiert und die Umgebung des Wasserstoffatoms bestimmt. Die Substruktursuche, die Addition der Wasserstoffatome, sowie alle weiteren Berechnungen auf den Liganden wurden mit Hilfe der Python Bibliotheken des *RDKit* (Version 2014.03) realisiert. Proteine hingegen wurden mit Hilfe der *Biopython* 1.62 Tools (http://www.biopython.org) prozessiert. Die chemische Umgebung wurde jeweils für beide Substituenten eines MMPs berechnet, wobei die vorab manuell definierten Rezeptor-Atomtypen (siehe Definition in **Tabelle A 1** des Anhangs) sowie die 3-Buchstaben-Identifizierung der Aminosäuren gespeichert wurden. Zusätzlich wurde gespeichert, ob sich das betrachtete Atom in der Seitenkette oder dem Rückgrat der Aminosäure befindet. Es wurden sechs unterschiedliche Atomtypen definiert, die sich aus dem PDB-Atomtypen und der Aminosäure, in der sich das Atom befindet, zusammensetzen (Hydrophob: HYD, Aromatisch: ARO, Polar: POL, H-Brücken-Donor: DON, H-Brücken-Akzeptor: ACC, H-Brücken-Donor oder Akzeptor: DON/ACC). Die PDB-Atomtypen, die 3-Buchstaben-Identifizierung der Aminosäuren sowie die Information ob sich das Atom in der Seitenkette oder dem Rückgrat der Aminosäure befindet, sind im *Protein Data Bank Format* (*.pdb) eines Proteins hinterlegt und können durch entsprechende *Biopython*-Funktionen abgefragt werden.

2.3 VAMMPIRE-LORD

2.3.1 Atompaar-Deskriptor LORD_FP

Basierend auf der 3D-MMP-Datenbank sollte ein Vorhersagemodell zur gezielten Leitstrukturoptimierung erstellt werden. Ein Atompaar Deskriptor (LORD_FP) wurde entwickelt um die Interaktionen zwischen Protein und Ligand, aber auch die direkte Umgebung des Substituenten innerhalb des Liganden, zu beschreiben. Aus diesem Grund wurde die chemische Umgebung, wie in Abschnitt 2.2.5 beschrieben, durch die

Einbeziehung der Ligand-Atome erweitert. Dazu wurde analog zur Rezeptorumgebung ein Radius von 4,5 Å um die Substituent-Atome gewählt, nur dass in diesem Fall lediglich die Ligand-Atome bzw. deren EState-Atomtypen[92] berücksichtigt wurden. EState-Atomtypen beschreiben das chemische Element des Atoms sowie die Anzahl der Bindungen zu den benachbarten Atomen. Der Atomtyp „aasC“ beispielsweise steht für ein Kohlenstoffatom mit zwei aromatischen Bindungen sowie einer Einfachbindung (engl. a: aromatic; s: single; C: carbon element).

Der LORD_FP besteht aus drei Teildeskriptoren, welche Ligand-Ligand-Interaktionen (LLIs), Protein-Protein-Interaktionen (PPIs) und Protein-Ligand-Interaktionen (PLIs) beschreiben. Für jeden Teildeskriptor wird ein Vektor mit allen möglichen Atomtyppaaren erstellt und mit 0 initialisiert. Jeder Teildeskriptor ist wiederum in drei Distanzklassen (engl. bins) unterteilt und beschreibt Interaktionen über kurze Distanz (bin1: 0-2 Å), mittlere Distanz (bin2: 2-3 Å) und hohe Distanz (bin3: 3-4,5 Å). Sobald ein Atompaar in entsprechender Distanz identifiziert wird, erhöht sich der Wert an der jeweiligen Stelle des Deskriptors.

Der LLI-Teildeskriptor besteht aus allen möglichen Paaren von EState-Atomtypen, welche in den Molekülen der VAMMPIRE Datenbank vorkommen. Es wurden 44 unterschiedliche Atomtypen gefunden, woraus sich 990 mögliche Paare ergeben. In **Abbildung 11** ist der Aufbau des LLI-Teildeskriptors exemplarisch dargestellt.

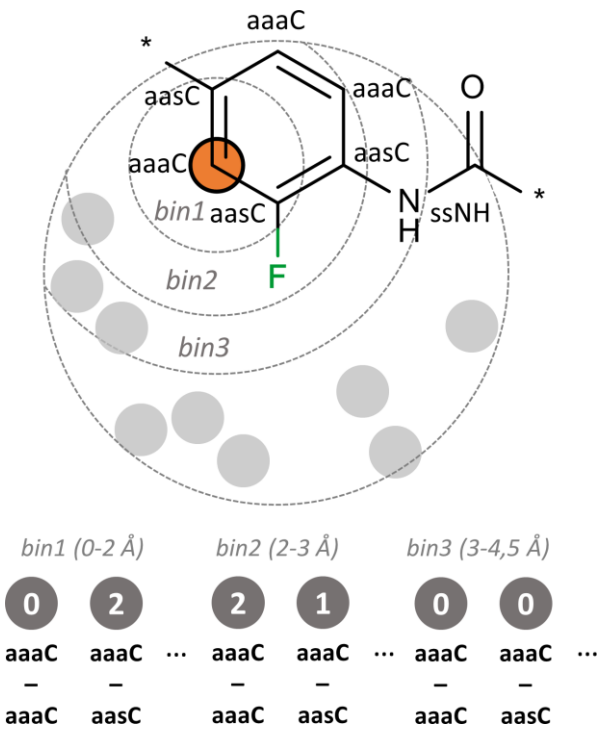


Abbildung 11. Aufbau des Teildeskriptors zur Darstellung der Ligand-Ligand-Interaktionen. Gezeigt ist die chemische Umgebung des Fluoratoms (äußerer Kreis). Für jedes Ligand-Atom der chemischen Umgebung werden alle möglichen Atompaare gebildet, je nachdem in welchem Abstand die Paare gefunden werden. Die Anzahl der Paare aaaC-aaaC und aaaC-aasC, ausgehend vom markierten Atom (orange), sind für die drei Distanzklassen (bin1, bin2, bin3) exemplarisch dargestellt.

Eine ähnliche Vorgehensweise wurde zur Darstellung der PPIs verwendet (Abbildung 12). Hier werden die Protein-Atome allerdings lediglich durch sechs mögliche Atomtypen beschrieben (Hydrophob: HYD, Aromatisch: ARO, Polar: POL, H-Brücken-Donor: DON, H-Brücken-Akzeptor: ACC, H-Brücken-Donor oder Akzeptor: DON/ACC). Die Zuweisung der Atomtypen zu den einzelnen Atomen der unterschiedlichen Aminosäuren ist in Tabelle A 1 im Anhang dargestellt. Der PPI-Teildeskriptor wird durch 21 mögliche Atompaare dargestellt.

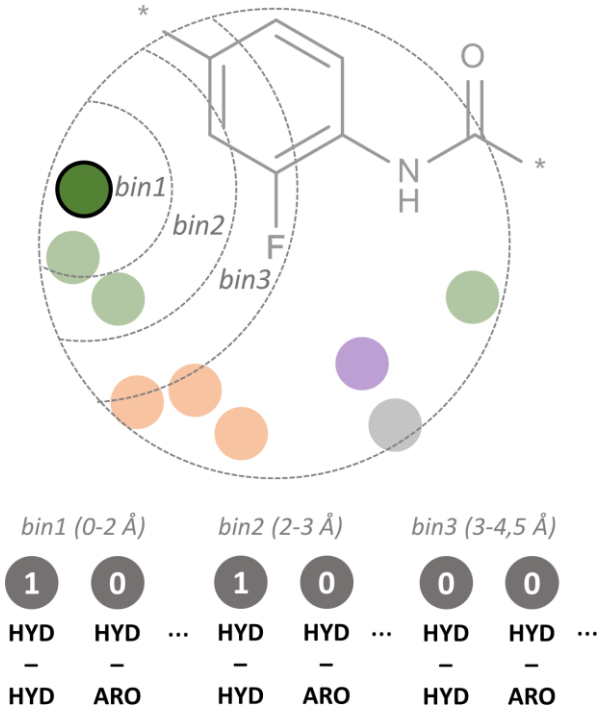


Abbildung 12. Aufbau des Teildeskriptors zur Darstellung der Protein-Protein-Interaktionen. Gezeigt ist die chemische Umgebung des Fluor Atoms (äußerer Kreis). Für jedes Protein-Atom der chemischen Umgebung werden alle möglichen Atompaare gebildet, je nachdem in welchem Abstand die Paare gefunden werden. Die Anzahl der Paare HYD-HYD und HYD-ARO, ausgehend vom markierten Atom (grün, schwarze Umrandung), sind für die drei Distanzklassen (bin1, bin2, bin3) exemplarisch dargestellt.

Für die Darstellung der PLIs wurde ein dritter Teildeskriptor entwickelt. Dieser beschreibt alle möglichen Paare zwischen Protein und Substituent und die zugehörige Distanz. Daraus ergeben sich 264 mögliche Paare (44 EState-Atomtypen * 6 Protein-Atomtypen). Die Erstellung des Teildeskriptors ist in **Abbildung 13** exemplarisch dargestellt.

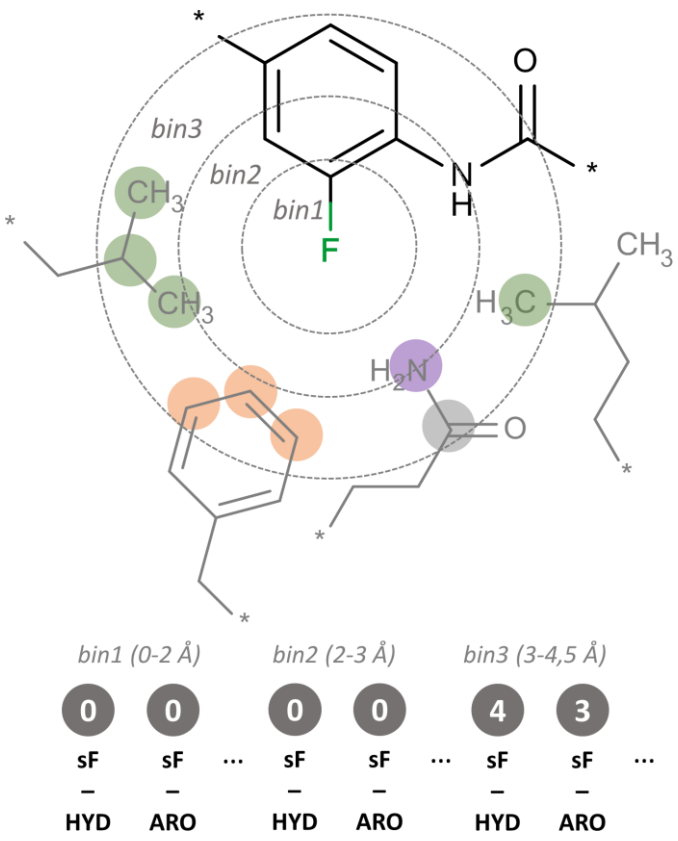


Abbildung 13. Aufbau des Teildeskriptors zur Darstellung der Protein-Protein-Interaktionen. Gezeigt ist die chemische Umgebung des Fluor Atoms (äußerer Kreis). Für jedes Substituent-Atom werden alle möglichen Atompaare mit Rezeptor-Atomen gebildet, je nachdem in welchem Abstand die Paare gefunden werden. Die Anzahl der Paare sF-HYD und sF-ARO, ausgehend vom Fluor Atom (sF), sind für die drei Distanzklassen (bin1, bin2, bin3) exemplarisch dargestellt.

Es resultiert ein umfangreicher Deskriptor (LORD_FP) der Länge 3.825, mit dem eine Transformation in einer spezifischen chemischen Umgebung beschrieben werden kann. Die Ähnlichkeit der chemischen Umgebung zweier Transformationen in unterschiedlichen Targets kann durch die Distanz der entsprechenden Deskriptoren ausgedrückt werden. Eine numerische Version des Tanimoto-Koeffizienten wurde zur mathematischen Darstellung der Ähnlichkeit gewählt. Er wird einzeln für die drei Teildeskriptoren berechnet. (Gleichung 2)

$$T(A,B) = \frac{\sum_{i=1}^{n} common(A_i, B_i)}{\sum_{i=1}^{n} A_i + \sum_{i=1}^{n} B_i - \sum_{i=1}^{n} common(A_i, B_i)} \qquad \text{Gleichung 2}$$

wobei *T(A,B)* den Tanimoto-Koeffizienten der Teildeskriptoren *A* und *B* darstellt. A_i und B_i stehen für die Häufigkeit mit der ein Atomtyp an Position *i* vorkommt und $Common(A_i,B_i)$ stellt die Anzahl der gemeinsamen Einträge der Teildeskriptoren *A* und *B* an Position *i* dar. Die Ähnlichkeit der Teildeskriptoren wird durch einen Wert zwischen null und eins ausgedrückt, wobei der Wert eins für die höchstmögliche Ähnlichkeit steht. Die Ähnlichkeit

der gesamten chemischen Umgebung wird durch den Mittelwert der Ähnlichkeiten der einzelnen Teildeskriptoren ausgedrückt

2.3.2 Validierung

2.3.2.1 Intrinsische Validierung

Um die target-übergreifende Vorhersagekraft des LORD_FP zu überprüfen, sollte eine Kreuzvalidierung anhand der VAMMPIRE Datenbank durchgeführt werden. Durch den Vergleich der chemischen Umgebung unterschiedlicher Targets sollte gezeigt werden, dass Transformationen genau dann einen ähnlichen Effekt auf die Bindungsaffinität eines Liganden haben, wenn die chemische Umgebung (der LORD_FP Deskriptor) ebenfalls ähnlich ist, auch wenn es sich um Targets mit unterschiedlichen Funktionen handelt.

Der Validierungsdatensatz unterscheidet sich vom VAMMPIRE Datensatz darin, dass die Target Klassifizierung nicht durch die CHEMBL_ID, sondern wenn vorhanden, durch die EC-Nummer (engl. Enzyme Commission number) oder durch einen manuell hinzugefügten Target Namen (extrahiert aus dem jeweiligen PDB Eintrag) dargestellt wird. Damit sollte sichergestellt werden, dass es sich bei der Bildung von Targetpaaren um unterschiedliche Targets handelt. Außerdem wurden pro Transformation und Target 20 zufällige Einträge gewählt, um zu verhindern dass überrepräsentierte Transformationen und Targets die Statistik verfälschen. Der Validierungsdatensatz wurde mithilfe der folgenden Prozedur erstellt:

```
Wiederhole 20x
      Für jede Transformation der Datenbank
            Wähle einen zufälligen Eintrag pro Target
            Berechne LORD_FP für alle Einträge
            Für jeden Eintrag
                  Berechne Ähnlichkeit zu jedem anderen Eintrag
```

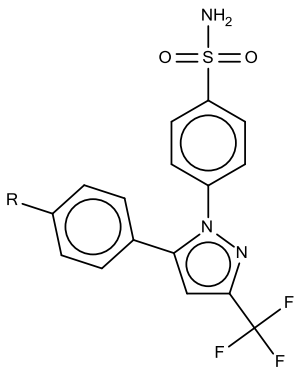
Der entstandene Validierungsdatensatz wurde anschließend in drei Teildatensätzen bezüglich der Vorhersagekraft des LORD_FP untersucht. Der erste Teildatensatz umfasst alle Paare, der zweite Teildatensatz umfasst alle Paare deren Transformationseffekt mindestens 0,5 Log-Einheiten beträgt und der dritte Datensatz umfasst jene Paare deren Transformationseffekt mindestens eine Log-Einheit beträgt. Die allgemeine Verteilung der

Ähnlichkeitswerte wurde in einem Balkendiagramm visualisiert und anhand der Ergebnisse verschiedene Ähnlichkeits-Schwellenwerte θ (θ=0,0, θ=0,4, θ=0,5, θ=0,6) definiert. Anschließend wurde für alle Teildatensätze und Schwellenwerte die Anzahl der Targetpaare bestimmt, bei denen die Tendenz des Transformationseffekts übereinstimmt (beide positiv oder beide negativ). Die Anzahl der Übereinstimmungen wurde in einem Balkendiagramm dargestellt und ausgewertet. Für alle Visualisierungen wurde die Software R (http://www.r-project.org/) verwendet.

2.3.2.2 Retrospektive Validierung

Anhand eines Targets, welches sich nicht in der VAMMPIRE Datenbank befindet, sollte die Vorhersagekraft des LORD_FP ebenfalls überprüft werden. Dazu wurde eine von Penning et al.[93] veröffentlichte SAR von Derivaten des Cyclooxygenase-2 (COX-2)-Hemmers Celecoxib gewählt. Diese SAR beinhaltet 16 Derivate mit IC_{50} Werten zwischen 5 nM und >100 µM (**Tabelle 4**) woraus sich 240 potentielle gerichtete MMPs ergeben.

Tabelle 4. Celecoxib Derivate. [a] Substituenten an der R-Gruppe, [b] IC_{50}-Werte aus einem rekombinanten humanen COX-2 Assay.[93]

	R-Gruppe [a]	IC_{50} (µM) [b]	R-Gruppe [a]	IC_{50} (µM) [b]
	-NMe_2	0.005	-NH2	0.34
	-OMe	0.008	-OEt	0.64
	-SMe	0.009	-Et	0.86
	-Cl	0.010	-NO_2	2.63
	-NHMe	0.016	-CF_3	8.23
	-H	0.032	-CO_2H	11.2
	-Me	0.040	-CH_2OH	93.3
	-F	0.041	-OH	>100

Im ersten Schritt wurden alle Derivate in die Bindetasche der COX-2 gedockt. Dazu wurde die Kristallstruktur der COX-2 mit dem Celecoxib Derivat 1-Phenylsulfonamid-3-Trifluormethyl-5-Parabromphenylpyrazol (PDBcode: 6COX) zunächst mit Hilfe der *Protonate3D* Routine in *MOE* protoniert. Die Liganden wurden mit Hilfe der *minimize* Funktion energieminimiert. Alle Moleküle wurden mit der Platzierungsmethode *Triangle Matcher* in die Bindetasche gedockt, mit der Bewertungsfunktion *London dG* bewertet und anschließend mittels *AMBER12:EHT* Kraftfeld energieminimiert. Zum Abschluss folgte eine weitere Bewertung der maximal 10 generierten Konformationen mit der Funktion

GBVI/WSA dG. Die Konformation mit der besten Bewertung wurde für die weiteren Berechnungen verwendet.

Für alle MMPs, die aus dem Datensatz hervorgehen, wurde im Anschluss der LORD_FP berechnet. Die VAMMPIRE Datenbank wurde nach den gefundenen Transformationen durchsucht, sofern der Transformationseffekt innerhalb der SAR mindestens 0,5 Log-Einheiten betrug. Wurde die gleiche Transformation in der VAMMPIRE Datenbank ebenfalls mit einem Effekt von mindestens 0,5 Log-Einheiten gefunden, so wurde die Ähnlichkeit der Deskriptoren berechnet. Bei einer Ähnlichkeit von mindestens 0,6 wurde notiert, ob die Tendenz des Effektes übereinstimmte (beide positiv oder beide negativ) oder nicht. Aus diesen Informationen wurde ein gerichteter Graph erstellt, bei dem die Knoten den Substituenten des Datensatzes entsprechen, während die gerichteten Kanten eine Transformation darstellen. Die Einfärbung der Kanten erfolgte je nachdem ob die Tendenz der Effekte übereinstimmte (grün) oder nicht (rot).

2.4 VAMMPIRE Webserver

Sowohl die VAMMPIRE Datenbank als auch der Assistent zur Leitstrukturoptimierung sind über eine webbasierte Benutzeroberfläche erreichbar (http://vammpire.pharmchem.uni-frankfurt.de). Der Webserver wurde mit Hilfe des Python Mikroframeworks Flask 0.10.1 (http://flask.pocoo.org/) implementiert. Die clientseitige 3D-Visualisierung wird mit Hilfe von GLmol (basierend auf WebGL/Javascript) realisiert (https://github.com/biochem-fan/GLmol/). Für die interne Verarbeitung der Eingabemoleküle wurde die Chemieinformatik Software *RDKit* verwendet (http://www.rdkit.org), für die clientseitige Eingabe von Molekülen die Marvin4JS Bibliothek (http://www.chemaxon.com/). Als Datenbank Server fungierte PostgreSQL 9.3 in Verbindung mit dem *RDKit* Datenbankmodul.

3 Ergebnisse und Diskussion

3.1 Analyse der ChEMBL Datenbank

Die ChEMBLdb enthielt zum Zeitpunkt der Anfertigung dieser Arbeit (September 2014) 1.638.394 chemische Verbindungen mit 12.843.338 experimentell bestimmten Aktivitätswerten zu 10.579 Targets. Da diese Aktivitätswerte wiederum aus 57.156 verschiedenen Publikationen und somit auch aus tausenden verschiedener Laboratorien stammen, sollte die Vergleichbarkeit der in der ChEMBLdb hinterlegten Messwerte überprüft werden. Dafür wurden experimentell bestimmte Aktivitätswerte aus unterschiedlichen Publikationen, Assays und Laboratorien paarweise für ein System (eine Molekül-Target-Relation) gegeneinander aufgetragen und der Korrelationskoeffizient (Pearson-Korrelation) bestimmt. Die statistische Analyse wurde separat für K_i-, K_d und IC_{50}-Werte durchgeführt, welche zusätzlich in drei Gütekriterien (alle Paare, Paare mit „CONFIDENCE_SCORE == 9“, Paare mit „CURATED_BY==‘Expert‘“) unterteilt wurden.

Innerhalb der ChEMBLdb ist ein Assay-Target-System genau dann mit der Bezeichnung „CONFIDENCE_SCORE == 9“ versehen, wenn der Assay eindeutig einem Protein zugewiesen wurde und nicht etwa einem homologen Protein oder Protein-Komplex. Ist ein Messwert mit „CURATED_BY==‘Expert‘“ versehen, bedeutet dies, dass der entsprechende Wert von einem Experten manuell in der Primärreferenz überprüft wurde. Da die in der ChEMBLdb hinterlegten Messwerte größtenteils automatisiert aus den unterschiedlichen Publikationen extrahiert werden, können Übertragungsfehler auftreten. Kramer et al.[78] identifizierten z. B. Fehler bei der Zuweisung der richtigen Stereoisomere und Assay-Bedingungen sowie Rundungsfehler und Werte mit falsch extrahierten Einheiten. Des Weiteren wurden redundante Messwerte identifiziert, die auftreten wenn ein Wert mehrfach aus einer früheren Publikation zitiert wurde. Diese Redundanzen würden in einer statistischen Analyse fälschlicherweise zu einer Verbesserung der Korrelation führen, da die entsprechenden Werte nicht *in vitro* reproduziert

wurden. Um die beschriebenen Fehler und Redundanzen einzugrenzen, wurden die in Abschnitt 2.1 beschriebenen Filterschritte eingeführt. Dabei wurden identische Messwerte oder Messwerte, die sich durch Rundung oder eine Änderung der Einheit ineinander überführen ließen, aus der Statistik ausgeschlossen, da es sich mit hoher Wahrscheinlichkeit um Übertragungsfehler oder Redundanzen handelt.

Verglichen mit der statistischen Analyse, die Kramer et al. 2012 veröffentlichten, wurde in dieser Arbeit pro Target nur ein zufälliges Molekül und dazu jeweils genau zwei zufällig ausgewählte Messwerte verglichen. Da einige Targets (z. B. die Carboanhydrase und die Phosphodiesterase) in der ChEMBLdb überrepräsentiert sind und die Varianzen innerhalb der Messwerte target- bzw. assayabhängig sein könnten, würde die Einbeziehung aller Daten unter Umständen eine verfälschte Statistik zur Folge haben. Die zufällige Auswahl des Messwertpaares wurde zehnfach wiederholt und anschließend der Mittelwert der Pearson-Korrelationen bestimmt. In **Abbildung 14** ist die Korrelation der pK_d-Werte in Abhängigkeit der drei Gütekriterien dargestellt.

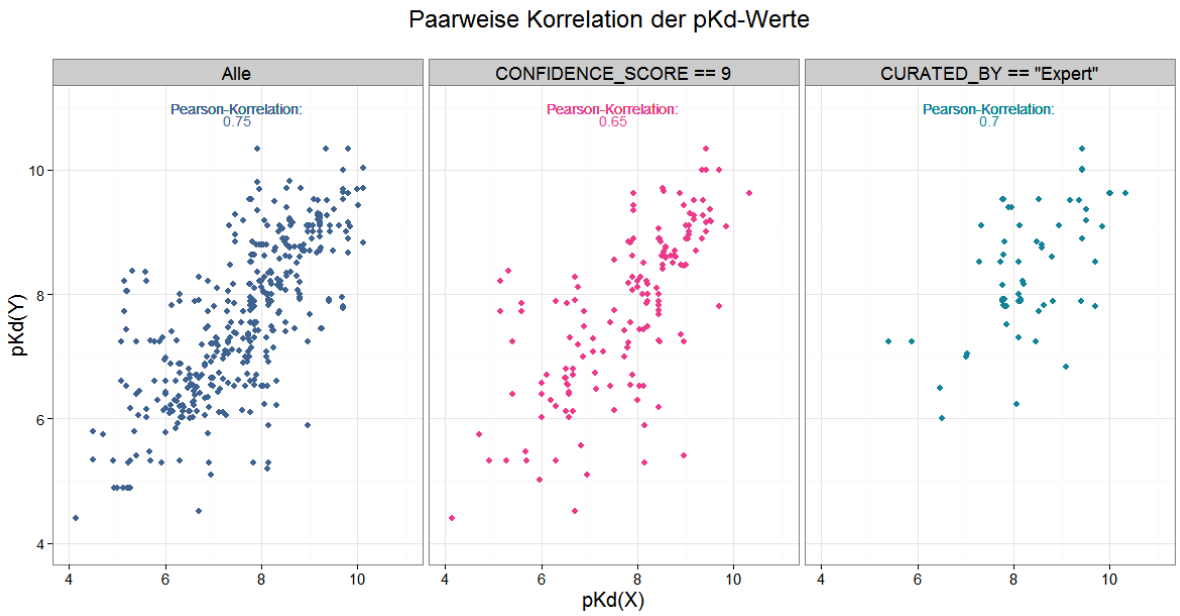


Abbildung 14. Korrelation der pK_d-Werte. Aufgetragen ist jeweils der Messwert aus Labor X gegen den Messwert aus Labor Y, für die gleiche Molekül-Target-Relation. Pro Target wurden zehn Paare zufällig ausgewählt. Die Daten wurden zusätzlich aufgeteilt in drei Güteklassen (alle Paare, Paare mit „CONFIDENCE_SCORE== 9“, Paare mit „CURATED_BY == ‘Expert‘“).

Zunächst fällt auf, dass vor allem bei pK_d-Werten zwischen vier und sechs (Messwerte zwischen 100 µM und 1 µM) größere Abweichungen innerhalb der Messwertpaare auftreten, als bei höheren pK_d-Werten. Der Korrelationskoeffizient für alle Paare liegt bei 0,75 und steigt auf 0,78

wenn jene pK_d-Werte kleiner als sechs nicht berücksichtigt werden. Die mittlere absolute Differenz der Messwerte liegt bei 0,67 und ist damit kleiner als eine Zehnerpotenz. Man kann also von einer guten Vergleichbarkeit der K_d-Werte sprechen, insbesondere wenn es sich um submikromolare Aktivitätswerte handelt. Betrachtet man lediglich jene Paare mit „CONFIDENCE_SCORE==9“, ist eine Verschlechterung der Pearson-Korrelation (0,65) zu erkennen. Dies entspricht nicht der Erwartung, da eine Verfeinerung der Daten (in diesem Fall eine eindeutige Target-Zuweisung) idealerweise zu einer Verbesserung der Korrelation führen sollte. Auch der Filter „CURATED_BY==‘Expert‘“ führt nicht zu einer Verbesserung der Korrelation (0,7), obwohl ein Großteil der Messwertpaare mit pK_d-Werten kleiner als sechs aus dem Datensatz entfernt wurden.

Auch die Ergebnisse für pK_i-Werte (**Abbildung 15**) entsprechen nur teilweise den Erwartungen.

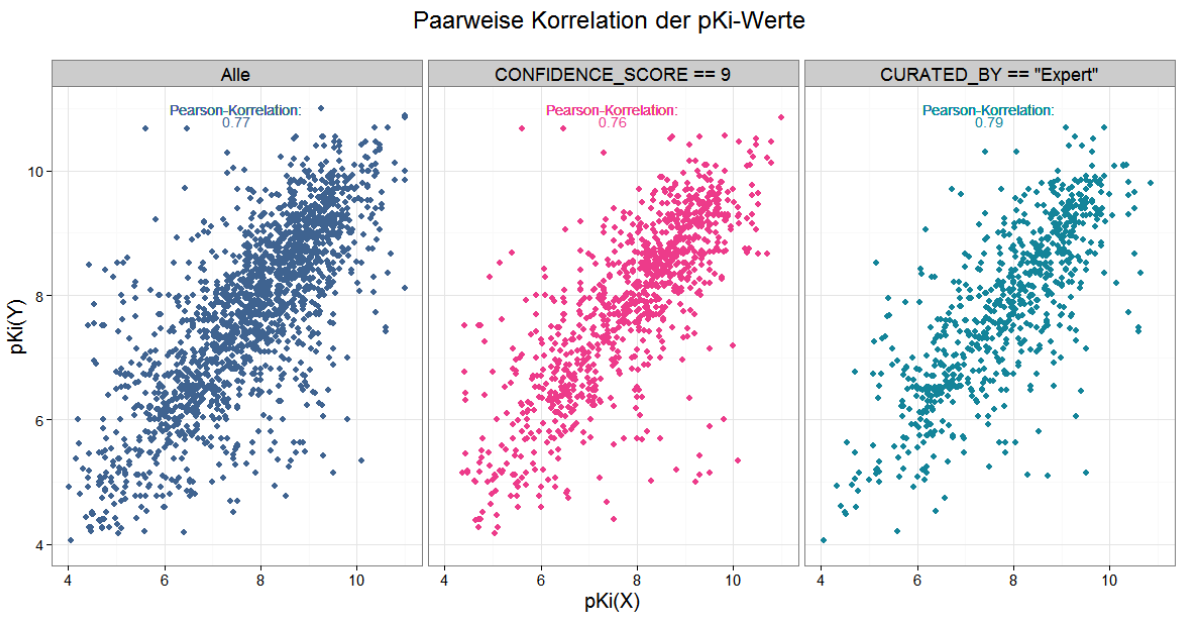


Abbildung 15. Korrelation der pK_d-Werte. Aufgetragen ist jeweils der Messwert aus Labor X gegen den Messwert aus Labor Y für die gleiche Molekül-Target-Relation. Pro Target wurden zehn Paare zufällig ausgewählt. Die Daten wurden zusätzlich aufgeteilt in drei Güteklassen (alle Paare, Paare mit „CONFIDENCE_SCORE== 9“, Paare mit „CURATED_BY == ‘Expert‘“).

Die Pearson-Korrelation für alle Paare beträgt 0,77 und wird durch die von Experten überprüften und gefilterten Werte noch verbessert (0.79). Auch hier wurden einige Ausreißer und einige pK_i-Werte kleiner als sechs aus dem Datensatz entfernt, was im Gegensatz zu den pK_d-Werten den erwarteten Effekt der Verbesserung zeigt. Der „CONFIDENCE_SCORE“ hat aber auch hier keinen positiven Einfluss auf die Korrelation der Messwerte.

Die höhere Variabilität der Messwerte bei niedrigen p-Werten ist am deutlichsten am Beispiel der IC_{50}-Werte zu beobachten (**Abbildung 16**).

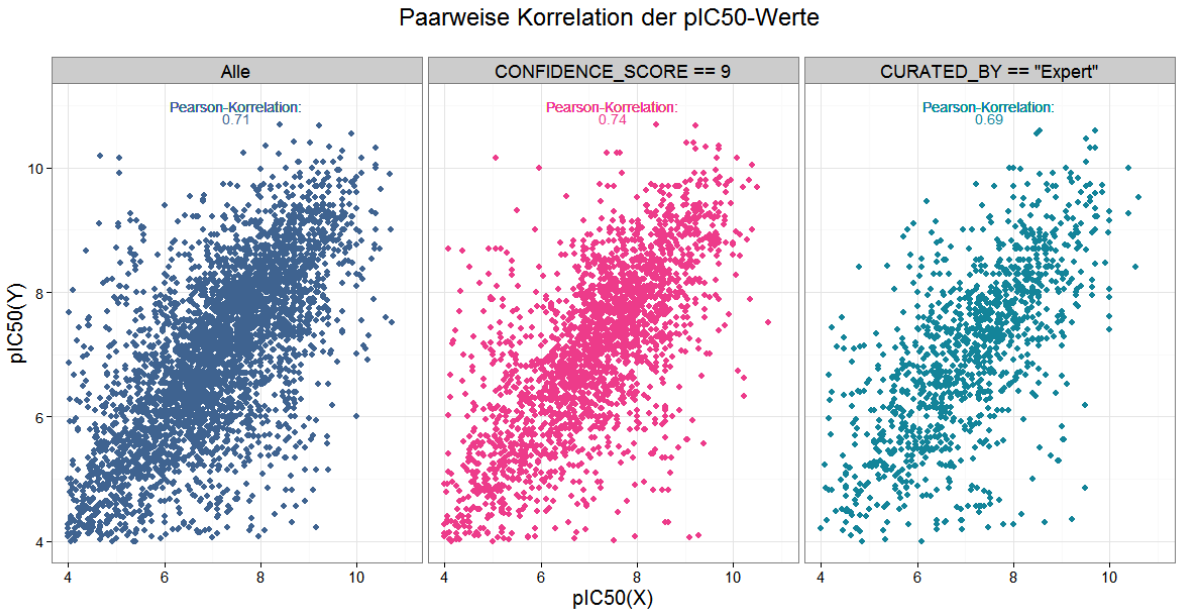


Abbildung 16. Korrelation der pIC_{50}-Werte. Aufgetragen ist jeweils der Messwert aus Labor X gegen den Messwert aus Labor Y für die gleiche Molekül-Target-Relation. Pro Target wurden zehn Paare zufällig ausgewählt. Die Daten wurden zusätzlich aufgeteilt in drei Güteklassen (alle Paare, Paare mit „CONFIDENCE_SCORE== 9“, Paare mit „CURATED_BY == ‘Expert‘“).

Da IC_{50}-Werte im Gegensatz zu K_i- und K_d-Werten keine Bindungskonstanten darstellen und somit von der Substratkonzentration abhängig sind, wurde eine Verschlechterung der Korrelation bei Messwerten aus unterschiedlichen Laboratorien erwartet. Die tatsächliche Korrelation liegt jedoch bei 0,71 und lässt vermuten, dass sich spezifische Assay-Bedingungen für bestimmte Targets durchgesetzt haben oder für die Optimierung und Kalibrierung der Assays in vielen Fällen ein und dieselbe Referenzverbindung verwendet wurde. Zwar sollte mit Hilfe der ChEMBLdb Assay_ID sichergestellt sein, dass die Assays innerhalb eines Messwertpaares nicht identisch sind (Abschnitt 2.1), jedoch wird bereits bei der kleinsten Änderung der Assay-Beschreibung eine neue Assay_ID vergeben. Nur mit Hilfe der Assay_ID ist es daher nicht möglich zwischen ähnlichen und unähnlichen Assay-Bedingungen zu differenzieren.

Zusammenfassend lässt sich sagen, dass im Vergleich zu Messwerten im höheren mikromolaren Bereich eine verbesserte Korrelation bei Aktivitätswerten im submikromolaren Bereich erreicht wird. Die höhere Abweichung der Messwerte bei niedrigeren pK-Werten, könnte man zum Beispiel mit einer schlechteren Löslichkeit der Verbindungen bei steigenden

Konzentrationen in variablen Puffersystemen erklären. Liegt eine Verbindung ab einer bestimmten Konzentration nicht mehr vollständig gelöst vor, kann dies zu einer Verfälschung der Assay-Ergebnisse führen.[94–96] Des Weiteren sind Verbindungen mit submikromolaren Aktivitäten meist sorgfältiger charakterisiert und durch Mehrfachbestimmung validiert. Die Filterung nach „CONFIDENCE_SCORE == 9“ und „CURATED_BY == ‘Expert‘“ führt zu keiner signifikanten Verbesserung der Messwert-Korrelation. Es ist allerdings auch nicht bekannt, in wie vielen Fällen ein Messwert innerhalb der ChEMBLdb tatsächlich dem falschen Target zugewiesen wurde. Da sich die Korrelation nicht signifikant ändert, lässt sich vermuten, dass nur wenige Targets davon betroffen sind. Auch bei den von Experten überprüften Daten lässt sich keine signifikante Verbesserung der Korrelation feststellen. Es entfallen zwar besonders auffällige Ausreißer, die Datenmenge wird aber insgesamt stark reduziert.

Die ChEMBLdb stellt aufgrund der beschriebenen Analyse eine gute Datengrundlage für die Erstellung von QSAR-Modellen dar. Für Modelle mit einer übersichtlichen Anzahl von Molekülen sollte jedoch möglichst die Primärreferenz der Messwerte überprüft werden. Im besten Fall sollten die Messwerte im gleichen Assay und Labor getestet worden sein. Außerdem sollte man sich dessen bewusst sein, dass ein Modell basierend auf den oben gezeigten Daten im Durchschnitt nicht besser sein kann als die intrinsische Korrelation.[78,97]

3.2 VAMMPIRE Datenbank

Die VAMMPIRE (engl. Virtually Aligned Matched Molecular Pairs Including Receptor Environment) Datenbank[98] wurde als Quelle für molekulare Transformationen im Kontext ihrer Rezeptorumgebung implementiert. In der aktuellen Version (VAMMPIRE092014) besteht sie aus 17.602 MMPs assoziiert mit dem Effekt einer Transformation auf die Bindungsaffinität zu 241 verschiedenen Targets. Jedem Eintrag der Datenbank wurden außerdem Informationen zu Publikation, Bindungstyp (K_i, K_d, IC_{50}) sowie über Aminosäuren in unmittelbarer Umgebung einer Transformation zugeordnet. In **Abbildung 17** ist ein Eintrag der VAMMPIRE Datenbank schematisch am Beispiel von zwei Faktor-Xa-Liganden dargestellt.

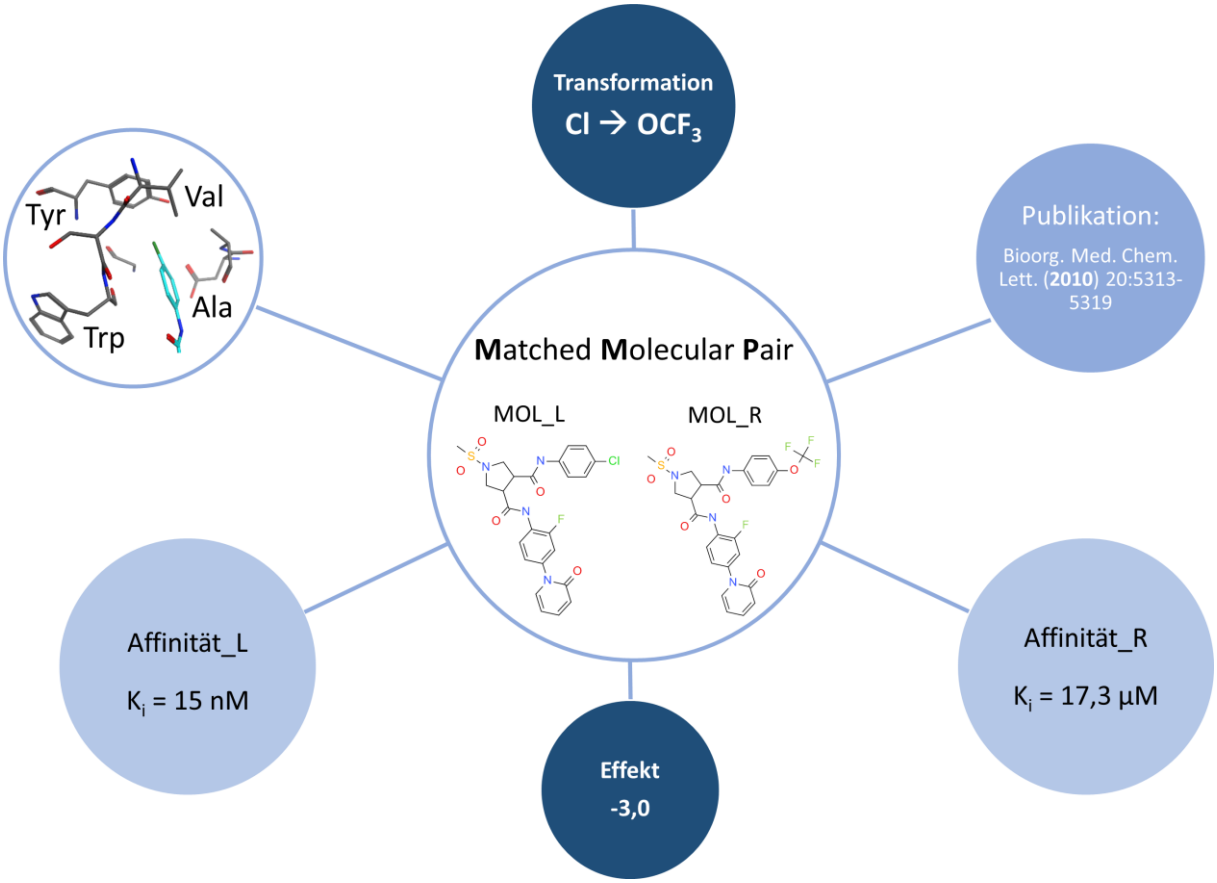


Abbildung 17. Informationen eines MMPs innerhalb der VAMMPIRE Datenbank. Dargestellt ist ein MMP mit einer gerichteten Transformation (Cl→OCF_3) am Beispiel von zwei Faktor-Xa-Liganden. Die Affinitätsdaten für beide Moleküle (Mol_L und Mol_R) wurden experimentell bestimmt und innerhalb einer Publikation veröffentlicht. Der Transformationseffekt von -3,0 entspricht einer Verschlechterung der Affinität um drei Zehnerpotenzen. Die bioaktive Konformation von Mol_L ist bekannt (PDBcode: 2XBX) und somit auch die unmittelbare Aminosäureumgebung des Substituenten „Cl“.

Im Laufe dieser Arbeit wurden drei Typen von MMPs definiert, die sich in ihrer Qualität unterscheiden. Bei Typ-I-MMPs handelt es sich um Molekülpaare mit bekannter bioaktiver Konformation, weshalb sie als 3D-MMPs von hoher Qualität angenommen werden. Bei MMPs vom Typ-II ist die bioaktive Konformation lediglich für eines der beiden Moleküle bekannt, welche als Basis für die Konformationsvorhersage des zweiten Moleküls dient. Diese vorhergesagte Konformation kann anschließend wieder als Basis dienen um MMPs vom Typ-III vorherzusagen. Da die bioaktive Konformation für keines der Moleküle innerhalb eines Typ-III-MMPs tatsächlich bekannt ist, wird dieser als unsicherster Typ angenommen. Die Anzahl der MMPs pro MMP-Typ und Bindungstyp sind im Folgenden (**Tabelle 5**) dargestellt:

Tabelle 5. Anzahl der MMPs pro MMP-Typ und experimentell bestimmter Bindungstypen.

	Alle	K_d	K_i	IC_{50}
Typ-I-MMPs	938	159	421	358

Typ-II-MMPs	3.341	230	931	2.180
Typ-III-MMPs	13.323	337	4.012	8.974

Im Mittel wurden 30 MMPs pro Target identifiziert, wobei die Anzahl der MMPs pro Target stark variiert. Für mehr als die Hälfte der Targets wurden maximal fünf MMPs gefunden während für die sieben häufigsten Targets jeweils mehr als 500 MMPs identifiziert wurden (**Abbildung 18**).

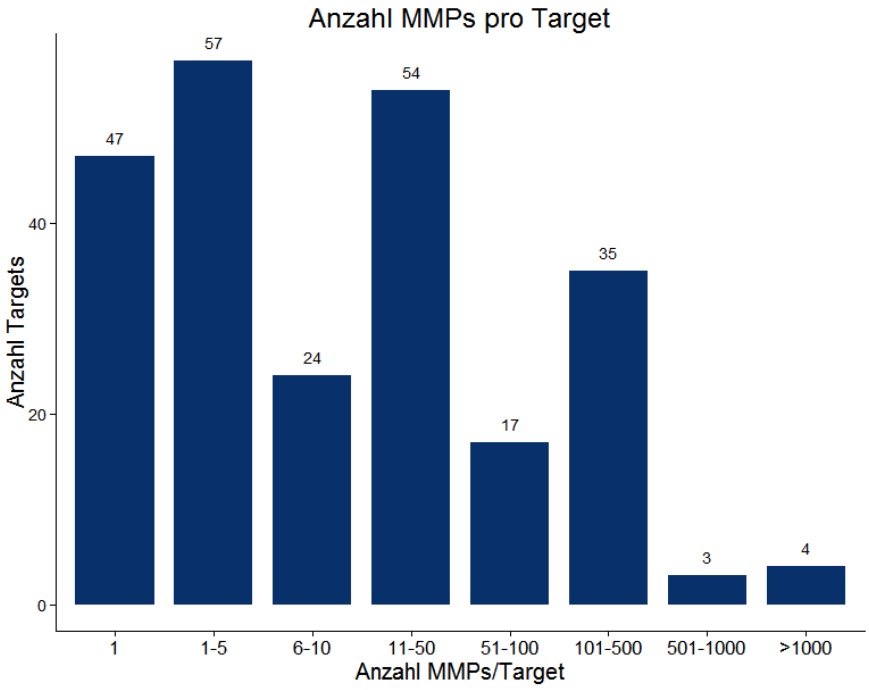


Abbildung 18. Anzahl gefundener MMPs pro Target. Dargestellt ist die Anzahl der Targets, die mit einer definierten Anzahl von MMPs assoziiert werden (aufgeteilt in acht Bins).

Zu den Targets mit den meisten MMPs gehören beispielsweise die Carboanhydrase II sowie der vaskuläre endotheliale Wachstumsfaktor Rezeptor II (Abk. VEGFR2, engl. Vascular endothelial growth factor receptor II). Die hohe Anzahl identifizierter MMPs lässt sich hier vor allem durch die große Anzahl publizierter Bioaktivitätsdaten erklären. Für die Carboanhydrase II sind 7.160 K_i- und 1.668 IC_{50}-Werte, für den VEGFR2 1.343 K_i- und 6.984 IC_{50}-Werte hinterlegt (Stand: März 2015). Die Carboanhydrase II ist zudem das Target mit den meisten Kristallstrukturen innerhalb der VAMMPIRE Datenbank (ursprünglich 554 in der PDB, davon 106 in der PDBbind).

Die maximale Größe der Substituenten einer Transformation wurde für zyklische Substituenten auf neun Nicht-Wasserstoffatome (z. B. 3-fach-substituierte Phenylgruppe) und für azyklische Substituenten auf 5 Nicht-Wasserstoffatome (z. B. Trifluormethoxy) festgelegt. Die

durchschnittliche Größe eines azyklischen Substituenten liegt bei 2,4 Nicht-Wasserstoffatomen. Am häufigsten (in 23 % der Fälle) bestehen Substituenten aus genau einem Nicht-Wasserstoffatom (dies entspricht z. B. einer Methylgruppe). Für zyklische Substituenten liegt die durchschnittliche Größe bei sieben Nicht-Wasserstoffatomen, welche gleichzeitig die häufigste Substituentgröße darstellt (in 29 % der Fälle). Die allgemeine Verteilung der Substituentgrößen innerhalb der VAMMPIRE Datenbank ist in **Abbildung 19** dargestellt.

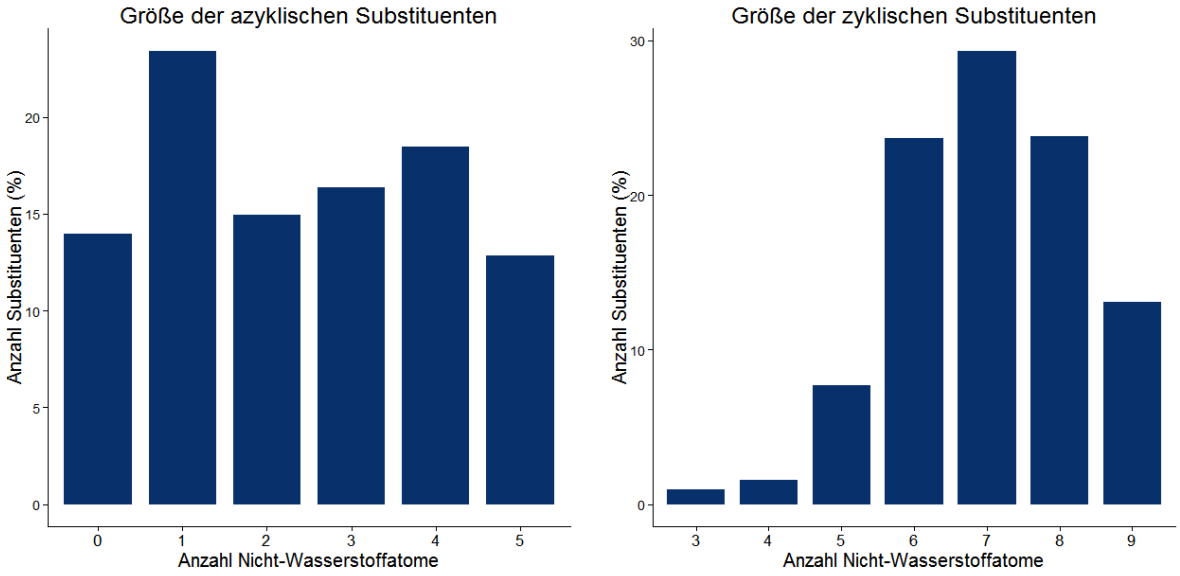


Abbildung 19. Anzahl der Substituenten pro Substituentgröße (Anzahl Nicht-Wasserstoffatome) für zyklische und azyklische Substituenten. Aufgetragen ist jeweils die Anzahl der Substituenten (in Prozent) gegen die Anzahl Nicht-Wasserstoffatome eines Substituenten. Null Nicht-Wasserstoffatome stehen für die Substitution eines Wasserstoffatoms. Es wurde die Gesamtheit aller Substituenten (Start- und Zielsubstituenten) innerhalb der VAMMPIRE Datenbank.

Um die Wahrscheinlichkeit zu erhöhen, dass die beiden Moleküle eines Typ-II- oder Typ-III-MMPs die gleiche Orientierung innerhalb der Bindetasche einnehmen wurde festgelegt, dass der Substituent maximal halb so viele Nicht-Wasserstoffatome enthalten darf wie der Rest des Moleküls (Kontext). Um diese Hypothese zu bestätigen wurden Typ-I-MMPs unter Berücksichtigung dieser Einschränkungen untersucht. Dazu wurde die Wurzel des mittleren quadratischen Abstands (Abk. RMSD, engl. root mean square deviation) der maximalen gemeinsamen Substruktur (Abk. MCS, engl. maximum common substructure) der Moleküle berechnet. In 81 % der Fälle betrug der RMSD der MCS beider Moleküle höchstens 1 Å, was auf eine sehr ähnliche Orientierung der Moleküle schließen lässt.

3.2.1 Vorhersage der bioaktiven Konformation

Für die Vorhersage der bioaktiven Konformation der nicht ko-kristallisierten Liganden wurde molekulares Docking mit „Pharmakophor-Platzierung“ verwendet (siehe Abschnitt 2.2.4). Dabei wird anhand des gemeinsamen Kontextes der Moleküle ein Pharmakophormodell definiert, welches durch die Beschreibung von potentiellen Interaktionen mit dem Rezeptor indirekt die Orientierung des zu platzierenden Liganden vorgibt. In einem solchen Modell werden Regionen definiert (Atome oder funktionelle Gruppen), in denen der Ligand spezifische Interaktionstypen aufweist. Ein Beispiel für ein solches Modell ist in **Abbildung 20** dargestellt.

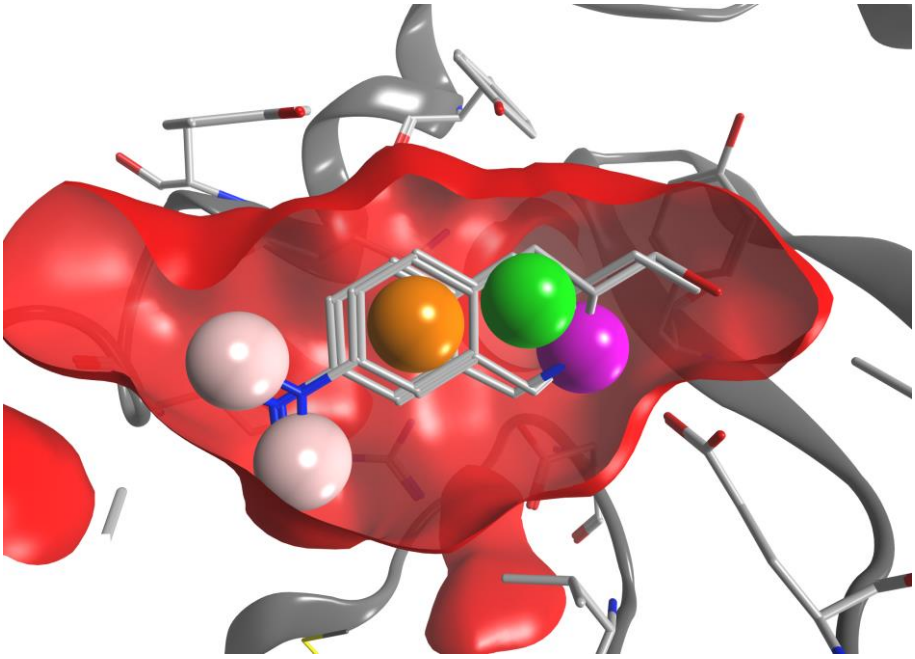


Abbildung 20. Exemplarische Darstellung eines MMPs im Kontext der Rezeptorumgebung. Bei der Transformation handelt es sich um den Austausch einer Methylgruppe durch eine Hydroxygruppe. Das Pharmakophormodell, am gemeinsamen Kontext ist in Form von farbigen Sphären dargestellt (rosa: H-Brücken-Donor oder H-Brücken-Akzeptor, orange: aromatisch, grün: hydrophob, violett: H-Brücken-Donor)

Das erzeugte Pharmakophormodell wird vom Docking-Platzierungsalgorithmus zur Erstellung der Konformationen verwendet. Beträgt der RMSD der MCS beider Moleküle nach einer anschließenden Energieminimierung innerhalb des Rezeptors weniger als 1 Å, so wird das erzeugte 3D-MMP als gültig angenommen.

Wie bereits erwähnt, zeigte die Untersuchung der Typ-I-MMPs, dass der RMSD der MCS zweier Moleküle innerhalb eines MMPs in 81 % der Fälle höchstens 1 Å beträgt. Diese Beobachtung deckt sich mit der Hypothese, dass sehr ähnliche Moleküle auch ähnliche Interaktionen mit dem Target eingehen und somit eine ähnliche Orientierung innerhalb der Bindetasche

einnehmen. Des Weiteren konnte in einer vorangegangenen Studie gezeigt werden, dass die Vorhersage der bioaktiven Konformation durch molekulares Docking besonders dann gute Ergebnisse liefert, wenn die verwendete Kristallstruktur einen ähnlichen Liganden enthält.[99] Je größer der gemeinsame Kontext zweier Moleküle eines MMPs und je kleiner die Substituenten der Transformation, umso höher ist die Wahrscheinlichkeit, dass beide Moleküle eine ähnliche Orientierung innerhalb der Bindetasche annehmen.

Die Änderung der Konformation durch kleinere Substituenten ist jedoch nicht ausgeschlossen. Ein Beispiel dafür sind die in **Abbildung 21** dargestellten Thrombin-Inhibitoren. Hier führt die Einführung eines Fluoratoms an einem aromatischen Ring zu einer Konformationsänderung sowie zu einer fünffachen Steigerung der Inhibition.[100] Diese drastische Änderung der Konformation kann durch die Entstehung einer dipolaren Interaktion zwischen dem Stickstoffatom der Peptidbindung und dem eingeführten Fluoratom in einer Distanz von 3,4 Å erklärt werden, welche ohne eine Änderung der Konformation nicht ausgebildet werden könnte.

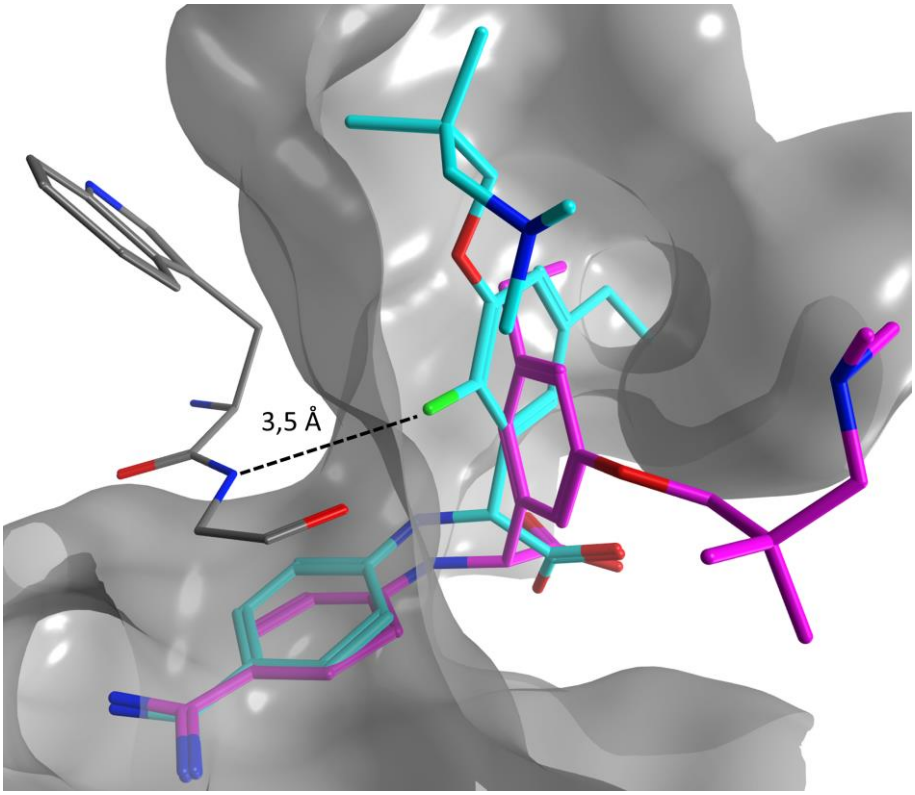


Abbildung 21. Drastische Änderung der Konformation durch Addition eines Fluoratoms. Das MMP wird durch zwei ko-kristallisierte Thrombin-Inhibitoren dargestellt. Das nicht fluorierte Analogon ist in violett (PDBcode: 2V3H), das fluorierte Analogon in hellblau (PDBcode 2V3O) dargestellt. Die dipolare Wechselwirkung zwischen dem Fluor- und dem Stickstoffatom der Peptidbindung ist durch eine gestrichelte Linie dargestellt.

Die Vorhersage einer solchen Konformationsänderung ist durch Docking mit Pharmakophor-Platzierung nicht möglich, da das Pharmakophormodell die Orientierung des Liganden vorgibt. An dieser Stelle würde das MMP entweder verworfen werden, wenn es z. B. durch die Einführung eines Fluoratoms zu einer Kollision mit dem Rezeptor kommt, oder es würde ein fehlerhaftes 3D-MMP erzeugt werden. Betrachtet man das oben gezeigte Beispiel nicht als Typ-I-MMP, sondern geht davon aus, dass lediglich die bioaktive Konformation des nicht-fluorierten Analogons bekannt ist, so würde die Vorhersage der Konformation durch Docking mit Pharmakophor-Platzierung zu einem fehlerhaften Typ-II-MMP führen (**Abbildung 22**). Da in diesem speziellen Fall genügend Raum für ein Fluoratom gegeben ist, würde lediglich eine kleinere Konformationsänderung der räumlich nahe gelegenen Säure-Gruppe hervorgerufen werden.

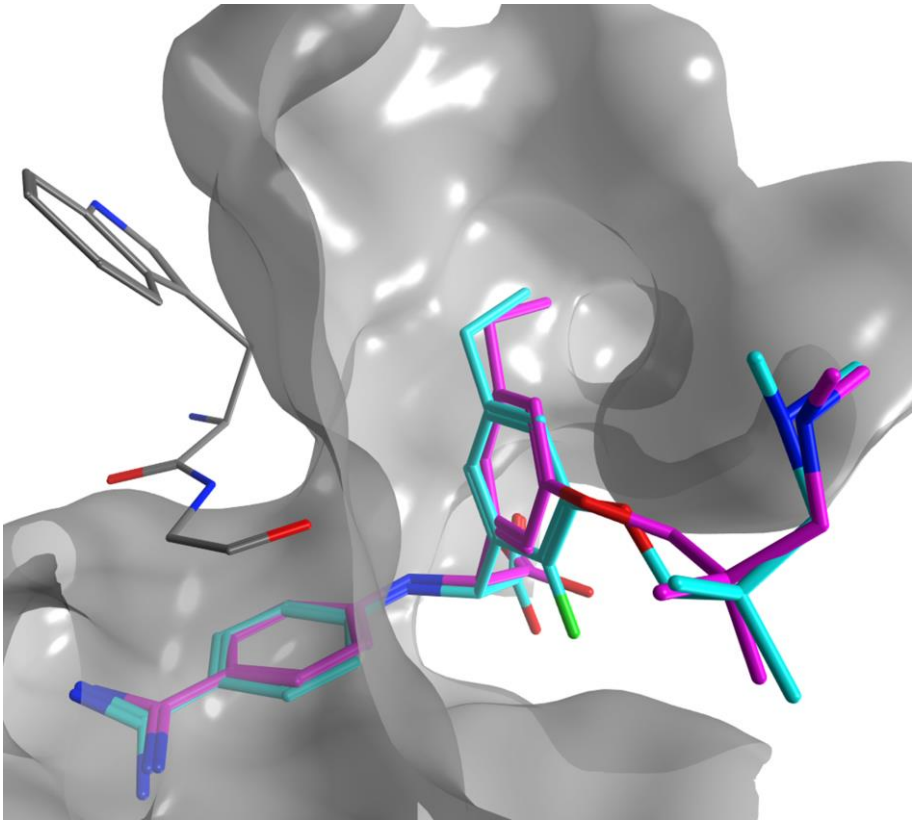


Abbildung 22. Falsche Vorhersage der Konformation durch Docking mit Pharmakophor-Platzierung. Das Typ-II-MMP entsteht durch den nicht fluorierten ko-kristallisierten Thrombin-Inhibitor mit dem PDBcode 2V3H (violett) und das fluorierte Analogon, dessen bioaktive Konformation durch Docking mit Pharmakophor-Platzierung vorhergesagt wurde (hellblau). Die tatsächliche bioaktive Konformation ist in **Abbildung 21** dargestellt.

Es gibt weitere Spezialfälle, bei denen das Docking mit Pharmakophor-Platzierung an seine Grenzen stößt. **Abbildung 23** zeigt zwei Tyrosinkinase-Inhibitoren bei denen sich die Orientierung der Substituenten nach dem Austausch einer Methyl- durch eine Aminogruppe

signifikant ändert. Auch hier entstehen durch den Austausch neue Interaktionsmöglichkeiten mit dem Rezeptor. Der Effekt auf die Aktivität des Liganden durch den Austausch beträgt 0,3 Log-Einheiten und liegt damit innerhalb der Messungenauigkeit. Die Änderung der Konformation kann durch die Ausbildung einer Wasserstoffbrückenbindung zwischen dem primären Amin und dem Sauerstoffatom der Peptidbindung erklärt werden, durch die eine Rotation des Benzensulfonamids erzwungen wird. Im Gegensatz zu den oben gezeigten Thrombin-Inhibitoren ändert sich zwar die Konformation des gemeinsamen Kontextes der Moleküle nicht (RMSD = 1,001 Å), trotzdem ist die lokale Umgebung der Transformation unterschiedlich und somit das MMP ungeeignet.

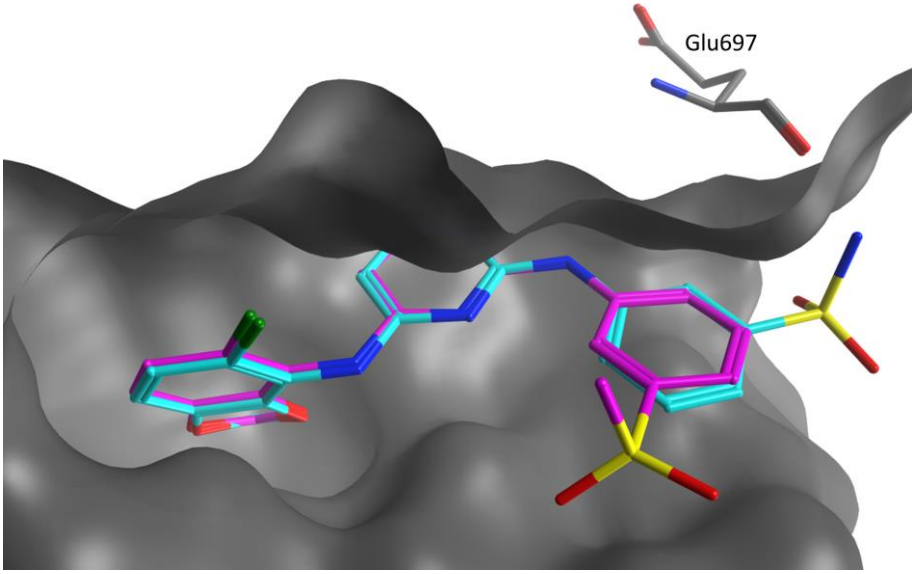


Abbildung 23. MMP Beispiel für die Änderung der chemischen Umgebung durch eine kleine Transformation. Gezeigt ist die Substitution einer Methylgruppe durch eine Aminogruppe zum Benzensulfonamid. Das MMP wird durch zwei Tyrosinkinase-Inhibitoren gebildet, welche in violett (PDBcode: 2VWY) und hellblau (PDBcode: 2VWX) dargestellt sind. Durch die Einführung der Aminogruppe ergibt sich eine Interaktion mit dem Rückgrat des Glutamat 697 (Glu697), welches eine Wasserstoffbrücke zum primären Amin ausbildet. Die Substituenten des MMPs liegen mit einer Distanz von 7,2 Å in verschiedenen Rezeptorumgebungen.

Die Änderung der Orientierung des Substituenten kann auch hier durch Docking mit Pharmakophor-Platzierung nicht vorhergesagt werden, da das Pharmakophormodell die Position der beiden Sauerstoffatome des Sulfonamids, welche zum gemeinsamen Kontext der Moleküle gehören, vorgibt. Die vorhergesagte Konformation des Benzensulfonamids ist in **Abbildung 24** dargestellt.

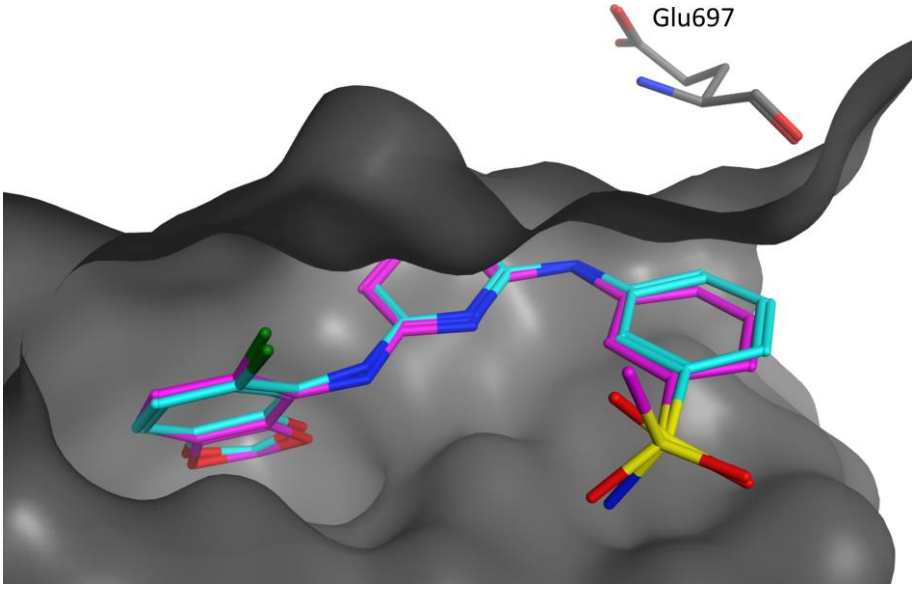


Abbildung 24. Falsche Vorhersage der Konformation durch Docking mit Pharmakophor-Platzierung. Das Typ-II-MMP entsteht durch den Austausch der Methylgruppe des ko-kristallisierten Tyrosinkinase-Inhibitors, dargestellt in violett (PDBcode: 2VWY) durch eine Aminogruppe, dessen bioaktive Konformation durch Docking mit Pharmakophor-Platzierung vorhergesagt wurde (hellblau). Die tatsächliche bioaktive Konformation ist in **Abbildung 23** dargestellt.

Starke Konformationsänderungen durch die Einführung von kleinen Substituenten wie im Beispiel der Fluorierung des Thrombin-Inhibitors sind sehr selten und setzen voraus, dass die neu entstandene Interaktion eine signifikante Steigerung der Affinität im Vergleich zu den zuvor eingegangenen Interaktionen des Grundgerüstes mit sich bringt. Aus diesem Grund wird das Risiko der Erzeugung fehlerhafter Typ-II- und Typ-III-MMs zugunsten der, durch Hinzunahme dieser MMP-Typen, 18-fach größeren 3D-MMP-Datenbank eingegangen. Durch die stetig wachsende Anzahl gelöster 3D-Strukturen von Protein-Ligand-Komplexen, kann auch die Qualität der Datenbank stetig verbessert werden.

Eine alternative Möglichkeit zur Vorhersage der bioaktiven Konformation ohne Vorgabe der Orientierung des Liganden wäre beispielsweise Docking ohne Definition eines Pharmakophormodells. Hier würden verschiedene Bindemodi erzeugt und anschließend mit einer Bewertungsfunktion beurteilt. In einer aktuellen Studie von Li et al. (2014)[37] konnte gezeigt werden, dass die meisten Bewertungsfunktionen, je nach verwendeter Methode, in der Lage sind 60-90 % der bioaktiven Konformationen unter den Top-Drei zu identifizieren. In etwa 50-85 % der Fälle stellt die am besten bewertete Konformation die bioaktive Konformation dar. Auch hier gibt es allerdings potentielle Fehlerquellen, je nachdem welche Bewertungsfunktion für welchen Protein-Ligand-Komplex verwendet wird. Die Kombination verschiedener Docking Programme und Scoring Funktionen könnte verwendet werden um mit einer sogenannten

Konsens-Entscheidung eine bessere Qualität der vorhergesagten Konformationen zu erhalten.[101–103] In einer Studie von Houston et al. (2013)[102] konnte anhand eines diversen Datensatzes von Protein-Ligand-Komplexen gezeigt werden, dass die Anzahl der korrekt vorhergesagten Konformationen von 64 % (Vorhersagerate des besten Docking Programms) auf 82 % (Vorhersagerate der Konsens-Entscheidung) gesteigert werden konnte.

3.2.2 Die chemische Umgebung einer Transformation

Mit der Definition der chemischen Umgebung sollen indirekt die relevanten Interaktionen eines Substituenten mit dem Target beschrieben werden. Alle Nicht-Wasserstoffatome des Proteins die innerhalb eines Radius von 4,5 Å um eines der Nicht-Wassersoffatome des Substituenten liegen, werden als potentielle Interaktionspartner angesehen. Da sich das Ausmaß der chemischen Umgebung eines Substituenten mit seiner Größe verändert und die Koordinaten der Substituenten eines MMPs nie identisch sind, wurde jeweils für beide Substituenten eines MMPs die chemische Umgebung bestimmt. Es wurden vorab sechs Rezeptor-Atomtypen definiert, die sich aus dem PDB-Atomtypen und der jeweiligen Aminosäure ergeben (ACC: H-Brücken-Akzeptor, DON: H-Brücken-Donor, ARO: Aromatisch, DON/ACC: H-Brücken-Donor oder Akzeptor, HYD: Hydrophob, POL: Polar). Das Stickstoffatom einer Peptidbindung (PDB-Atomtyp: „N“) ist beispielsweise grundsätzlich als H-Brücken-Donor, das Sauerstoffatom (PDB-Atomtyp: „O“) als H-Brücken-Akzeptor, das daran gebundene Kohlenstoffatom (PDB-Atomtyp: „C“) als polar und das Cα-Atom (PDB-Atomtyp: „CA“) als hydrophob klassifiziert. Der PDB-Atomtyp CD1 hingegen ist z. B. als hydrophob definiert, wenn er in der Aminosäure Leucin (L) vorkommt, während er in der Aminosäure Phenylalanin (F) als aromatisch definiert ist (für die vollständige Zuweisung der Atomtypen siehe Tabelle A1 im Anhang). Die Vereinfachung der chemischen Umgebung mit Hilfe dieser Atomtypen soll eine Abstraktion der potentiellen Interaktionen des Substituenten mit seiner chemischen Umgebung darstellen. Diese Klassifizierung von Atomtypen birgt allerdings auch Risiken, wenn ein zugewiesener Atomtyp der potentiellen Interaktion nicht gerecht wird. So kann beispielsweise das Schwefelatom in Methionin am ehesten als hydrophob klassifiziert werden, die potentiellen Interaktionen sind jedoch nicht mit den hydrophoben Seitenketten eines Valins oder Leucins gleichzusetzen. Quantenchemische Berechnungen konnten zum Beispiel zeigen, dass das Schwefelatom in Methionin in der Lage ist gerichtete Halogenbindungen einzugehen.[104] Auch für geladene Atome wie beispielsweise das Stickstoffatom in der Seitenkette von Arginin (H-Brücken-Donor),

welches in der Lage ist starke Kation-Aryl-Interaktionen auszubilden, wurde kein spezifischer Atomtyp definiert. Aus diesem Grund wurde zusätzlich zu den Atomtypen auch die 3-Buchstaben-Identifizierung der entsprechenden Aminosäuren in der Datenbank hinterlegt. Dabei wurde zusätzlich zwischen Rückgrat- und Seitenketten-Interaktionen unterschieden, je nachdem ob sich das Atom im Rückgrat oder in der Seitenkette der Aminosäure befindet.

Die exakte Modellierung einzelner Protein-Ligand-Interaktionen und der damit verbundene Beitrag zur Affinität eines Liganden ist auch heute noch eine große Herausforderung. Zwar können typische Energien von Salzbrücken (~2 kcal/mol), H-Brücken-Bindungen (~1 kcal/mol), hydrophoben (~0,7 kcal/mol) und aromatischen Interaktionen (~1-3 kcal/mol) ungefähr abgeschätzt werden[105–107], jedoch ist die Formulierung von Termen zur Beschreibung von Desolvatisierung sowie entropischen und repulsiven Beiträgen immer noch problematisch.[108] Kraftfeldbasierte Methoden zur Abschätzung der freien Bindungsenergie werden durch die nicht-kovalenten Terme der klassischen molekularen Mechanik dargestellt. Dabei werden Van-der-Waals-Wechselwirkungen durch das Lennard-Jones-Potential beschrieben, während die Coulomb-Energie elektrostatische Wechselwirkungen darstellt. Empirische Methoden schätzen ebenfalls die freie Energie der Bindung, basieren jedoch auf experimentellen Daten, welche als Grundlage für die Parametrisierung der Funktion dienen.[109] Der in dieser Arbeit verfolgte Ansatz gehört zu den wissensbasierten Methoden. Hier dient klassischerweise die inverse Formulierung des Boltzmann-Gesetzes als Basis, wobei die Häufigkeit eines Atompaares in einer bestimmten Distanz im Vergleich zu einer Hintergrundverteilung statistisch ausgewertet wird. Da im Falle der MMPs eine direkte Beziehung zwischen einer Transformation in einer spezifischen Umgebung und dem Effekt auf die Bindungsaffinität hergestellt werden soll, ist eine Abstraktion dieser Umgebung notwendig, welche durch die Vereinfachung der Atomtypen umgesetzt wurde. Damit erhöht sich die Anzahl der observierten Transformationen in einer nicht identischen, aber ähnlichen chemischen Umgebung. Im Vergleich zu klassischen wissensbasierten Methoden besteht hier der Vorteil, dass auch negative Beiträge zur Bindungsaffinität einbezogen werden können.

Die Größe der VAMMPIRE Datenbank reicht zwar noch nicht aus um eine allgemeine wissensbasierte- oder empirische Funktion zur Bewertung von Protein-Ligand-Komplexen abzuleiten, jedoch können die gewonnen Informationen z. B. in Kombination mit anderen

Bewertungsfunktionen genutzt werden. Die Anwendung der Datenbank auf das spezifische Problem der Leitstrukturoptimierung wird in Kapitel 3.3 dargestellt.

3.2.3 Webserver

Für den Zugriff auf die VAMMPIRE Datenbank wurde ein Webserver implementiert und eine Schnittstelle für die gezielte Datenbanksuche bereitgestellt. Hier kann nach einzelnen Substituenten oder Transformationen wahlweise exakt oder als Substruktur gesucht werden. Außerdem kann nach PDB_IDs, ChEMBLdb_IDs sowie nach spezifischen Aminosäuren, die in der chemischen Umgebung eines Substituenten vorkommen, gesucht werden. Ein Beispiel für die Nutzung der VAMMPIRE Datenbank ist in **Abbildung 25** gegeben. Hier wird die gezielte Suche nach Transformationen von Fluor nach Chlor in beliebigen Molekülen, Targets und chemischen Umgebungen dargestellt. Die Ergebnisse werden aufgelistet und nähere Details können durch die Auswahl eines Eintrags angezeigt werden. Dazu gehört die Darstellung des MMPs im Kontext der Rezeptorumgebung durch eine interaktive 3D-Visualisierung sowie Details zu den Aminosäuren, die sich in der unmittelbaren Umgebung der Transformation (4,5 Å) befinden. Des Weiteren führen Web-Links zu den Details der entsprechenden Targets, Molekülen und Publikationen, aus denen die Aktivitätswerte extrahiert wurden.

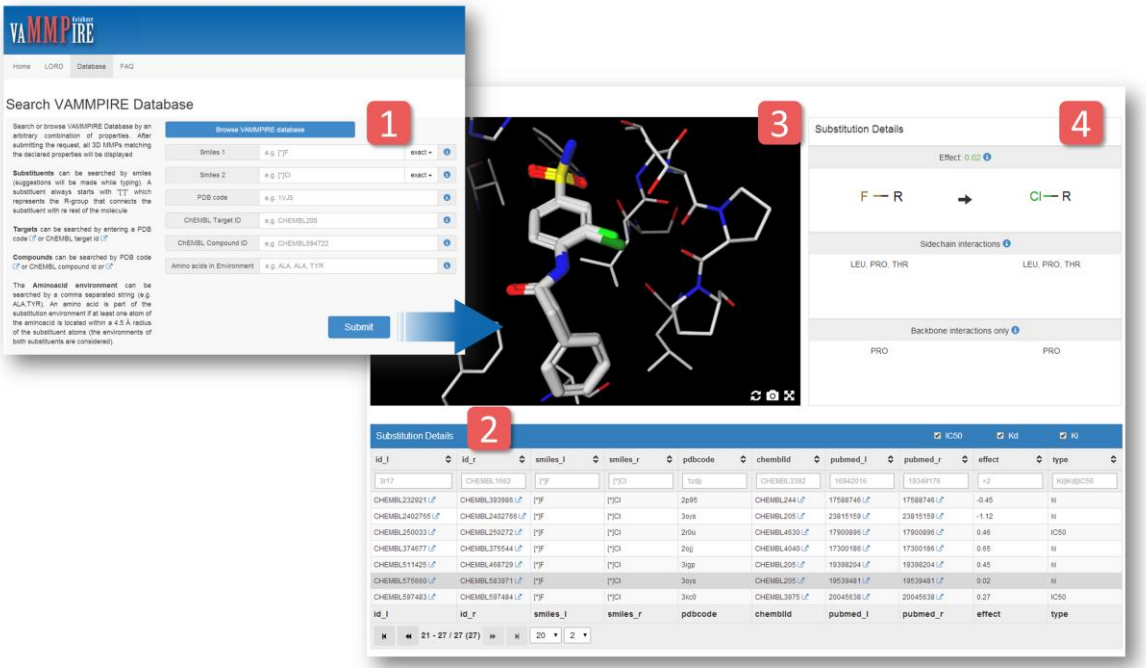


Abbildung 25. VAMMPIRE Database Webserver. 1) Suchfelder für die Datenbank Anfrage. 2) Tabellarische Zusammenfassung der Ergebnisse. 3) Interaktive 3D-Darstellung des ausgewählten MMPs. 4) Details zur Transformation des Ausgewählten MMPs (Seitenketten- und Rückgrat-Interaktionen mit entsprechenden Aminosäuren).

Die VAMMPIRE Datenbank wurde für die Nutzung von Medizinalchemikern entwickelt, die beispielsweise Effekte von spezifischen Transformationen allgemein oder an bestimmten Targets untersuchen möchten. Sie enthält wertvolle Informationen, die zum Beispiel für gezielte Leitstrukturoptimierungen eingesetzt werden können, sofern man die 3D-Struktur des betrachteten Targets bzw. die interagierenden Aminosäuren kennt.

3.3 VAMMPIRE-LORD

VAMMPIRE-LORD (kurz LORD, engl. Lead Optimization by Rational Design)[110] ist ein auf der VAMMPIRE Datenbank basierender Ansatz zur gezielten Leitstruktur-Optimierung, der die folgende Annahme voraussetzt:

> *Eine molekulare Transformation führt in ähnlichen Rezeptor- und Ligand-Umgebungen zu ähnlichen Effekten auf die Bindungsaffinität des Liganden, unabhängig davon um welchen Rezeptor oder Liganden es sich handelt.*

Wie bereits in Abschnitt 1.2 erwähnt, beschränken sich die meisten MMP-Methoden, die zur Vorhersage von Transformationseffekten entwickelt wurden, auf spezifische Targets. Diese

Methoden setzen voraus, dass eine Vielzahl von Liganden für das untersuchte Target bereits bekannt sind und somit genügend MMPs für eine Vorhersage gebildet werden können. VAMMPIRE-LORD stellt eine neuartige, targetübergreifende Methode dar, welche lediglich die unmittelbare chemische Umgebung einer molekularen Transformation berücksichtigt. Dies hat den Vorteil, dass eine vielfach höhere Anzahl MMPs erzeugt werden kann, auf deren Basis auch Targets mit nur wenigen bekannten Liganden untersucht werden können.

3.3.1 Atompaar-Deskriptor LORD_FP

Um sowohl die chemische Umgebung einer Transformation innerhalb des Rezeptors als auch die lokale Umgebung einer Transformation innerhalb des Liganden mathematisch zu beschreiben, wurde der LORD-Fingerprint (Abk. LORD_FP) entwickelt. Dabei handelt es sich um einen neuartigen Deskriptor zur Darstellung von Protein-Ligand-Interaktionen. Er gehört zur Klasse der topologischen Atompaar-Deskriptoren und ist angelehnt an den 2003 veröffentlichten CATS3D.[45] LORD_FP wird aus drei verschiedenen Interaktionstypen zusammengesetzt. Protein-Protein-Interaktionen (PPIs) beschreiben Atompaare aus Proteinatomen, während Ligand-Ligand-Interaktionen (LLIs) Atompaare aus Ligand-Atomen beschreiben, die in unmittelbarer Umgebung des betrachteten Substituenten auftreten und eine definierte Distanz aufweisen (siehe Abschnitt 2.3.1). Protein-Ligand-Interaktionen (PLIs) hingegen beschreiben Atompaare in definierter Distanz zwischen Protein- und Ligand-Atomen. Neben der Beschreibung von Protein-Ligand-Interaktionen sollen durch den LORD_FP auch die Vorteile der 2D-MMP Methoden genutzt werden, welche den unmittelbaren Kontext einer Transformation innerhalb des Liganden berücksichtigen (siehe Abschnitt 1.2.1).

Die eindeutige Zuweisung der in Abschnitt 3.2.2 definierten Rezeptor-Atomtypen ist für die Atome innerhalb des Liganden nicht möglich. Wie bei den Atomen der Aminosäuren ist die Zuweisung des Typen abhängig vom Kontext in dem sich das entsprechende Atom befindet und würde deshalb eine Substruktursuche voraussetzen. Zudem würde eine Eingrenzung auf sechs Interaktionstypen den vielfältigen Interaktionsmöglichkeiten eines Liganden nicht gerecht werden. Aus diesem Grund wurden die sogenannten EState-Atomtypen[92] zur Typisierung der Ligand-Atome eingesetzt. EState-Atomtypen beschreiben das chemische Element des Atoms sowie die Anzahl der Bindungen zu den benachbarten Atomen und stellen im Vergleich zu den Rezeptoratomtypen eine spezifischere Typisierung dar. Der Nachteil hierbei ist die relativ große

Anzahl verschiedener Typen (44 verschiedene Typen innerhalb der VAMMPIRE Datenbank) und somit die Vergrößerung des Deskriptors auf insgesamt 3.825 verschiedene Atompaare.

3.3.2 Validierung

3.3.2.1 Intrinsische Validierung

Für die intrinsische Validierung des LORD_FP Deskriptors sollte untersucht werden, ob ein Zusammenhang zwischen der LORD_FP Ähnlichkeit zweier identischer Transformationen und dem Effekt auf die Bindungsaffinität hergestellt werden kann, wenn die Transformationen für unterschiedliche Targets beobachtet wurden. Dabei soll ausgeschlossen werden, dass es sich um das gleiche Target oder Isoformen des gleichen Targets handelt. Ein Targetpaar, welches diese Eigenschaften aufweist wird im Folgenden als „gültiges Targetpaar“ bezeichnet. Die ChEMBLdb Target_ID ist für die Validierung ungeeignet, da Targets existieren, die durch unterschiedliche ChEMBLdb_IDs repräsentiert werden, aber für die hier dargestellte Anwendung als identisch angesehen werden. Das Targetpaar CHEMBL1907605/CHEMBL303470 z. B. stellt in beiden Fällen die Cyclin-abhängige Kinase 2 (CDK2) dar. In der entsprechenden Target-Beschreibung der ChEMBLdb unterscheiden sie sich lediglich in den Cyclinen, die sie komplexieren (Cyclin-A1 bzw. Cyclin-E1). Die Bindetasche, in die der hier relevante CDK2 Inhibitor bindet, ist jedoch in beiden Kristallstrukturen identisch. Um sicher zu gehen, dass bei der statistischen Analyse tatsächlich die targetübergreifende Vorhersagekraft des LORD_FP dargestellt wird, wurde ein Vergleich lediglich paarweise für unterschiedliche Targets durchgeführt. Die EC-Nummer (engl. enzyme commission number), welche eine Klassifikation für Enzyme darstellt, wurde deshalb im Rahmen der Validierung als Target-Identifikation angenommen. Für Targets die keine EC-Nummer besitzen wurde anhand des Proteinnamens, welcher in der PDB hinterlegt ist, ausgeschlossen, dass es sich um identische Targets handelt.

Zu Beginn sollte die Verteilung der LORD_FP Ähnlichkeiten innerhalb der gültigen Targetpaare untersucht werden. Dazu wurde für alle Targetpaare, die eine Transformation teilen, die Tanimoto-Ähnlichkeit des LORD_FP berechnet (siehe Abschnitt 2.3.1). In mehr als 90 % der Fälle wurde eine LORD_FP Ähnlichkeit kleiner als 0,5 beobachtet (Abbildung 26 A). Ähnlichkeiten über 0,8 oder sogar Ähnlichkeiten von 1,0, welche vor der Einschränkung der gültigen Targetpaare durch die EC-Nummer noch im Datensatz vorhanden waren, treten in der

dargestellten Ähnlichkeitsverteilung nicht auf. Im Anschluss wurde die Übereinstimmung der LORD_FP Ähnlichkeit mit den beobachteten Transformationseffekten auf die Bindungsaffinität untersucht. (Abbildung 26 B).

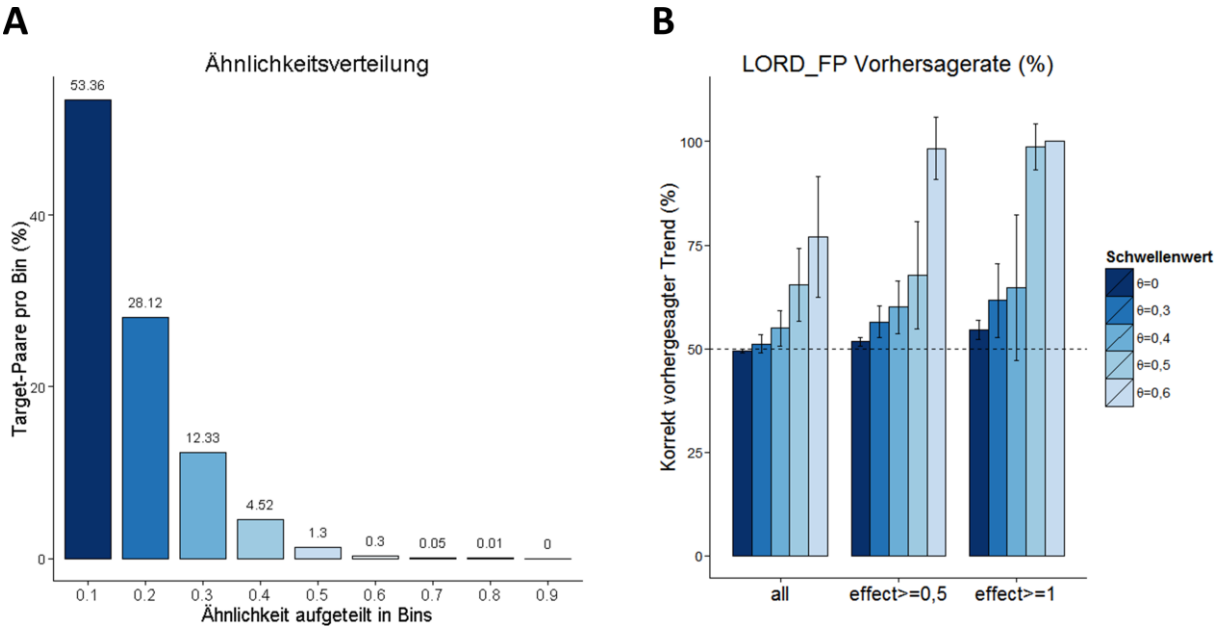


Abbildung 26. A) Ähnlichkeitsverteilung der LORD_FP-Deskriptoren innerhalb der VAMMPIRE Datenbank für alle gültigen Targetpaare. B) Die Vorhersagerate des LORD_FP in Anhängigkeit vom Teildatensatz (alle Paare, alle Paare mit einem Effekt von mindestens 0,5 Log-Einheiten, alle Paare mit einem Effekt von mindestens einer Log-Einheit) und vom Schwellenwert, der die minimale Ähnlichkeit der Targetpaare angibt.

Es wurde keine Korrelation der absoluten Effekte erwartet, da es sich um ähnliche und nicht etwa um identische chemische Umgebungen handelt. Aus diesem Grund wurde lediglich die Übereinstimmung des Trends (in beiden Fällen ein positiver bzw. negativer Effekt auf die Bindungsaffinität) betrachtet. Gleichzeitig wurde untersucht, ob die Stärke des Effekts einen Einfluss auf die Vorhersagekraft des LORD_FP hat.

Es konnte gezeigt werden, dass sich die Übereinstimmung des Trends (Vorhersagerate) mit steigender Ähnlichkeit der chemischen Umgebung (LORD_FP Ähnlichkeit) erhöht. Zudem ist eine Verbesserung der Vorhersagerate zu beobachten, wenn der Effekt in beiden Fällen stärker ist als die erwartete experimentelle Ungenauigkeit von 0,5 Log-Einheiten.[97] Ab einer LORD_FP Ähnlichkeit von 0,5 (θ=0,5) liegt die mittlere Vorhersagerate unabhängig von der Stärke des Effekts bereits bei 65 %. Bei einem Effekt von mindestens 0,5 Log-Einheiten auf beiden Targets und einer LORD_FP Ähnlichkeit von mindestens 0,5, liegt die mittlere Vorhersagerate bei 68 %. Besonders auffällig ist die Verbesserung der Vorhersagerate bei einer LORD_FP Ähnlichkeit von mindestens 0,6. Hier ist allerdings zu berücksichtigen, dass es innerhalb der Datenbank nur

wenige Paare gibt, die eine derart hohe Ähnlichkeit aufweisen. Im Durchschnitt werden fünf Paare mit einer Ähnlichkeit (θ) von mindestens 0,6 beobachtet. Im Vergleich dazu werden durchschnittlich 262 Paare mit θ=0,3, 77 Paare mit θ=0,4 und 21 Paare mit θ=0,5 beobachtet. Nur durchschnittlich zwei Paare mit θ=0,6 weisen auch gleichzeitig einen Effekt ≥ 0,5 Log-Einheiten auf.

Trotz der geringen Anzahl gültiger Targetpaare mit hoher Ähnlichkeit, lässt sich ein Zusammenhang zwischen Ähnlichkeit und Vorhersagerate des LORD_FP beobachten. Die zu Beginn formulierte Hypothese, dass eine Transformation in ähnlicher chemischer Target- und Ligand-Umgebung auch zu einem ähnlichen Effekt auf die Bindungsaffinität führt, wird durch die dargestellten Ergebnisse gestützt. Dennoch ist die Größe des Datensatzes wie bei den meisten MMP-Methoden ein limitierender Faktor und lässt die Frage aufkommen, ob die geringe Anzahl gültiger Targetpaare mit ausreichend hoher Ähnlichkeit repräsentativ ist. Hinzu kommen die Ungenauigkeit der publizierten Messwerte und die Tatsache, dass die verglichenen Werte teilweise aus unterschiedlichen Laboratorien stammen. Die Vorhersagerate des LORD_FP ist außerdem abhängig von der Stärke des Effekts und die Methode deshalb nicht in der Lage kleine Effekte (< 0,5 Log-Einheiten) auf die Bindungsaffinität eines Liganden bzw. den korrekten Trend vorherzusagen. Dies war allerdings auch nicht zu erwarten, da Effekte < 0,5 Log-Einheiten innerhalb der Messungenauigkeit liegen. Um die Vorhersagerate auch für kleinere Effekte zu verbessern, müsste die Bildung von MMPs auf Messwerte, die innerhalb eines Labors entstanden sind, beschränkt werden. Gleichzeitig müsste die Standardabweichung der publizierten Messwerte berücksichtigt werden, welche in den bisherigen Versionen der ChEMBLdb nicht hinterlegt sind.

Mit wachsender Anzahl publizierter Affinitätsdaten zu unterschiedlichen Molekülen und Targets sowie der Strukturaufklärung von Protein-Ligand-Komplexen durch Röntgenkristallographie oder NMR-Spektroskopie kann die Qualität des Modells stetig verbessert werden.

Um den Einfluss der einzelnen Teildeskriptoren auf die Vorhersagekraft zu untersuchen, wurde die Vorhersagerate analog zu **Abbildung 26** jeweils für LLIs, PPIs und PLIs bestimmt (**Abbildung 27**).

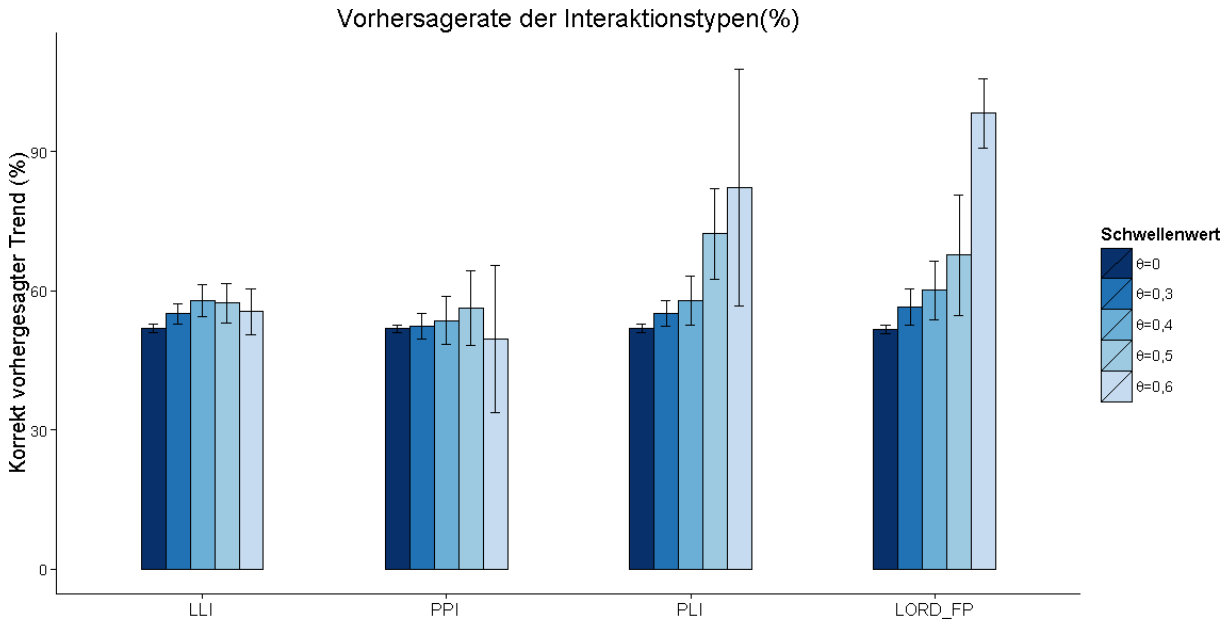


Abbildung 27. Vorhersagerate der einzelnen Interaktionstypen im Vergleich zu LORD_FP. Gezeigt sind die Vorhersageraten der Ligand-Ligand-Interaktionen (LLI), Protein-Protein-Interaktionen (PPIs), Protein-Ligand-Interaktionen (PLI) und des LORD_FP. Die Vorhersagerate des jeweiligen Deskriptors ist in Abhängigkeit der Ähnlichkeit (für alle gültigen Targetpaare) dargestellt.

Es ist deutlich zu erkennen, dass die alleinige Berücksichtigung der LLIs und PPIs keinen Einfluss auf die Vorhersagekraft haben. Die alleinige Berücksichtigung der PLIs hingegen zeigt mit dem LORD_FP vergleichbare Ergebnisse. Die deutlich stärkere Standardabweichung bei hoher Ähnlichkeit wird vermutlich durch jene Fälle ausgelöst, in denen die nähere Ligand-Umgebung die Interaktion mit dem Rezeptor beeinflusst. Ein Beispiel dafür ist in **Abbildung 28** dargestellt. Dabei handelt es sich um zwei Wasserstoff → Fluor Transformationen in ähnlicher hydrophober Rezeptorumgebung, wobei im ersten Fall (A) ein positiver Effekt (Verbesserung um eine Zehnerpotenz) und im zweiten Fall (B) ein negativer Effekt (Verschlechterung um eine Zehnerpotenz) auf die Bindungsaffinität zu beobachten ist. Die Ähnlichkeit des PLI-Teildeskriptors ist mit 0,71 sehr hoch, betrachtet man jedoch die Ähnlichkeit des LLI-Teildeskriptors (0,19), so findet man den Grund für den konträren Effekt auf die Bindungsaffinität. Der aromatische Ring ist hier bereits zweifach mit Fluor substituiert, was zu einer Erhöhung der Elektronegativität und somit zu einer Abnahme der Lipophilie des Fluoratoms führt.

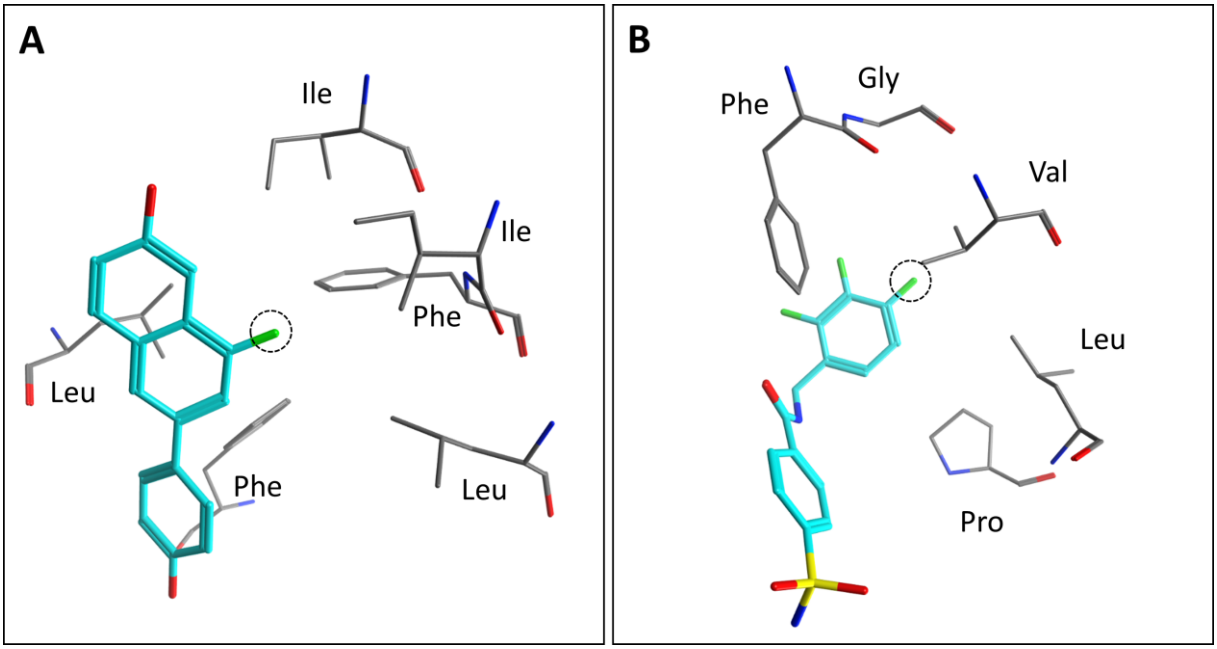


Abbildung 28. Fluorierung in ähnlicher Rezeptorumgebung mit konträrem Effekt auf die Bindungsaffinität. A) Typ-III-MMP (CHEMBL191974/ CHEMBL192386) in der unmittelbaren Rezeptorumgebung (Aminosäuren im Radius von 4.5 Å) des Estrogenrezeptor-β. Die Fluorierung (gestrichelte Linie) hat einen positiven Effekt auf die Bindungsaffinität. B) Typ-II-MMP (1G52/CHEMBL66907) in der unmittelbaren Rezeptorumgebung (Aminosäuren im Radius von 4.5 Å) der Carboanhydrase II. Die Fluorierung (gestrichelte Linie) hat einen negativen Effekt auf die Bindungsaffinität.

Zusammenfassend lässt sich sagen, dass der PLI-Teildeskriptor der Wichtigste für die Vorhersage des Transformationseffekts in einer spezifischen Umgebung ist. Die Hinzunahme des PPI- und LLI-Teildeskriptors ist jedoch nützlich, wenn in speziellen Fällen die lokale Umgebung des Proteins bzw. Liganden die Interaktionen eines Substituenten beeinflussen.

3.3.2.2 Retrospektive Validierung

Anhand einer publizierten Struktur-Aktivitäts-Beziehung (SAR) der Cyclooxygenase-2 (COX-2), sollte der LORD_FP ebenfalls validiert werden. Dazu wurden die von Penning et al.[93] veröffentlichten Celecoxib-Derivate (**Tabelle 4**) zunächst in die Bindetasche der COX-2 gedockt (siehe Abschnitt 2.3.2). Die Ergebnisse des Dockings sind im Anhang (**Abbildung A 1**) dargestellt. Anschließend wurde für jeden Substituenten der LORD_FP berechnet und die VAMMPIRE Datenbank nach Transformationen in ähnlicher Umgebung durchsucht. 83 Transformationen mit einem Effekt von mindestens 0,5 wurden identifiziert, von denen 25 eine LORD_FP Ähnlichkeit von mindestens 0,5 aufwiesen. Als Ergebnis wurde ein Transformationsnetzwerk erhalten, welches in **Abbildung 29** dargestellt ist.

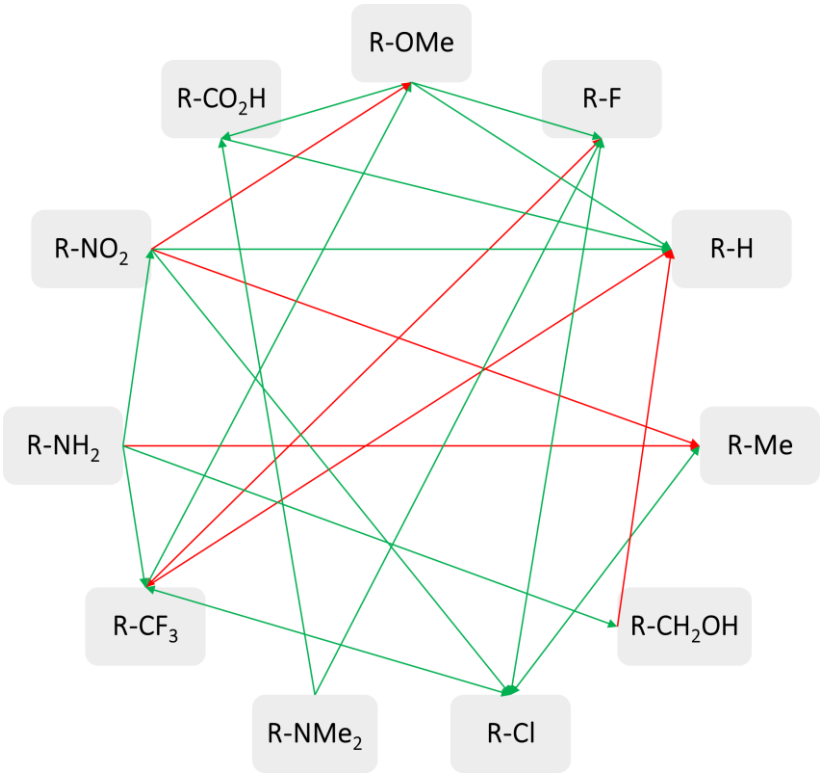


Abbildung 29. Transformationsnetzwerk welches die Ergebnisse der LORD_FP Vorhersage repräsentiert. Die Substituenten der COX-2 SAR sind als Knoten dargestellt, während Transformationen (mit einem Effekt ≥ 0,5) durch gerichtete Kanten dargestellt werden. Bei korrekter Vorhersage des Trends sind die Kanten grün, bei falscher Vorhersage rot dargestellt.

18 der 25 in der VAMMPIRE Datenbank identifizierten Transformationen zeigen einen übereinstimmenden Trend mit den Transformationen der COX-2 SAR. Dies entspricht der erwarteten Vorhersagerate von 70 %, welche durch die intrinsische Validierung ermittelt wurde.

Die chemische Umgebung, welche die höchste Ähnlichkeit aufweist, wurde für die Transformation $-CF_3$ → -Cl gefunden, welche sowohl in der COX-2 SAR als auch bei der Transformation in der VAMMPIRE Datenbank mit einem positiven Effekt auf die Bindungsaffinität hinterlegt ist. Die Transformation wurde für das Protein EPAS1 (engl. Endothelial PAS domain-containing protein 1) gefunden, wobei die LORD_FP Ähnlichkeit 0,63 beträgt. Die Transformation ist das Ergebnis eines Typ-III-MMPs (CHEMBL2311960 → CHEMBL2311959), welches in die Kristallstruktur mit dem PDBcode 4GHI platziert wurde. Der ursprünglich ko-kristallisierte Ligand ist N-(3-chloro-5-fluorophenyl)-4-nitro-2,1,3-benzoxadiazol-5-amin. Die überwiegend hydrophobe und aromatische Umgebung der Transformationen in beiden Proteinen sowie die aromatische Ligand-Umgebung sind in **Abbildung 30** dargestellt.

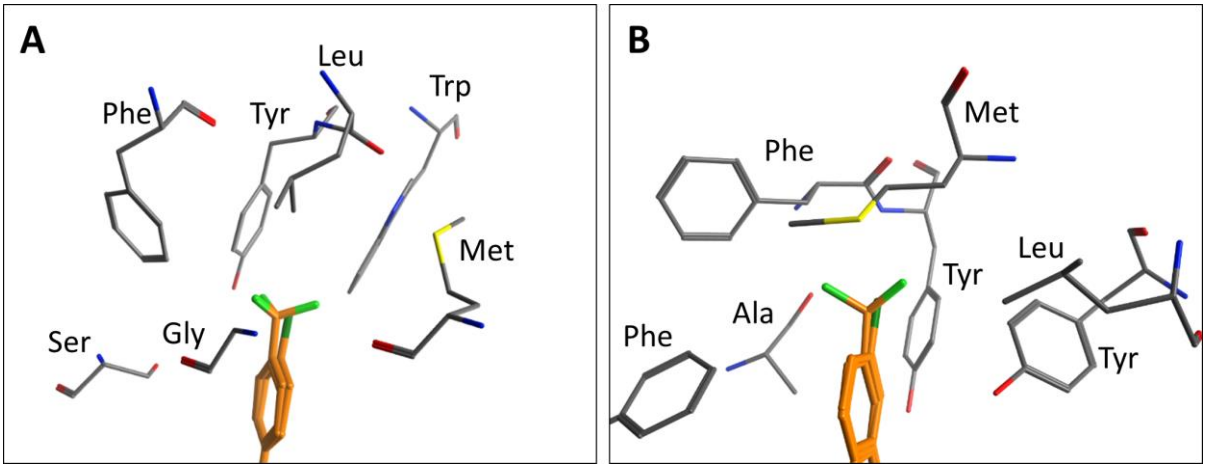


Abbildung 30. Vergleich der chemischen Umgebung der Transformationen -CF_3 → -Cl in der Bindetasche von COX-2 (A) und EPAS1 (B). Alle Aminosäuren innerhalb eines Radius von 4,5 Å um die -CF_3 –Gruppe sind dargestellt. Ein Ausschnitt der Liganden ist in orange gezeigt.

In beiden Fällen ist die Transformation von Methionin, Phenylalanin und Tyrosin umgeben und obwohl die räumliche Anordnung unterschiedlich ist, ist die LORD_FP Ähnlichkeit hoch. Unter der Berücksichtigung, dass die Sequenzidentität der beiden Proteine bei lediglich 4,3 % liegt, ist dies ein sehr gutes Beispiel für die targetübergreifende Vorhersagekraft des LORD_FP

3.3.2.3 Vergleich mit gängigen Bewertungsfunktionen

Eine gängige Methode um Protein-Ligand-Interaktion und deren Beiträge zur Bindungsaffinität eines Liganden zu modellieren ist molekulares Docking und der Einsatz von sogenannten „Bewertungsfunktionen“ (siehe Abschnitt 1.1 und Abschnitt 3.2.2). Im Rahmen einer Masterarbeit, welche parallel zu dieser Arbeit angefertigt wurde, sollte die Vorhersagekraft unterschiedlicher Bewertungsfunktionen für MMPs im Kontext der Rezeptorumgebung untersucht werden.[111] Dazu wurden 108 MMPs auf Basis eines diversen Datensatzes von ko-kristallisierten Protein-Ligand-Komplexen gebildet und die Vorhersagerate der Bewertungsfunktionen, analog zur intrinsischen Validierung des LORD_FP (Abschnitt 3.3.2.1), in Bezug auf die Anzahl korrekt vorhergesagter Trends untersucht. Elf Bewertungsfunktionen aus den Docking Programmen MOE 2014.09 (London dG, ASE, Affinity dG, Alpha HB, GBVI/WSA dG)[83], GOLD v5.2 (GoldScore, ChemScore, ASP und ChemPLP)[31,112,113], AutoDock 4.2[114] und AutoDock Vina 1.1.2[115] sowie die Bewertungsfunktion X-Score[116] und DSX[117] wurden für die Bewertung der Protein-Ligand-Komplexe eingesetzt. Die Komplexe wurden zum einen ohne die Berücksichtigung von Wasser-Molekülen und zum anderen unter der Berücksichtigung von Wasser-Molekülen bewertet, sofern es die Parametrisierung der jeweiligen Funktion zuließ.

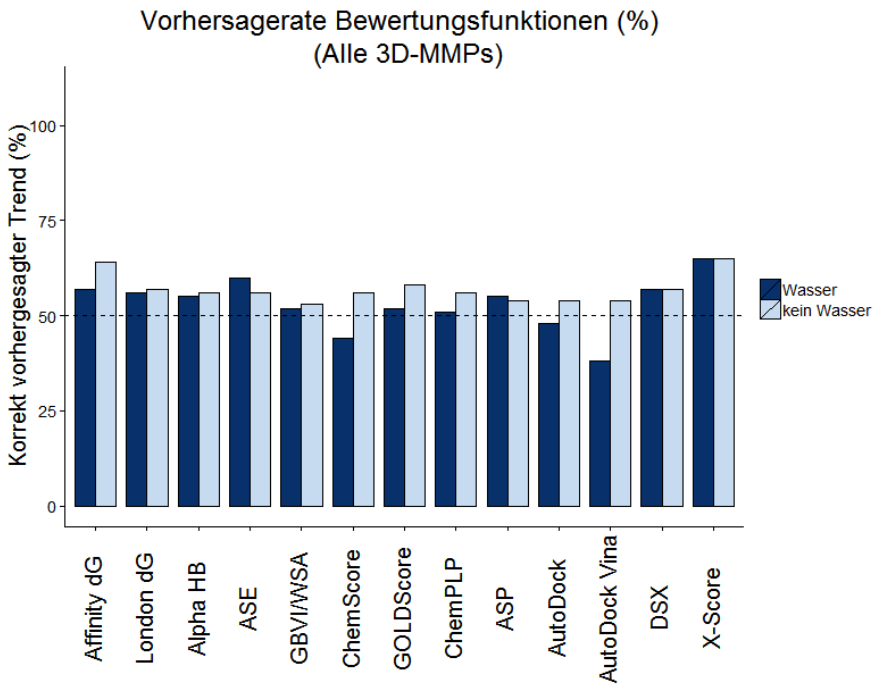


Abbildung 31. Vorhersagerate der 13 Bewertungsfunktionen (Alle 3D-MMPs). Dargestellt ist die Anzahl korrekt vorhergesagter Trends (%) der unterschiedlichen Bewertungsfunktionen, durchgeführt auf einem diversen Datensatz von ko-kristallisierten MMPs. Die Studie wurde unter Berücksichtigung von Wasser-Molekülen und ohne die Berücksichtigung von Wasser-Molekülen durchgeführt (Aus einer Thesis von Lena Kalinowski)[111]

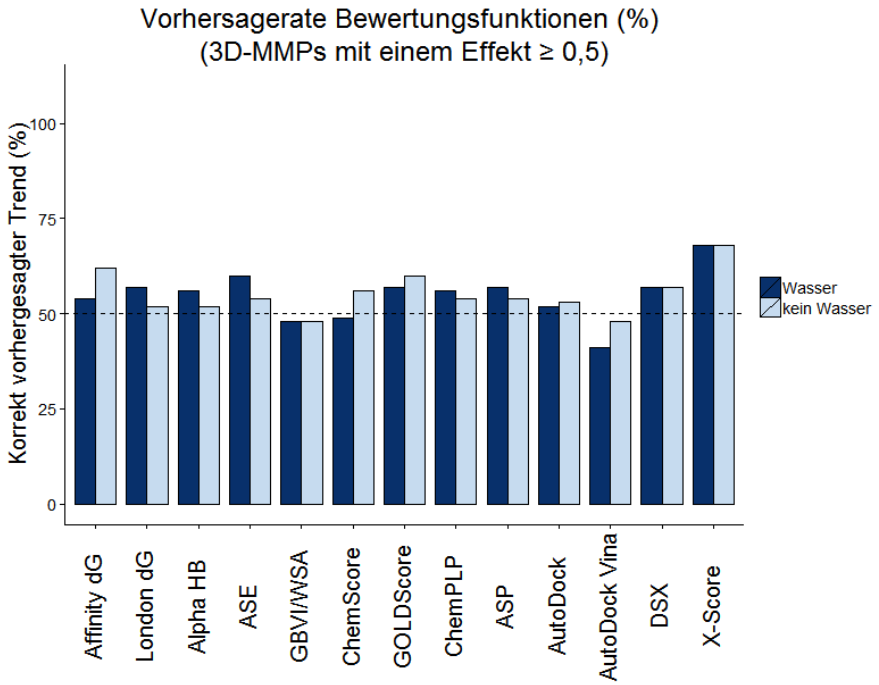


Abbildung 32. Vorhersagerate der 13 Bewertungsfunktionen (3D-MMPs mit einem Effekt ≥ 0,5). Dargestellt ist die Anzahl korrekt vorhergesagter Trends (%) der unterschiedlichen Bewertungsfunktionen, durchgeführt auf einem diversen Datensatz von ko-kristallisierten MMPs mit einem Transformationseffekt von mindestens 0,5 Log-Einheiten. Die Studie wurde unter Berücksichtigung von Wasser-Molekülen und ohne die Berücksichtigung von Wasser-Molekülen durchgeführt (Aus einer Thesis von Lena Kalinowski)[111].

Abbildung 31 zeigt die Vorhersagerate der unterschiedlichen Bewertungsfunktionen bezogen auf alle 3D-MMP-Paare und **Abbildung 32** zeigt die Vorhersagerate für jene 3D-MMP-Paare mit einem Effekt von mindestens 0,5 Log-Einheiten. Lediglich vier der dreizehn Bewertungsfunktionen erreichten eine Vorhersagerate von mindestens 60 % bei einem Transformationseffekt von mindestens 0,5 Log-Einheiten (X-Score, Affinity dG, ASE und GOLDScore). X-Score zeigte mit einer Vorhersagerate 68% das beste Ergebnis.

Diese Erkenntnisse unterstützen die Aussage einer aktuellen Studie von Li et al.,[37] dass die meisten Bewertungsfunktionen zwar in 60 - 80 % der Fälle in der Lage sind die bioaktive Konformation eines Liganden unter den Top-3 der bewerteten Konformationen zu identifizieren, aber nur wenige Bewertungsfunktionen eine Korrelation von Aktivitätsdaten und berechnetem „Score“ (engl. für Auswertung) erreichen (Korrelationskoeffizienten zwischen 0,22 und maximal 0,61 wurden erreicht).

Ein direkter Vergleich zwischen der Vorhersagerate der untersuchten Bewertungsfunktionen und der Vorhersagerate des LORD_FP ist nicht möglich. Zwar liefert der LORD_FP bei einer Tanimoto-Ähnlichkeit von mindestens 0,5 und einem Transformationseffekt von mindestens 0,5 Log-Einheiten eine Vorhersagerate von 70 %, allerdings ist der Umfang der VAMMPIRE Datenbank ein limitierender Faktor. Nicht für jede Anfrage kann ein Eintrag mit ausreichend hoher LORD_FP-Ähnlichkeit identifiziert werden. Die meisten Bewertungsfunktionen hingegen liefern in der Regel immer ein Ergebnis. Des Weiteren sind Bewertungsfunktionen für einzelne Targets oder Targetfamilien oft sehr gut parametrisiert und können hier sehr gute Vorhersageraten liefern. Im allgemeinen basieren Bewertungsfunktionen auf der Annahme, dass die Bindungsaffinität eines Liganden als eine Summe von unabhängigen Termen beschrieben werden kann.[118] Das bedeutet, dass die meisten Bewertungsfunktionen mit der Größe der Moleküle korrelieren. Sieben der in dieser Studie untersuchten Bewertungsfunktionen zeigten bei steigender Molekülgröße in über 65 % der Fälle auch einen höheren Score (X-Score, DSX, London dG, Alpha HB, ASE, ASP). Diese Additivität wird zwar durch den VAMMPIRE-LORD Ansatz umgangen, jedoch können hier lediglich die Beiträge einzelner Transformationen und nicht etwa die Gesamtaffinität eines Moleküls vorhergesagt werden. Aus diesem Grund stellt VAMMPIRE-LORD eine schnelle und einfache Methode zur gezielten Leitstrukturoptimierung dar, die als orthogonale Methode zu molekularem Docking verwendet werden kann.

3.3.3 Webserver

Für die gezielte Vorhersage von vielversprechenden molekularen Transformationen zur Verbesserung der Aktivität einer Leitstruktur zu einem spezifischen Target, wurde der VAMMPIRE-LORD Web-Assistent implementiert. Dabei handelt es sich um eine Webschnittstelle, die den Nutzer Schritt-für-Schritt durch den Prozess der Leitstrukturoptimierung leitet. Voraussetzung hierbei ist die kristallisierte oder modellierte Konformation der Leitstruktur mit dem betrachteten Target. Im Folgenden werden die einzelnen Schritte des VAMMPIRE-LORD Web-Assistenten dargestellt:

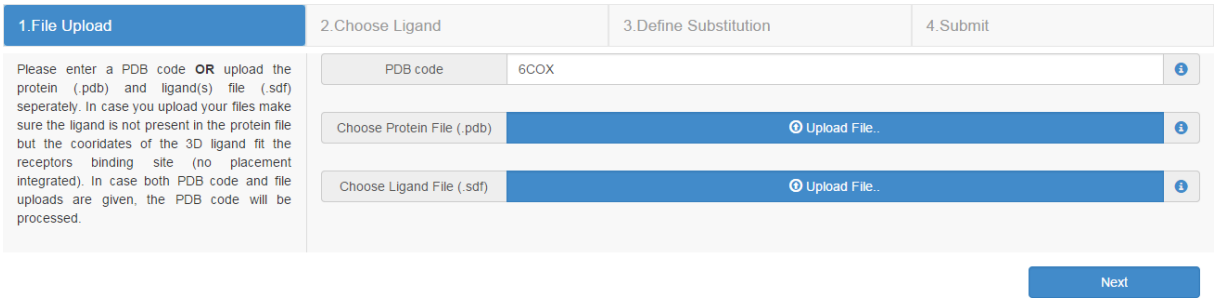


Abbildung 33. Schritt 1: Auswahl von Protein und Ligand.

Im ersten Schritt wird der Nutzer aufgefordert das Target im *Protein Data Bank Format* (*.pdb) und den/die Liganden im *Structure Data Format* (*.sdf) hochzuladen (**Abbildung 33**). Alternativ kann der PDBcode der ko-kristallisierten Struktur eingegeben werden, sofern diese in der PDB hinterlegt ist.

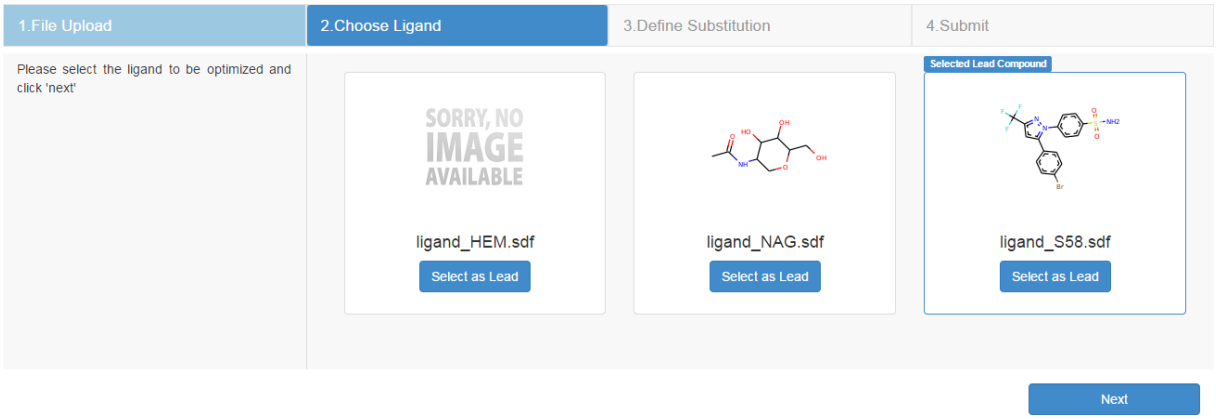


Abbildung 34. Schritt 2: Auswahl der Leitstruktur.

Anschließend werden alle hochgeladenen Liganden aufgelistet (**Abbildung 34**). Für den Fall, dass die Struktur als PDBcode angegeben wurde, wird diese direkt aus der PDB

heruntergeladen und temporär gespeichert. Die Liganden werden auch hier extrahiert und aufgelistet. Der gewünschte Ligand kann aus der Liste ausgewählt werden.

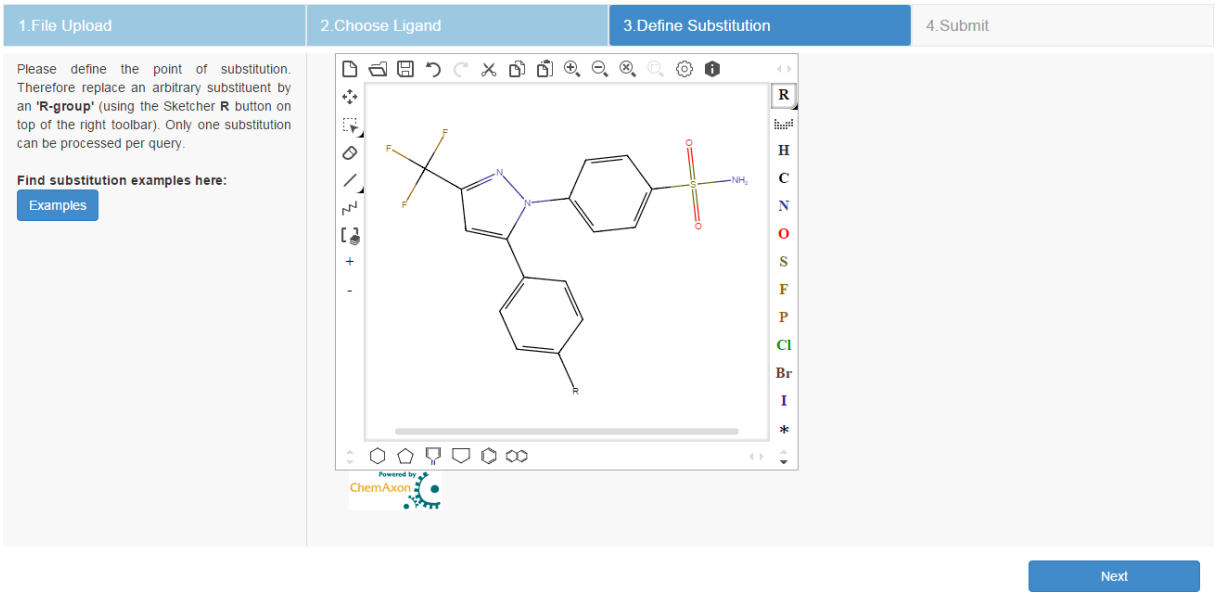


Abbildung 35. Schritt 3: Definition der Substitution.

Nun kann die gewünschte Position der Transformation definiert werden (**Abbildung 35**). Hierzu wird der unerwünschte Substituent entfernt und durch einen Rest („R“) ersetzt. Möchte man zum Beispiel wie im dargestellten Beispiel eine geeignete Substitution für Brom finden, so ersetzt man dieses mit Hilfe des eingebetteten Molekül-Zeichenprogramms durch ein („R“). Pro Anfrage kann genau eine Substitution definiert werden.

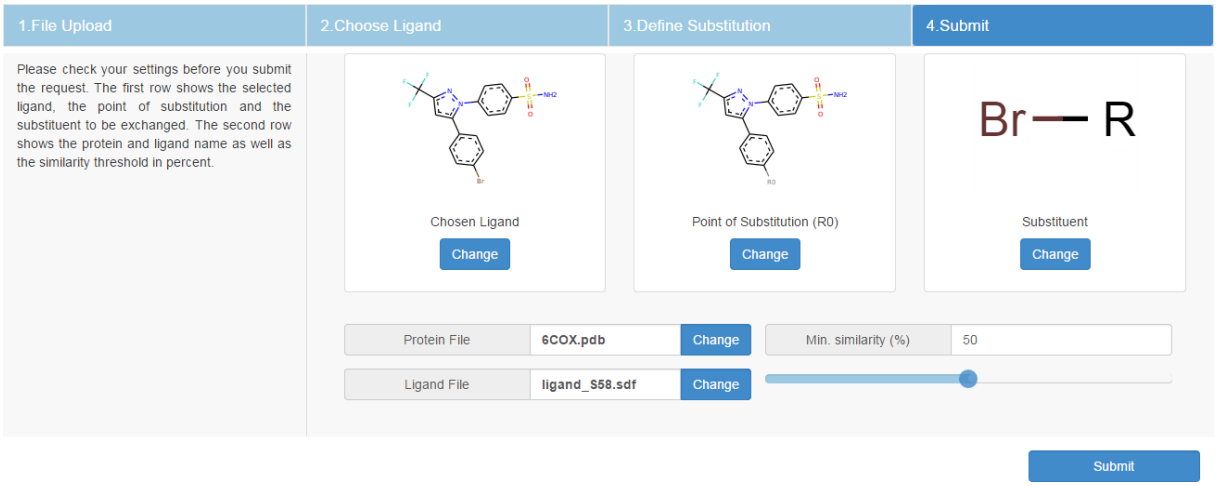
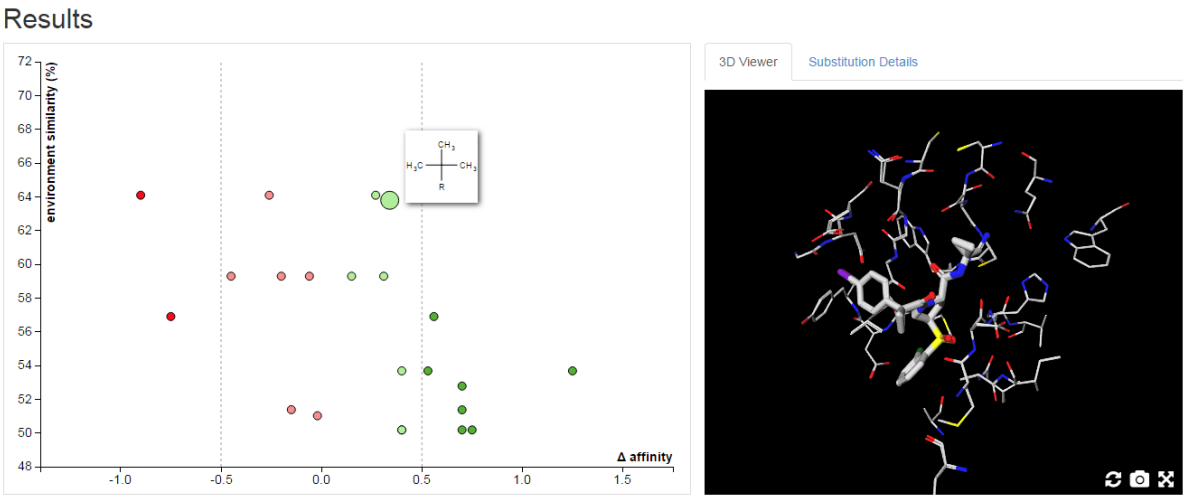


Abbildung 36. Schritt 4: Zusammenfassung der ausgewählten Einstellungen.

Abschließend wird eine Zusammenfassung der ausgewählten Einstellungen angezeigt. (**Abbildung 36**). Der Nutzer kann hier seine Eingaben überprüfen und gegebenenfalls Änderungen vornehmen. Die ausgewählte Position der Substitution wird durch „R0“ dargestellt

und der resultierende Substituent wird separat angezeigt. Des Weiteren wird die minimale Ähnlichkeit (in Prozent) definiert, die zwischen dem LORD_FP des Eingabekomplexes und dem LORD_FP der Datenbank gegeben sein soll. Wie bereits in Kapitel 2.3.2 beschrieben, reicht bereits eine Tanimoto-Ähnlichkeit von 0,5 (50 %) aus, um in durchschnittlich 68 % der Fälle einen korrekten Trend vorherzusagen.

Nach dem Absenden der Anfrage an den Server wird zunächst die chemische Umgebung (siehe Abschnitt 2.2.5) und der LORD_FP (siehe Abschnitt 2.3.1) des Eingabekomplexes berechnet. Anschließend wird die VAMMPIRE Datenbank nach Transformationen durchsucht, die zum einen den ausgewählten Substituenten (im dargestellten Beispiel Brom) enthalten und zum anderen eine LORD_FP Ähnlichkeit von mindestens dem angegebenen Schwellenwert (im dargestellten Beispiel 50 %) aufweisen.



Substitution Details ☑ IC50 ☑ Kd ☑ Ki

ID 1	ID 2	Subst 1	Subst 2	PDBcode	Target ID	Pubmed ID 1	Pubmed ID 2	Effect	Type	Similarity (%)
3r17	CHEMBL1662	[*]F	[*]Cl	1zdp	CHEMBL3392	9888838	19171483	>2	IC50||Ki||Kd	>30
2YJ2	2YJ9	[*]Br	[*]C(F)(F)F	2YJ2_2YJ9	CHEMBL2111380	21898833	21898833	-0.90	IC50	64.09
2YJ2	2YJ8	[*]Br	[*]I	2YJ2_2YJ8	CHEMBL2111380	21898833	21898833	0.27	IC50	64.09
2YJ2	2YJC	[*]Br	[*]Cl	2YJC_2YJ2	CHEMBL2111380	21898833	21898833	-0.26	IC50	64.09
CHEMBL2333443	4I4F	[*]Br	[*]C(C)(C)C	4I4F	CHEMBL2695	23414845	23414845	0.34	IC50	63.79
CHEMBL123404	CHEMBL121729	[*]Br	[*][N+](=O)[O-]	3T5U	CHEMBL205	11020282	11020282	0.31	Ki	59.29
CHEMBL123404	CHEMBL122868	[*]Br	[*]OC	3T5U	CHEMBL205	11020282	11020282	-0.45	Ki	59.29
CHEMBL123404	CHEMBL123233	[*]Br	[*]NC(C)=O	3T5U	CHEMBL205	11020282	11020282	-0.20	Ki	59.29

Abbildung 37. VAMMPIRE-LORD Auflistung der Ergebnisse.

Die identifizierten Substitutionen, welche die in der Anfrage definierten Anforderungen erfüllen, werden auf der Ergebnisseite präsentiert (**Abbildung 37**). Eine Übersicht der Ergebnisse ist in Form einer Grafik dargestellt, in dem grüne und rote Punkte Substitutionen

mit positiven bzw. negativen Effekten auf die Bindungsaffinität des Moleküls darstellen. Dabei wird zwischen Effekten kleiner als 0,5 Log-Einheiten (hellgrün bzw. hellrot) und mindestens 0,5 Log-Einheiten (dunkelgrün bzw. dunkelrot) unterschieden. Die Position der Punkte ist abhängig von der Ähnlichkeit zum LORD_FP der Anfrage und der Stärke des Effekts (Δ Affinität). Transformationen mit hoher Ähnlichkeit zum LORD_FP der Anfrage, die gleichzeitig einen starken positiven Effekt auf die Bindungsaffinität des Moleküls haben, sind in der rechten oberen Ecke des Graphen lokalisiert. Beim Auswählen eines Punktes wird die Transformation sowie die Ligand- und Rezeptorumgebung in einer interaktiven 3D-Visualisierung angezeigt.

Der implementierte Webserver soll bei der Optimierung einer gegebenen Leitstruktur assistieren. Es handelt sich dabei jedoch nicht um ein Klassifizierungsmodell, welches spezifische Transformationen vorhersagt, sondern vielmehr um die Aufbereitung einer Übersicht von bereits in ähnlichen chemischen Umgebungen durchgeführten Transformationen und deren Effekte auf die Bindungsaffinität eines Liganden. Die 3D-Visualisierung und die zu den Transformationen hinterlegten Informationen können dem Medizinalchemiker bei der Entscheidung assistieren, welcher nächste Schritt im Zuge der Leitstrukturoptimierung den erwünschten Effekt auf die Bindungsaffinität mit sich bringen könnte.

3.4 Fazit

Das Ziel dieser Arbeit war die Erstellung eines targetübergreifenden Modells für die gezielte Leitstrukturoptimierung auf der Basis von experimentell bestimmten Affinitätsdaten sowie der strukturellen Information diverser Protein-Ligand-Komplexe. Anhand einer umfangreichen Datenbank von MMPs im Kontext ihrer Rezeptorumgebung konnte gezeigt werden, dass der Effekt einer Transformation auf die Bindungsaffinität eines Liganden sowohl von der der chemischen Umgebung als auch von der Stärke eines Transformationseffekts abhängig ist. Mit Hilfe eines topologischen Atompaar-Deskriptors (LORD_FP) konnte eine mathematische Repräsentation der Rezeptor- und Ligand-Umgebung einer Transformation entwickelt und so gezielt targetübergreifende Vorhersagen von Transformationseffekten realisiert werden.

Anhand einer intrinsischen Validierung konnte gezeigt werden, dass Transformationen in ähnlichen chemischen Umgebungen ähnliche Effekte auf die Bindungsaffinität eines Liganden

aufweisen, auch wenn es sich dabei um unterschiedliche Targets handelt. Bei einer LORD_FP-Ähnlichkeit von mindestens 0,5 (Tanimoto-Koeffizient) und einem Transformationseffekt von mindestens 0,5 Log-Einheiten konnte außerdem am Beispiel einer COX-2-SAR eine Vorhersagerate von 72 % erreicht werden.

Wie bei den meisten MMP-Methoden ist auch bei dieser Methode die Anzahl der MMPs bzw. die Anzahl der ko-kristallisierten Protein-Ligand-Komplexe ein limitierender Faktor. Die Vorhersage für einen Transformationseffekt kann nur dann erfolgen, wenn die entsprechende Transformation in einer Umgebung mit ausreichend hoher Ähnlichkeit (bezüglich des LORD_FP) in der VAMMPIRE Datenbank hinterlegt ist. Unter der Annahme, dass ähnliche Moleküle, die an einem spezifischen Target binden, eine ähnliche bioaktive Konformation aufweisen, wurde die gemeinsame Teilstruktur der Moleküle innerhalb eines MMPs genutzt, um die Konformation eines Liganden mit unbekannten Koordinaten vorherzusagen. So konnte zwar eine umfangreiche 3D-MMP-Datenbank von 17.602 3D-MMPs erzeugt werden, dennoch decken die dadurch repräsentierten Transformationen und Target-Umgebungen nicht den möglichen chemischen Raum ab.

Auch die Vertrauenswürdigkeit der in der ChEMBLdb hinterlegten experimentellen Daten und die Vergleichbarkeit der in unterschiedlichen Laboratorien gemessenen Affinitätswerte stellen eine bislang unvermeidbare Fehlerquelle dar. Die intrinsische Analyse der ChEMBLdb zeigte, dass die Pearson-Korrelation der in unterschiedlichen Laboratorien bestimmten Affinitätswerte zwischen 0,71 (für IC_{50}-Werte) und 0,77 (für K_i-Werte) liegt. Es ist festzuhalten, dass ein auf dieser Datengrundlage erstelltes mathematisches Modell im Mittel nicht besser sein kann, als diese intrinsische Korrelation.

Dennoch stellt die in dieser Arbeit entwickelte Methode einen neuartigen und vielversprechenden Ansatz zur Vorhersage von bevorzugten Protein-Ligand-Interaktionen dar. Bei ausreichend hoher Ähnlichkeit der chemischen Umgebung einer Transformation und bei einem starken Transformationseffekt können gute Vorhersageraten erreicht werden, die mit gängigen Bewertungsfunktionen mithalten oder sie sogar übertreffen können. Insbesondere im Hinblick auf die stetig wachsende Anzahl an Protein-Ligand-Komplexen in der PDB sowie experimentell bestimmter Affinitätsdaten in der ChEMBLdb könnten 3D-MMP-Ansätze in Zukunft von großer Bedeutung sein.

Zusammenfassung

Die Anzahl experimentell bestimmter Aktivitätsdaten zu einer Vielzahl von Molekülen und krankheitsrelevanten Targets steigt Jahr für Jahr. Die ChEMBL Datenbank (ChEMBLdb),[76] eine Sammlung von bioaktiven, wirkstoffartigen Molekülen ist allein in den letzten fünf Jahren um mehr als das doppelte gewachsen und enthält in der aktuellen Version *ChEMBL_20* (Stand: März 2015) 1.463.270 Verbindungen mit 13.520.737 experimentell bestimmten Aktivitätswerten zu 10.774 Targets, extrahiert aus 59.610 unterschiedlichen Publikationen. Gleichzeitig steigt auch die Anzahl der überwiegend durch Röntgenkristallographie und Kernspinresonanzspektroskopie (NMR-Spektroskopie) aufgeklärten 3D-Strukturen von Protein-Ligand-Komplexen. Die Protein Data Bank (PDB)[9], eine Sammlung von biologischen makromolekularen Strukturen, enthält zum aktuellen Zeitpunkt (Stand: März 2015) 106.858 Einträge und zeigte in den vergangenen 40 Jahren einen exponentiellen Anstieg. Diese Sammlungen von experimentellen Daten sind für die Entwicklung von mathematischen Modellen zur Vorhersage von Protein-Ligand-Interaktionen und Bindungsaffinitäten von großer Bedeutung.

In der hier vorgestellten Arbeit wurden die oben genannten Datenquellen genutzt um einen neuartigen, strukturbasierten Ansatz zur gezielten Leitstrukturoptimierung zu entwickeln. Die Grundlage dafür bilden die sogenannten Matched Molecular Pairs (MMPs). Dabei handelt es sich um Paare von Molekülen, welche sich lediglich in einer wohldefinierten Modifikation (Transformation) unterscheiden und sich in einer Datenbank mit gemessenen Moleküleigenschaften befinden.[61] Diese Transformationen können in Verbindung mit der Änderung einer Moleküleigenschaft (Transformationseffekt) gebracht und statistisch analysiert werden. Der Unterschied zu den bisher bekannten MMP-Methoden[59–75] ist, dass die MMPs im Kontext der unmittelbaren Rezeptor- und Ligand-Umgebung untersucht werden und daraus ein targetübergreifendes Modell zur Vorhersage des Effekts auf die Bindungsaffinität eines Liganden abgeleitet werden kann. Das Modell beruht auf der Annahme, dass eine molekulare

Transformation in ähnlichen Rezeptor- und Ligand-Umgebungen zu ähnlichen Effekten auf die Bindungsaffinität eines Liganden führt, unabhängig davon um welchen Rezeptor oder Liganden es sich handelt.

Um ein solches Modell zu generieren wurde zunächst eine umfangreiche Datenbank von MMPs mit bioaktiven Konformationen (3D-MMPs) benötigt. Da die Anzahl der Moleküle mit gemessenen Aktivitätsdaten die Anzahl der ko-kristallisierten Protein-Ligand-Komplexe weit übersteigt, wurde zunächst eine Methode zur Modellierung von 3D-MMPs entwickelt. Unter der Annahme, dass ähnliche Moleküle, die an einem spezifischen Target binden, eine ähnliche bioaktive Konformation aufweisen, wurde die gemeinsame Teilstruktur der Moleküle innerhalb eines MMPs genutzt um die Konformation des Liganden mit unbekannten Koordinaten vorherzusagen. Dazu wurde zunächst auf Basis der gemeinsamen Teilstruktur der Moleküle ein Pharmakophormodell generiert, welches dem Docking-Platzierungsalgorithmus anschließend als Orientierung diente. Mit Hilfe dieser Methode wurde die 3D-MMP-Datenbank von ursprünglich 938 auf 17.602 MMPs erweitert.

Um eine mathematische Beziehung zwischen einer molekularen Transformation, in einer spezifischen chemischen Umgebung, und dessen Effekt auf die Bindungsaffinität eines Liganden herzustellen, wurde ein topologischer Atompaar-Deskriptor entwickelt, der indirekt die Interaktionen zwischen Protein und Ligand (bzw. Protein und Substituent) und gleichzeitig die Komposition der unmittelbaren Rezeptor- und Ligand-Atome einer Transformation beschreibt. Auf Basis der 3D-MMP-Datenbank sollte schließlich untersucht werden, ob eine spezifische Transformation in ähnlichen chemischen Umgebungen (ausgedrückt durch die Tanimoto-Ähnlichkeit der Atompaar-Deskriptoren) zu einem ähnlichen Effekt auf die Bindungsaffinität der Liganden führt, auch wenn es sich dabei um unterschiedliche Targets und Liganden handelt. Anhand einer intrinsischen Kreuzvalidierung und am Beispiel einer Struktur-Aktivitätsbeziehung von Cyclooxygenase-2-Inhibitoren konnte gezeigt werden, dass die Vorhersagerate des Modells sowohl von der Ähnlichkeit der Deskriptoren als auch von der Stärke des Transformationseffekts abhängig ist.

Auf Basis der mathematischen Beziehung zwischen Deskriptor-Ähnlichkeit und Transformationseffekt wurde schließlich eine Methode zur gezielten Leitstrukturoptimierung implementiert und als Webservice zur Verfügung gestellt. Die 3D-MMP-Datenbank dient hier

als Wissensbasis für die Vorhersage von vielversprechenden Transformationen zur Optimierung der Bindungsaffinität einer Leitstruktur mit bekannter bioaktiver Konformation.

Literaturverzeichnis

1. Imming, P.; Sinning, C.; Meyer, A. Drugs, Their Targets and the Nature and Number of Drug Targets. *Nat. Rev. Drug Discov.* 2006, *5*, 821–834.

2. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How Many Drug Targets Are There? *Nat. Rev. Drug Discov.* 2006, *5*, 993–996.

3. Rask-Andersen, M.; Almén, M. S.; Schiöth, H. B. Trends in the Exploitation of Novel Drug Targets. *Nat. Rev. Drug Discov.* 2011, *10*, 579–590.

4. Afshar, M.; Prescott, C. D.; Varani, G. Structure-Based and Combinatorial Search for New RNA-Binding Drugs. *Curr. Opin. Biotechnol.* 1999, *10*, 59–63.

5. Gallego, J.; Varani, G. Targeting RNA with Small-Molecule Drugs: Therapeutic Promise and Chemical Challenges. *Acc. Chem. Res.* 2001, *34*, 836–843.

6. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem Biol* 2003, *10*, 787–797.

7. Kroemer, R. T. Structure-Based Drug Design: Docking and Scoring. *Curr. Protein Pept. Sci.* 2007, *8*, 312–328.

8. Gane, P. J.; Dean, P. M. Recent Advances in Structure-Based Rational Drug Design. *Curr. Opin. Struct. Biol.* 2000, *10*, 401–404.

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

10. Hajduk, P. J.; Greer, J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discov.* 2007, *6*, 211–219.

11. Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. Fragment-Based Lead Discovery Using X-Ray Crystallography. *J. Med. Chem.* 2005, *48*, 403–413.

12. Joseph-McCarthy, D.; Campbell, A. J.; Kern, G.; Moustakas, D. Fragment-Based Lead Discovery and Design. *J. Chem. Inf. Model.* 2014, *54*, 693–704.

13. Farmer, B. T.; Reitz, A. B. Fragment-Based Drug Discovery. *Pract. Med. Chem.* 2008, 228–243.

14. Albert, J. S. Fragment-Based Lead Discovery. *Lead Gener. Approaches Drug Discov.* **2010**, 105–139.

15. Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nat. Chem.* **2009**, *1*, 187–192.

16. Yokoyama, S. Protein Expression Systems for Structural Genomics and Proteomics. *Curr. Opin. Chem. Biol.* **2003**, *7*, 39–43.

17. Heinemann, U.; Büssow, K.; Mueller, U.; Umbach, P. Facilities and Methods for the High-Throughput Crystal Structural Analysis of Human Proteins. *Acc. Chem. Res.* **2003**, *36*, 157–163.

18. Goulding, C. W.; Perry, L. J. Protein Production in Escherichia Coli for Structural Studies by X-Ray Crystallography. *J. Struct. Biol.* **2003**, *142*, 133–143.

19. Stewart, L.; Clark, R.; Behnke, C. High-Throughput Crystallization and Structure Determination in Drug Discovery. *Drug Discov. Today* **2002**, *7*, 187–196.

20. Sharff, A.; Jhoti, H. High-Throughput Crystallography to Enhance Drug Discovery. *Curr. Opin. Chem. Biol.* **2003**, *7*, 340–345.

21. Tickle, I.; Sharff, A.; Vinkovic, M.; Yon, J.; Jhoti, H. High-Throughput Protein Crystallography and Drug Discovery. *Chem. Soc. Rev.* **2004**, *33*, 558–565.

22. Williams, S. P.; Kuyper, L. F.; Pearce, K. H. Recent Applications of Protein Crystallography and Structure-Guided Drug Design. *Curr. Opin. Chem. Biol.* **2005**, *9*, 371–380.

23. Gileadi, O.; Burgess-Brown, N. A.; Colebrook, S. M.; Berridge, G.; Savitsky, P.; Smee, C. E. A.; Loppnau, P.; Johansson, C.; Salah, E.; Pantic, N. H. High Throughput Production of Recombinant Human Proteins for Crystallography. *Methods Mol. Biol.* **2008**, *426*, 221–246.

24. Ziegler, S.; Pries, V.; Hedberg, C.; Waldmann, H. Target Identification for Small Bioactive Molecules: Finding the Needle in the Haystack. *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 2744–2792.

25. Rabilloud, T.; Chevallet, M.; Luche, S.; Lelong, C. Two-Dimensional Gel Electrophoresis in Proteomics: Past, Present and Future. *J. Proteomics* **2010**, *73*, 2064–2077.

26. Keserü, G. M.; Makara, G. M. The Influence of Lead Discovery Strategies on the Properties of Drug Candidates. *Nat. Rev. Drug Discov.* **2009**, *8*, 203–212.

27. Keseru, G. M.; Makara, G. M. Hit Discovery and Hit-to-Lead Approaches. *Drug Discov. Today* **2006**, *11*, 741–748.

28. Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of Successful Lead Generation. *Curr. Top. Med. Chem.* **2005**, *5*, 421–439.

29. Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining Docking and Molecular Dynamic Simulations in Drug Design. *Med. Res. Rev.* 2006, *26*, 531–568.

30. Kalyaanamoorthy, S.; Chen, Y.-P. P. Structure-Based Drug Design to Augment Hit Discovery. *Drug Discov. Today* 2011, *16*, 831–839.

31. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* 1997, *267*, 727–748.

32. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* 2004, *47*, 1739–1749.

33. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* 1996, *261*, 470–489.

34. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* 2004, *3*, 935–949.

35. Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking: Current Status and Future Challenges. *Proteins* 2006, *65*, 15–26.

36. Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.* 2014, *54*, 1700–1716.

37. Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* 2014, *54*, 1717–1736.

38. Caffrey, M. Membrane Protein Crystallization. *J. Struct. Biol.* 2003, *142*, 108–132.

39. Caffrey, M.; Cherezov, V. Crystallizing Membrane Proteins Using Lipidic Mesophases. *Nat. Protoc.* 2009, *4*, 706–731.

40. Acharya, C.; Coop, A.; Polli, J. E.; Mackerell, A. D. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr. Comput. Aided. Drug Des.* 2011, *7*, 10–22.

41. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* 2014, *57*, 3186–3204.

42. Auer, J.; Bajorath, J. Molecular Similarity Concepts and Search Calculations. *Methods Mol. Biol.* 2008, *453*, 327–347.

43. Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discov. Today* 2007, *12*, 225–233.

44. Willett, P. Similarity-Based Approaches to Virtual Screening. *Biochem. Soc. Trans.* 2003, *31*, 603–606.

45. Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of Correlation Vector Methods for Ligand-Based Similarity Searching. *J. Comput. Aided. Mol. Des.* 2003, *17*, 687–698.

46. Schneider, G. B. K.-H. *Molecular Design*; WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2008.

47. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* 2014, *57*, 4977–5010.

48. Goodnow, R. A. Hit and Lead Identification: Integrated Technology-Based Approaches. *Drug Discov. Today Technol.* 2006, *3*, 367–375.

49. Topliss, J. G. Utilization of Operational Schemes for Analog Synthesis in Drug Design. *J. Med. Chem.* 1972, *15*, 1006–1011.

50. Patrick, G. L. *An Introduction to Medicinal Chemistry*; 5th edit.; Oxford University Press, 2013.

51. Kar, A. *Medicinal Chemistry*; 3rd edit.; Anshan Ltd., 2006.

52. Friedman, H. Influence of Isosteric Replacements upon Biological Activity. *Symp. Chem. Correl. …* 1951, *206*, 295–358.

53. Ciapetti, P.; Giethlen, B. Molecular Variations Based on Isosteric Replacements. *Pract. Med. Chem.* 2008, 290–342.

54. Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Model.* 2002, *42*, 103–108.

55. Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of Intersubstituent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* 2003, *43*, 406–411.

56. Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* 2006, *46*, 677–685.

57. Krier, M.; Hutter, M. C. Bioisosteric Similarity of Molecules Based on Structural Alignment and Observed Chemical Replacements in Drugs. *J. Chem. Inf. Model.* 2009, *49*, 1280–1297.

58. Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of Reduced Graphs to Encode Bioisosterism for Similarity-Based Virtual Screening. *J. Chem. Inf. Model.* 2009, *49*, 1330–1346.

59. Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. SwissBioisostere: A Database of Molecular Replacements for Ligand Design. *Nucleic Acids Res.* 2013, *41*, 1137–1143.

60. Haubertin, D. Y.; Bruneau, P. A Database of Historically-Observed Chemical Replacements. *J. Chem. Inf. Model.* 2007, *47*, 1294–1302.

61. Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* 2006, *49*, 6672–6682.

62. Griffen, E.; Leach, A.; Robb, G.; Warner, D. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* 2011, *54*, 7739–7750.

63. Geppert, T.; Beck, B. Fuzzy Matched Pairs: A Means to Determine the Pharmacophore Impact on Molecular Interaction. *J. Chem. Inf. Model.* 2014, *54*, 1093–1102.

64. Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discov. Today* 2013, *18*, 724–731.

65. Zhang, L.; Zhu, H.; Mathiowetz, A.; Gao, H. Deep Understanding of Structure-Solubility Relationship for a Diverse Set of Organic Compounds Using Matched Molecular Pairs. *Bioorganic Med. Chem.* 2011, *19*, 5763–5770.

66. O'Boyle, N. M.; Boström, J.; Sayle, R. a; Gill, A. Using Matched Molecular Series as a Predictive Tool to Optimize Biological Activity. *J. Med. Chem.* 2014, *57*, 2704–2713.

67. Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; et al. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of HERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* 2010, *50*, 1872–1886.

68. Posy, S. L.; Claus, B. L.; Pokross, M. E.; Johnson, S. R. 3D Matched Pairs: Integrating Ligand- and Structure-Based Knowledge for Ligand Design and Receptor Annotation. *J. Chem. Inf. Model.* 2013, *53*, 1576–1588.

69. Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* 2010, *50*, 1350–1357.

70. Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* 2012, *52*, 1138–1145.

71. Beck, J. M.; Springer, C. Quantitative Structure-Activity Relationship Models of Chemical Transformations from Matched Pairs Analyses. *J. Chem. Inf. Model.* **2014**, *54*, 1226–1234.

72. Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Bioisosteric Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3*, 425–436.

73. De la Vega de León, A.; Bajorath, J. Prediction of Compound Potency Changes in Matched Molecular Pairs Using Support Vector Regression. *J. Chem. Inf. Model.* **2014**, *54*, 2654–2663.

74. Bradley, A. R.; Wall, I. D.; Green, D. V. S.; Deane, C. M.; Marsden, B. D. OOMMPPAA: A Tool To Aid Directed Synthesis by the Combined Analysis of Activity and Structural Data. *J. Chem. Inf. Model.* **2014**, *54*, 2636–2646.

75. Wassermann, A. M.; Bajorath, J. Chemical Substitutions That Introduce Activity Cliffs across Different Compound Classes and Biological Targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.

76. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *44*, 1–8.

77. Willighagen, E. L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A. J.; Tkachenko, V.; Hastings, J.; Chen, B.; Wild, D. J. The ChEMBL Database as Linked Open Data. *J. Cheminform.* **2013**, *5*.

78. Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K I Data. *J. Med. Chem.* **2012**, *55*, 5165–5173.

79. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. *Stud. Classif. Data Anal. Knowl. Organ. (GfKL 2007)* **2007**, 319–326.

80. Weber, J. Implementierung Einer Wissensbasierten Bewertungsfunktion Für Protein-Ligand Komplexe, Goethe Universität Frankfurt, 2012.

81. Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 1675–1679.

82. Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

83. Chemical Computing Group Inc. Molecular Operating Environment (MOE), 2014.09. *Molecular Operating Environment (MOE), 2014.09*, 2014.

84. Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* 2004, *32*, D115–D119.

85. Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2005, *33*.

86. Landrum, G. RDKit: Open-source cheminformatics, http://www. rdkit.org.

87. Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* 2010, *50*, 339–348.

88. Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small Molecule Subgraph Detector (SMSD) Toolkit. *J. Cheminform.* 2009, *1*, 12–25.

89. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* 2003, *43*, 493–500.

90. Case, D, A. AMBER 12. *Univ. California, San Fr.* 2012.

91. Gerber, P. R.; Müller, K. MAB, a Generally Applicable Molecular Force Field for Structure Modelling in Medicinal Chemistry. *J. Comput. Aided. Mol. Des.* 1995, *9*, 251–268.

92. Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Model.* 1995, *35*, 1039–1045.

93. Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; et al. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1H-Pyrazol-1-Yl]benze Nesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* 1997, *40*, 1347–1365.

94. Ferreira, R. S.; Bryant, C.; Ang, K. K. H.; McKerrow, J. H.; Shoichet, B. K.; Renslo, A. R. Divergent Modes of Enzyme Inhibition in a Homologous Structure-Activity Series. *J. Med. Chem.* 2009, *52*, 5005–5008.

95. Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* 2010, *53*, 37–51.

96. Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.* 2007, *50*, 2385–2390.

97. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PLoS One* 2013, *8*, e61007.

98. Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization. *J. Med. Chem.* 2013, *56*, 5203–5207.

99. Weber, J.; Rupp, M.; Proschak, E. Impact of X-Ray Structure on Predictivity of Scoring Functions: PPARγ Case Study. *Mol. Inform.* 2012, *31*, 631–633.

100. Müller, K.; Faeh, C.; Diederich, F. Fluorine in Pharmaceuticals: Looking beyond Intuition. *Science* 2007, *317*, 1881–1886.

101. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* 1999, *42*, 5100–5109.

102. Houston, D. R.; Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.* 2013, *53*, 384–390.

103. Plewczynski, D.; Łaźniewski, M.; Grotthuss, M. Von; Rychlewski, L.; Ginalski, K. VoteDock: Consensus Docking Method for Prediction of Protein-Ligand Interactions. *J. Comput. Chem.* 2011, *32*, 568–581.

104. Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur–Halogen Bonds. *J. Chem. Theory Comput.* 2011, *7*, 2307–2315.

105. Molecular Interactions, Consulting Cambridge MedChem http://www.cambridgemedchemconsulting.com/resources/molecular_interactions.html.

106. Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional Group Contributions to Drug-Receptor Interactions. *J. Med. Chem.* 1984, *27*, 1648–1657.

107. Carver, F. J.; Hunter, C. A.; Seward, E. M. Structure–activity Relationship for Quantifying Aromatic Interactions†. *Chem. Commun.* 1998, 775–776.

108. Hunter, C. A. Quantifying Intermolecular Interactions: Guidelines for the Molecular Recognition Toolbox. *Angew. Chem. Int. Ed. Engl.* 2004, *43*, 5310–5324.

109. Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chemie Int. Ed.* 2002, *41*, 2644–2676.

110. Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE-LORD: A Webserver for Straightforward Lead Optimization Using Matched Molecular Pairs. *J. Chem. Inf. Model.* 2015, *55*, 207–213.

111. Kalinowsky, L. Validierung von Bewertungsfunktionen Für Protein–Ligand Komplexe Mit Hilfe von Matched Molecular Pairs, Goethe Universität Frankfurt, 2015.

112. Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

113. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein – Ligand Docking Using GOLD. *Proteins Struct. Funct. Bioinforma.* **2003**, *623*, 609–623.

114. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28*, 1145–1152.

115. Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.

116. Renxiao Wang. X-Score http://sw16.im.med.umich.edu/software/xtool.

117. Neudert, G.; Klebe, G. DSX: A Knowledge-Based Scoring Function for the Assessment of Receptor-Ligand Complexes. *J. Chem. Inf. Model.* **2011**, *In Press*, 2731–2745.

118. Schulzgasch, T.; Stahl, M. Scoring Functions for Protein-Ligand Interactions: A Critical Perspective. *Drug Discov. Today Technol.* **2004**, *1*, 231–239.

Abbildungsverzeichnis

Eidesstattliche Erklärung

Die vorliegende Dissertation wurde selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Alle Stellen die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche gekennzeichnet. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung nicht vorgelegt worden.

Frankfurt am Main, den …… …….

(Julia Weber)

Publikationsliste

1. Julia Weber; Matthias Rupp; Ewgenij Proschak, Impact of X-Ray Structure on Predictivity of Scoring Functions: PPARγ Case Study, *Mol. Inf.*, **2012**, *31*, 631-633.
2. Julia Weber; Janosch Achenbach; Daniel Moser; Ewgenij Proschak, VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization, *J. Med. Chem.,* **2013**, 56, 5203-5207.
3. Estel La Buscató; Dominik Büttner; Astrid Brüggerhoff; Franca M. Klingler; Julia Weber; Bastian Scholz; Aleksandra Živković; Rolf Marschalek; Holger Stark; Dieter Steinhilber, Helge B. Bode; Ewgenij Proschak, From a Multipotent Stilbene to Soluble Epoxide Hydrolase Inhibitors with Antiproliferative Properties. *ChemMedChem* **2013**, 8, 919-923.
4. Dominik Vogt; Julia Weber; Katja Ihlefeld; Astrid Brüggerhoff; Ewgenij Proschak; Holger Stark, Design, synthesis and evaluation of 2-aminothiazole derivatives as sphingosine kinase inhibitors, *Bioorg. Med. Chem.*, **2014**, 22, 5354-5367.
5. Daniel Merk; Christina Lamers; Julia Weber; Daniel Flesch; Matthias Gabler; Ewgenij Proschak; Manfred Schubert-Zsilavecz, Anthranilic acid derivatives as nuclear receptor modulators-Development of novel PPAR selective and dual PPAR/FXR ligands. *Bioorg. Med. Chem.* **2015**, *23*, 499-514.
6. Julia Weber; Janosch Achenbach; Daniel Moser; Ewgenij Proschak, VAMMPIRE-LORD: A Webserver for Straightforward Lead Optimization using Matched Molecular Pairs. *J. Chem. Inf. Model.* **2015**, *55*, 207-2013.

Anhang

Tabellen und Abbildungen

Tabelle A 1 Zuweisung der Pharmakophor-Typen. Die sechs Pharmakophor-Typen werden durch eine Kombination von PDB-Atomtypen und der entsprechenden Aminosäure definiert.

Farbe	Pharmakophor Type
	H-Bindungs-Akzeptor
	Aromatisch
	H-Bindungs-Donor oder H-Bindungs-Akzeptor
	H-Bindungs-Donor
	Hydrophob
	Polar

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Rückgrat																				
CA																				
N																				
O																				
C																				
Seitenkette																				
CB																				
CD																				
CD1																				
CD2																				
CE																				
CE1																				
CE2																				
CE3																				
CG																				
CG1																				
CG2																				

CH2																				
CZ																				
CZ2																				
CZ3																				
ND1																				
ND2																				
NE																				
NE1																				
NE2																				
NH1																				
NH2																				
NZ																				
OD1																				
OD2																				
OE1																				
OE2																				
OG																				
OG1																				
OH																				
SD																				
SG																				

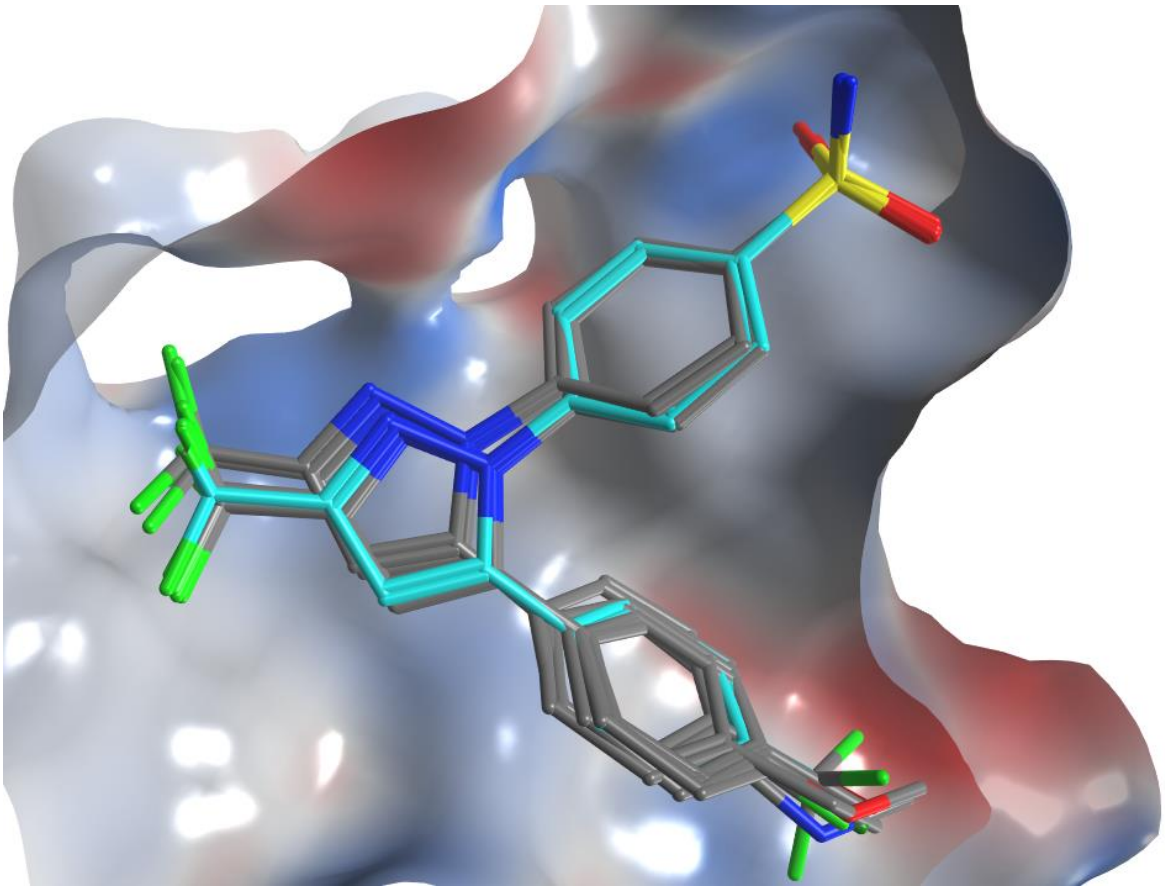


Abbildung A 1 Ergebnis des Dockings der COX-2 Derivate. Der Referenzligand ist in hellblau dargestellt, die übereinander gelagerten Liganden der SAR in grau.

Publikationen

Journal of Medicinal Chemistry

Brief Article

pubs.acs.org/jmc

VAMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization

Julia Weber, Janosch Achenbach, Daniel Moser, and Ewgenij Proschak*

Institute of Pharmaceutical Chemistry, Goethe-University, Max-von-Laue Strasse 9, Frankfurt D-60438, Germany

Supporting Information

ABSTRACT: Structure-based optimization to improve the affinity of a lead compound is an established approach in drug discovery. Knowledge-based databases holding molecular replacements can be supportive in the optimization process. We introduce a strategy to relate the substitution effect within matched molecular pairs (MMPs) to the atom environment within the cocrystallized protein−ligand complex. Virtually Aligned Matched Molecular Pairs Including Receptor Environment (VAMMPIRE) database and the supplementary web interface (http://vammpire.pharmchem.uni-frankfurt.de) provide valuable information for structure-based lead optimization.

INTRODUCTION

Hit-to-lead and lead optimization is a crucial task in drug discovery campaigns. The improvement of the affinity of a small chemical compound to its target lies in the focus of this procedure.[1] Often the information about the three-dimensional (3D) complex derived from X-ray crystallography or NMR spectroscopy is used to support rational optimization. The understanding of the principles in protein−ligand interactions is fundamental for structure-based drug design and protein engineering. Although a wide range of different approaches have been developed to predict the binding affinity of a small-molecule ligand to a protein,[2] the development of these techniques is still ongoing and the predictive power is still far from experimental techniques. However, due to the great amount of available binding affinity data of small molecules to their protein targets, the logical consequence is the use of this data for structure-based drug design and prediction of binding affinity.[2,3] Leach et al. introduced the valuable concept of matched molecular pairs (MMPs) for lead optimization.[4,5] MMPs are defined as two molecules, which differ in one particular substituent and exhibit different properties. The underlying assumption of MMPs is that the difference in properties can be extrapolated to another pair of molecules exhibiting the same substitution pattern. Originally, MMPs have been employed to optimize aqueous solubility, plasma protein binding, and oral exposure.[4] The applicability of MMPs to affinity optimization is limited by the fact that the exchange of a functional group which leads to improvement of binding affinity for one pharmacological target might cause a quite opposite effect in another system. However, the huge amount of information about such exchanges available led to the development of the SwissBioisostere database by Wirth et al.,[6] providing a useful platform for systematic studies of MMPs related to binding affinity. Our present study closes the gap between the MMPs and the structural data, linking the exchange of substituents and the associated binding affinity data with the chemical environment in the protein−ligand complexes. We assume that a change in binding affinity caused by the exchange of the substituent depends on the surrounding atoms in the complex. This assumption provides the possibility to extrapolate from one biological system to another one and consequently might be useful for structure-based drug design and lead optimization as well as for fundamental studies of protein−ligand interactions.

RESULTS AND DISCUSSION

Database Preparation. The creation of the VAMMPIRE database involves the processing of a structural database on the one hand and a large compound library with assigned affinity data on the other hand. The PDBbind v2012[7,8] provides an extensive collection of binding affinity data for biomolecular complexes deposited in the Protein Data Bank (PDB)[9] and is subdivided in three sets of different size and quality (Figure 1). For the preparation of the general set, every primary reference has been reviewed manually and 7986 entries with K_i, K_d, or IC_{50} values were identified. IC_{50} values are largely affected by corresponding assay conditions and therefore not suitable for an independent comparison of binding affinities. For this purpose, the refined set, a subset of the general set, was created containing 3172 entries with K_i or K_d values. It defines additional prerequisites for the complexes, including that only noncovalently bound ligands are accepted, only one ligand is allowed in the binding pocket, and the ligand must contain only common organic elements, i.e., C, N, O, P, S, F, Cl, Br, I, and H. Nonstandard amino acids in the binding pocket of the protein and an X-ray resolution higher than 2.5 Å are not accepted as well. The refined set marks the structural basis for our calculations, while the ChEMBLdb[10] v15 provides 1254575 distinct drug-like compounds with given 2D structure and measured binding affinity of specific targets.

Matched Molecular Pairs. To identify potential matched molecular pairs (pMMPs), we extracted all molecules from ChEMBLdb that exhibit affinity data (K_i/K_d − values) for one of the targets stored in the PDBbind refined set. If more than one affinity value was reported for a molecule, and the difference between the minimum and maximum value was smaller than 1

Received: February 13, 2013
Published: June 4, 2013

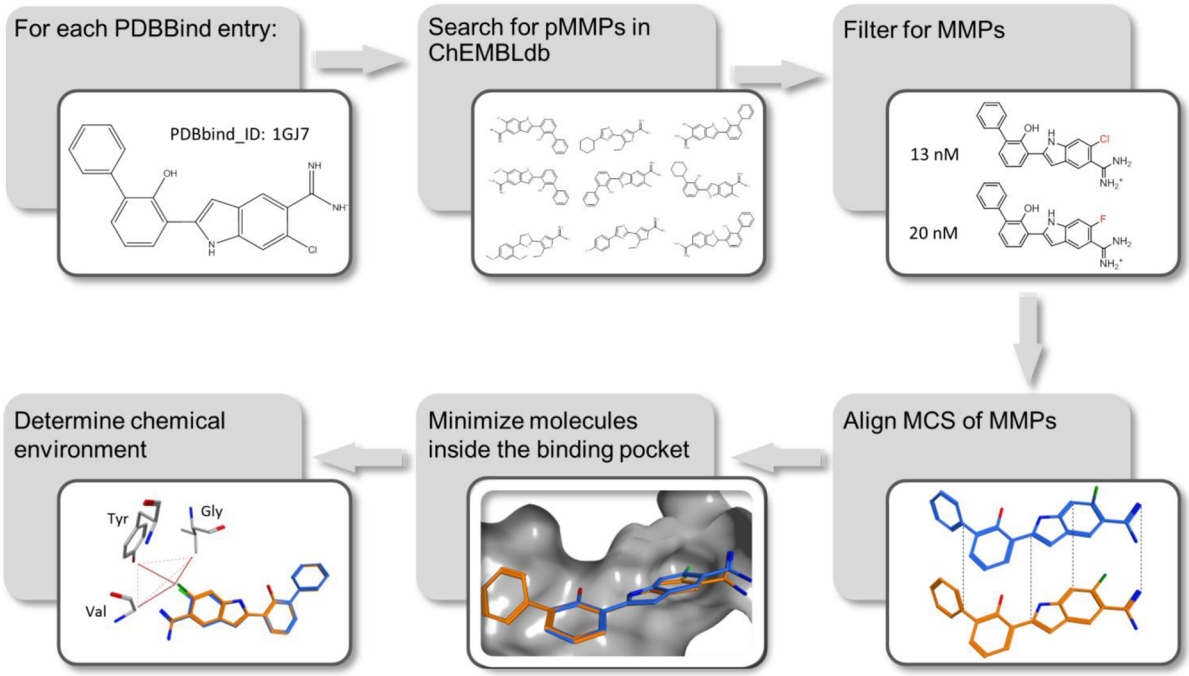


Figure 1. Workflow carried out for each PDBbind ligand. First step: search for ligands binding the same target in ChEMBLdb (pMMPs). Second step: find MCS between molecules of each pMMP and keep valid MMPs. Third step: align MMPs in order to predict 3D coordinates of ChEMBLdb ligands. Fourth step: perform energy minimization within the receptor. Fifth step: determine chemical environment by building tetrahedrons between each substituent atom and the three closest receptor atoms.

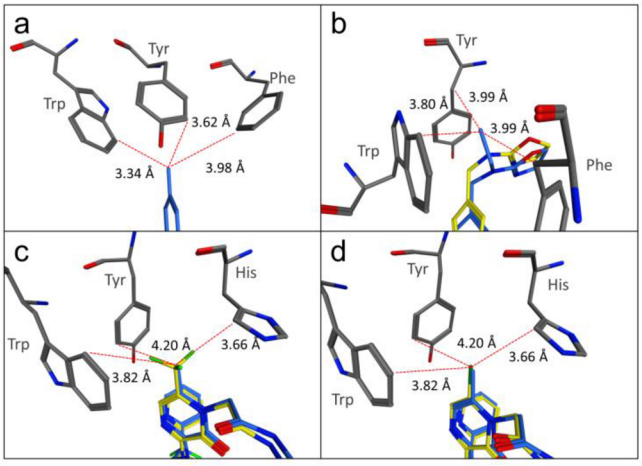


Figure 2. Methyl substitutions in similar chemical environments and distances for the three closest atoms are shown. (a) Extract of Celecoxib bound to COX-2. The 4-methyl group of Celecoxib surrounded by the residues Trp, Tyr, and Phe. (b) Extract of an MMP (−methyl → −ethyl) within factor Xa (PDB code, 1MQ6; CHEMBL_ID, 226564). (c) Extract of an MMP (−methyl → $-CF_3$) within thrombin (PDB code, 1MU6; CHEMBL_ID, 142106). (d) Extract of an MMP (−methyl → −Cl) within thrombin (PDB code, 1MU6; CHEMBL_ID, 30054).

order of magnitude, we took the median value. Otherwise, we discarded the molecule from our data set. To derive the substitution effect of a specific substituent replacement, it is necessary that the molecules differ at exactly one position. Therefore we calculated the maximum common substructure (MCS) between the two molecules of a pMMP (excluding partial ring matching). Additionally, we introduced the restriction that the common core has to be at least twice the size of both substituents, and the substituents contain a maximum of five non-hydrogen atoms for none-cyclic substituents (allowing, e.g., OCF_3 as a substituent) and a maximum of nine non-hydrogen atoms for cyclic substituents (allowing, e.g., a six-membered ring substituted at three positions). We excluded bicyclic rings as the subsequent prediction of the binding mode of big rigid substituents turned out to fail in many cases. The resulting MMPs and the change in their affinity values were used to calculate the effect of the substitution to the overall affinity of the ligand. This effect was calculated as follows:

$$E_{s1,s2} = \log_{10}(a_{m2}) - \log_{10}(a_{m1}) \quad (1)$$

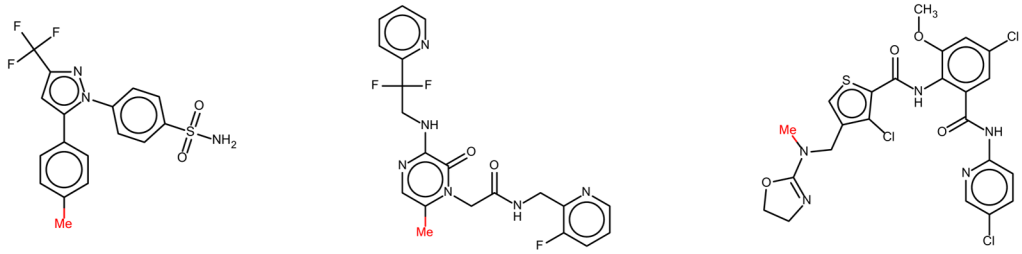
where $E_{s1,s2}$ describes the effect of the replacement of substituent s1 by substituent s2, and a_{m1} and a_{m2} represent the affinity values of molecule one and molecule two respectively.

Prediction of Target Binding Mode. Because of the high similarity between the two molecules of an MMP, we assumed that the binding mode to the target is also very similar.[11] Because ChEMBLdb does not contain 3D conformational information, we transferred the MCS coordinates (including fractional ring matching) of the native PDBbind ligand to the mapped atoms of the ChEMBLdb molecule. The atoms of the substituents cannot be superposed; therefore we applied a coordinate translation by identifying the MCS atom which is connected to the substituent and adopted the associated translation vector to all atoms of the substituent. Subsequently, we applied an energy minimization step using MOE Pose Refinement (AMBER12: EHT force field) to avoid conformational errors in consequence of the displacement, and to predict the position of the substituent inside the binding pocket. ChEMBLdb molecules with a predicted 3D conformation can subsequently act as templates themselves.

Chemical Environment and Statistical Evaluation. We implemented three different representations to describe the chemical environment of the substituents. The most intuitive description is given by the amino acids surrounding the

Table 1. Substitution Effects for Experimental and Predicted Methyl Substitutions in the Environment TYR, TRP

Experimental Data(COX-2)[a] *Data found in VAMMPIRE Database*[b]



Celecoxib (COX-2 IC_{50} = 40 nM)[a] | PDBcode: CDA (factor Xa K_i = 0.007 nM)[c] | PDBcode: XLD (thrombin K_i = 4.2 nM)[c]

substituent	IC_{50} (nM)[a]	effect[d]	K_i (nM)[e]	effect[d]	ChemblID[f]	target[g]
Cl	10	**+0.60**	5.2	**−0.09**	30054	thrombin (1MU6)
Et	860	**−1.33**	0.012	**−0.23**	226564	factor Xa (1MQ6)
CF_3	8230	**−2.31**	44	**−1.02**	142106	thrombin (1MU6)

[a]IC_{50} values determined for COX-2.[12] [b]Search query: methyl-substitution in environment TYR,TRP. [c]K_i values stored in PDBbind.[7,8] [d]Calculated using equation eq 1. [e]K_i values stored in ChEMBLdb for corresponding target. [f]CHEMBL_ID. [g]Target name and PDBcode.

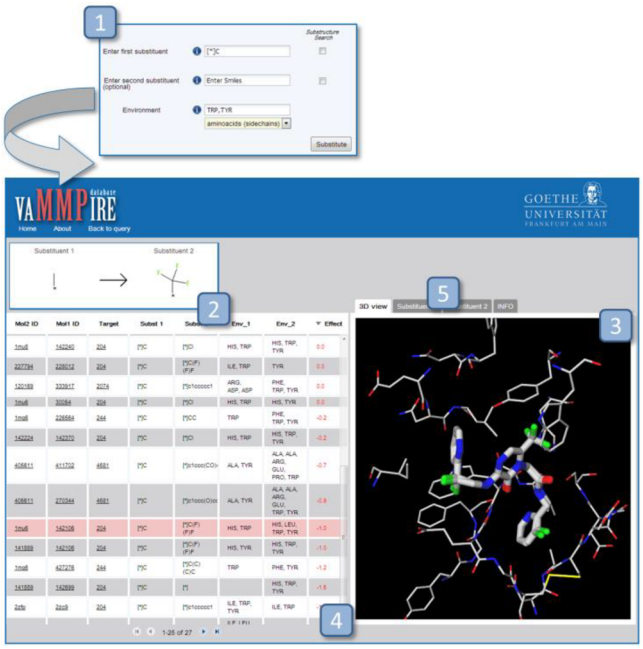


Figure 3. Web interface to access VAMMPIRE database. (1) Search form to add one or two substituents (as smiles) and optionally an arbitrary number of amino acids (three-letter code) or atom types to refine the search. Substructure search can be activated to consider all substituents containing the defined substructure. (2) Depiction of the substituents within the currently selected MMP. (3) 3D molecular viewer showing the aligned MMP within receptor environment. (4) Results table containing molecule and substituent information as well as the corresponding substitution effect. (5) Substitution effect statistics for the selected substituent.

substituent. For each atom of the substituent, we stored the three closest amino acids that are located within a distance of 5 Å (ignoring hydrogen atoms). For the second representation, only amino acids potentially forming side chain interactions are included in the environment description. The three-letter code of amino acids can be used to refine the search for a specific substitution on the webserver.

To generate a more general representation which can be used to establish a relation between an atom type in a specific environment and an associated change in ligand affinity, we used SYBYL Atom Types (see Supporting Information (SI) Table 1) to describe both the ligand and receptor atoms. For each atom in the substituent and its three closest receptor atoms (within a distance of 5 Å, ignoring hydrogen atoms), a tetrahedron was formed. Each tetrahedron was assigned either a positive or negative value depending on whether the associated substitution caused a positive or negative effect to the ligand affinity. A factor of 2 was set as minimum difference between affinity values for a tetrahedron to be included in the statistics. For each substituent, we conveyed a statistical analysis by observing the frequencies of tetrahedrons in conjunction with positive and negative substitution effects. The atom type statistics for each substitution is accessible through the web interface.

On the basis of 2892 complexes stored in the PDBbind refined set, we found 8972 matched molecular pairs (MMPs) within a total of 142 unique targets represented in 589 different crystal structures. For most targets, only one MMP could be found. Approximately one-third of the substituents consist of only one atom.

Database Validation. We demonstrate the usefulness of VAMMPIRE Database by reproducing the structure−activity relationship (SAR) of a target with published affinity data and a cocrystallized structure which is not present in our database. We chose cyclooxygenase 2 (COX-2) cocrystallized with the well characterized inhibitor Celecoxib (PDBcode: 3LN1) and observed the substitution of the 4-methyl group. The first step was to determine the chemical environment of the methyl group inside the binding pocket of the cocrystallized structure. We observed interactions with the aromatic residues Trp373, Tyr371, and Phe367 (Figure 2a) and therefore searched VAMMPIRE Database for methyl substitutions in similar environments.

Searching the exact environment (query: Trp,Tyr,Phe), we found a substitution of the methyl by an ethyl group within the target factor Xa (Figure 2b), and by reducing the environment definition to the closest two amino acids (query: Trp,Tyr), we found two more substitutions showing a very similar interaction

profile with His instead of Phe as a third interaction partner (Figure 2c,d).

The substitution effects (eq 1) calculated for the measured IC_{50} values on the one hand[12] and the effects deposited in VAMMPIRE database on the other hand are shown in Table 1. In both cases, the substitution by chlorine did not affect the compound affinity significantly. Chlorine is known to reveal an electronegative as well as an electropositive potential. Interactions with nucleophiles are carried out by the so-called σ-hole effect,[13−17] which enables the halogens to form nearly linear interactions with electronegative binding partners, in this case with three aromatic moieties. In contrast, fluorine is not able to form "halogen bonds", and therefore the substitution by the electronegative CF_3 in both cases lead to a loss of affinity by more than 1 order of magnitude.

In this study, we were able to collect a database of MMPs suitable for application in structure-based drug design. We extended the classical MMPs approach by incorporating structural information available from cocrystallized protein−ligand complexes and combining it with the large quantity of binding affinity data available for molecules known to bind the same targets. We are aware of the fact that the information about the chemical environment of the substituents is based on a prediction of the three-dimensional coordinates of its atoms and the underlying implication of a similar binding mode. Our preliminary analysis suggests that our initial assumption considering changes in binding affinity caused by the substituent replacement, independently from biological target, might hold true. We therefore feel confident that the VAMMPIRE database and Web Interface (Figure 3) might not only provide valuable information for structure-based lead optimization but also for studies engaged in fundamental understanding of protein−ligand interactions.

EXPERIMENTAL SECTION

Database Preparation. The data was processed using the workflow management tool KNIME (Konstanz Information Miner, KNIME 2.7.2, KNIME.com GmbH, 2011). In the first step, we prepared the refined set by splitting it into two parts, depending on whether K_i- or K_d-values were denoted. All calculations were carried out separately on the two generated subsets in order to ensure the comparability of measured affinity data. The protonation state of the receptors was assigned by means of the function Protonate 3D,[18] available in the software package Molecular Operating Environment (MOE) 2012.10. The MOE Energy Minimization was used to rebuild 3D coordinates of the ChEMBLdb molecules, while the Wash function was used to assign the protonation state to the ligands in both PDBbind refined set and ChEMBLdb. MMPs were detected using the RDKit KNIME integration 2.1.0 (Matched Pairs Detector−Node), which calculates all possible (single-point) chemical transformations.[19−21] The MCS was calculated using the Small Molecule Subgraph Detector (SMSD) toolkit[22] available in the Java library CDK (Chemistry Development Kit 1.4.11).[23] The energy minimization inside the binding pocket was calculated using MOE Pose Refinement (available as KNIME Node) using AMBER12: Extended Hueckel Theory (EHT) force field (as implemented in MOE 2012.10). As the target information in the refined set is only given in terms of the PDBcode, we performed a PDBcode to Chembl_ID mapping via UniProt_ID to identify the given targets in the ChEMBLdb.

Web Interface. We built a web interface to provide the obtained data to public and adapted the database to the requirements we expect from the user (Figure 3). VAMMPIRE was developed using the Google Web Developer Toolkit (GWT 2.5) extended by the Java UI Library GXT (http://www.sencha.com/products/gxt). It employs Java, JavaScript, and PostgreSQL and runs on an Oracle GlassFish 2.1 (http://glassfish.dev.java.net/) application server with PostgreSQL 9.2 (http://www.postgresql.org/) as underlying RDBMS. In principle, every Java EE 5 application server can be used. In the current version, however, VAMMPIRE uses the RDKit PostgreSQL cartridge (http://www.rdkit.org) for substructure search and therefore depends on PostgreSQL. Molecule depiction is handled by the ChemDoodle Web Components (http://web.chemdoodle.com/), while GL mol (http://webglmol.sourceforge.jp/) is used as 3D-viewer.

ASSOCIATED CONTENT

Supporting Information

Table of SYBYL atom types. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49 69 798 29301. E-mail: proschak@pharmchem.uni-frankfurt.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (Sachbeihilfe PR 1405/2-1), Oncogenic Signaling Frankfurt (OSF), Deutsches Konsortium für Translationale Krebsforschung (DKTK), and LOEWE-Schwerpunkt: Anwendungsorientierte Arzneimittelforschung. J.A. thanks Merz Pharmaceuticals and J.W. the Beilstein-Institut zur Förderung der Chemischen Wissenschaften (Beilstein Institute for the Advancement of Chemical Sciences) for a fellowship.

ABBREVIATIONS USED

EHT, extended Hueckel theory; MCS, maximum common substructure; MMP, matched molecular pair; pMMP, potential matched molecular pair; SAR, structure−activity relationship; VAMMPIRE, virtually aligned matched molecular pairs including receptor environment

REFERENCES

(1) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nature Rev. Drug Discovery* **2003**, *2*, 369−378.

(2) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2644−2676.

(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Rev. Drug Discovery* **2004**, *3*, 935−949.

(4) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties: a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* **2006**, *49*, 6672−6682.

(5) Griffen, E.; Leach, A.; Robb, G.; Warner, D. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(6) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41*, 1137−1143.

(7) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 1675−1679.

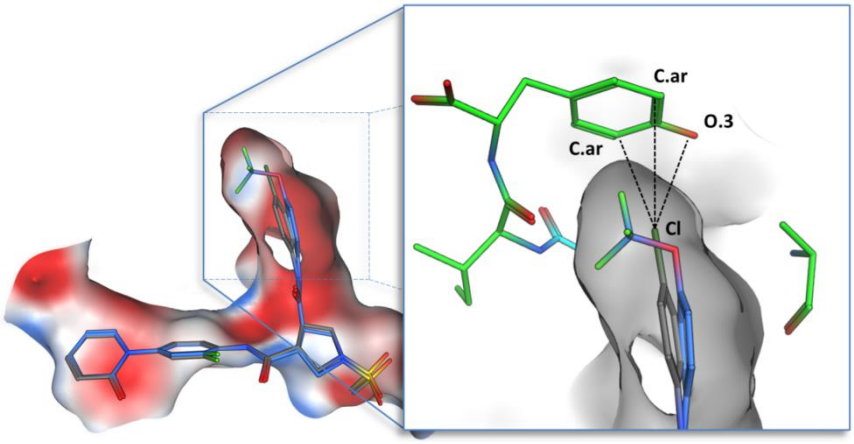
(8) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(9) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(10) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *44*, 1−8.
(11) Kim, K. H. Outliers in SAR and QSAR: Is unusual binding mode a possible source of outliers? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 63−86.
(12) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1*H*-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.
(13) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the sigma-hole. *J. Mol. Model.* **2007**, *13*, 291−296.
(14) Eskandari, K.; Zariny, H. Halogen bonding: A lump−hole interaction. *Chem. Phys. Lett.* **2010**, *492*, 9−13.
(15) Zhang, Y.; Ma, N.; Wang, W. A new class of halogen bonds that avoids the s-hole. *Chem. Phys. Lett.* **2012**, *532*, 27−30.
(16) Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S. An overview of halogen bonding. *J. Mol. Model.* **2007**, *13*, 305−11.
(17) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2012**, *56*, 1363−1388.
(18) Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **2009**, *75*, 187−205.
(19) Wagener, M.; Lommerse, J. P. M. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.
(20) Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(21) Papadatos, G.; Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; MacDonald, S. J. F. Lead optimization using matched molecular pairs: Inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872−1886.
(22) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.* **2009**, *1*, 12.
(23) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

SUPPORTING INFORMATION

VAMMPIRE: a matched molecular pairs database for structure based drug design and optimization

Julia Weber, Janosch Achenbach, Daniel Moser, Ewgenij Proschak



Supplementary Table 1. SYBYL Atom Types (http://www.tripos.com/mol2/atom_types.html)

Code	Definition	Code	Definition	Code	Definition
C.3	carbon sp3	**S.2**	sulfur sp2	**Li**	lithium
C.2	carbon sp2	**S.O**	sulfoxide sulfur	**Na**	sodium
C.1	carbon sp	**S.O2**	sulfone sulfur	**Mg**	magnesium
C.ar	carbon aromatic	**P.3**	phosphorous sp3	**Al**	aluminum
C.cat	carbocation (C^+) used only in a guadinium group	**F**	fluorine	**Si**	silicon
N.3	nitrogen sp3	**Cl**	chlorine	**K**	potassium
N.2	nitrogen sp2	**Br**	bromine	**Ca**	calcium
N.1	nitrogen sp	**I**	iodine	**Cr.th**	chromium (tetrahedral)
N.ar	nitrogen aromatic	**H**	hydrogen	**Cr.oh**	chromium (octahedral)
N.am	nitrogen amide	**H.spc**	hydrogen in Single Point Charge (SPC) water model	**Mn**	manganese
N.pl3	nitrogen trigonal planar	**H.t3p**	hydrogen in Transferable intermolecular Potential (TIP3P) water model	**Fe**	iron
N.4	nitrogen sp3 positively charged	**LP**	lone pair	**Co.oh**	cobalt (octahedral)
O.3	oxygen sp3	**Du**	dummy atom	**Cu**	copper
O.2	oxygen sp2	**Du.C**	dummy carbon	**Zn**	zinc
O.co2	oxygen in carboxylate and phosphate groups	**Any**	any atom	**Se**	selenium
O.spc	oxygen in Single Point Charge (SPC) water model	**Hal**	halogen	**Mo**	molybdenum
O.t3p	oxygen in Transferable Intermolecular Potential (TIP3P) water model	**Het**	heteroatom = N, O, S, P	**Sn**	tin
S.3	sulfur sp3	**Hev**	heavy atom (non hydrogen)		

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

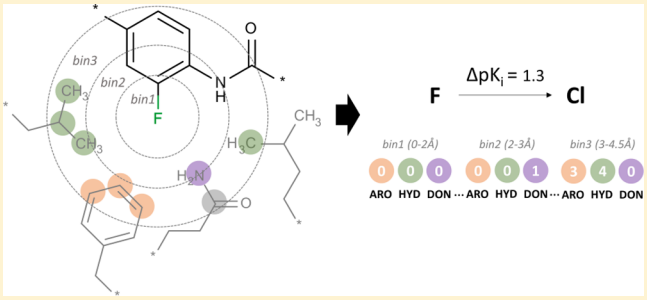
VAMMPIRE-LORD: A Web Server for Straightforward Lead Optimization Using Matched Molecular Pairs

Julia Weber, Janosch Achenbach, Daniel Moser, and Ewgenij Proschak*

Institute of Pharmaceutical Chemistry, Goethe University, Frankfurt 60438, Germany

Supporting Information

ABSTRACT: VAMMPIRE-LORD (lead optimization by rational design) describes an innovative strategy to improve the binding affinity of a defined lead compound using 3D matched molecular pairs (3D-MMPs). 3D-MMPs are defined as pairs of molecules that differ in exactly one structural transformation and have a known bioactive conformation. We developed a novel atom-pair descriptor (LORD_FP) that represents the ligand—as well as the receptor environment—of a chemical transformation and built a predictive model based on 17 602 3D-MMPs. We demonstrate that the created model is able to extrapolate the knowledge of a chemical transformation and the associated effect on ligand affinity to any similar system. VAMMPIRE-LORD was implemented as a web server that guides the user step-by-step through the optimization process of a defined lead compound.



INTRODUCTION

Small structural transformations within a molecule, along with the effect on an arbitrary molecular property sustain valuable information. Statistical models derived from these observations have been used to rationally improve lead compounds regarding a range of particular properties like solubility, plasma protein binding, or oral exposure.[1−3] The so-called matched molecular pairs (MMPs) have gained increasing interest in recent years not only in terms of optimizing physicochemical properties but also with regard to improving the binding affinity of a lead compound to a specific target.[4−7] 2D-MMP methods have been developed to identify transformations improving the binding affinity of a ligand more frequently than the average,[5,8] and it has recently been shown that considering the local environment of a transformation within the ligand can improve the predictive power of 2D-MMPs.[7,9,10] Nevertheless most of the 2D-MMP methods are designed for the characterization of specific targets or target families with a large number of known ligands and the resulting statistics are not transferable to any other target or target family. There are very few methods incorporating the protein environment into MMP calculations.[11] One of these methods is the OOMMPPAA tool,[12] a 3D-MMP method which identifies favorable positions for pharmacophoric features within a specific protein. OOMMPPAA is able to identify promising suggestions for the synthesis of novel compounds which arise from a combined analysis of structural and activity data of known ligands. However, a comprehensive data set of ligands with activity data for the considered target is necessary to generate an adequate number of MMPs for the prediction.

In order to create a target-independent data set we implemented the VAMMPIRE database,[13] a collection of 3D-MMPs in receptor context which enables us to compare the immediate amino acid environment of an arbitrary transformation independently from the rest of the protein structure. In the following we present VAMMPIRE-LORD, a Web server for rational lead optimization, which is based on the VAMMPIRE database and follows the principle that substitutions in similar chemical environments cause similar effects on the ligand affinity. We implemented a novel method to translate the context of a substitution, including the ligand- as well as the receptor environment, into an atom-pair fingerprint (LORD_FP). LORD_FP is a descriptor based on topological atom-pairs inspired by CATS3D[14] and comprises protein−ligand and protein−protein as well as ligand−ligand atom pairs. We extended VAMMPIRE database by the LORD_FP which serves as a basis for the prediction of promising substitutions on particular lead compounds. VAMMPIRE-LORD was implemented as an easy to use wizard which supports the user step-by-step through the process of lead optimization and presents the results in a 3D viewer together with the corresponding substitution information. The Web server is freely available at http://vammpire.pharmchem.uni-frankfurt.de.

RESULTS

We obtained 17 602 MMPs which are deposited in the current version of VAMMPIRE database. The entries are subdivided into three different MMP-types (Figure 1) depending on their reliability. Type-I-MMPs represent about 10% of the database and are of the highest quality as both molecules are available as cocrystallized structures in the PDBbind.[15,16] In a Type-II-

Received: August 30, 2014
Published: January 28, 2015

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207−213

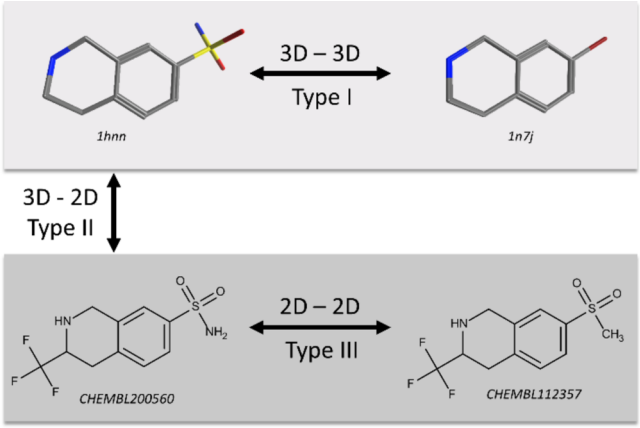


Figure 1. Definition of the MMP types: (Type-I-MMP) both molecules are available as cocrystallized structures in the PDBbind; (Type-II-MMP) one of the molecules is available as cocrystallized structure and serves as a basis for the prediction of the second (2D) molecule; (Type-III-MMP) both molecules are available in 2D only. The predicted coordinates of a molecule within a Type-II-MMP forms the basis for a second prediction.

MMP (20% of the database) the bioactive conformation is known for one of the molecules which forms the basis for the prediction of the conformation of the second molecule. We optimized the procedure to predict the bioactive conformation using MOE[17] (Molecular Operating Environment) docking with pharmacophore placement instead of aligning and energy minimizing the molecules (Figure 2).

In a recently published comparative assessment of docking and scoring functions,[18] it has been shown that most of the scoring functions are able to the predict 60–90% of the bioactive conformations within the top 3 ranked poses. In a previous study,[19] we could also show that docking into a cocrystallized structure containing a similar ligand improves the quality of the docking results. Therefore, molecular docking is in our opinion the most suitable method for the prediction of the bioactive conformation and at the same time serves as a filter for MMPs that most likely have different binding modes. For further improvement of the quality of 3D-MMPs we calculated the root mean square deviation between the common core atoms of the two molecules after the docking placement (core-RMSD). All MMPs with a core-RMSD greater than 1 Å were rejected. The third type of MMPs, the Type-III-MMPs, represent the major part of the database (about 70%) and are of the lowest quality as both molecules originate from ChEMBL database and are not available as cocrystallized structures. Nevertheless we accepted these pairs as Type-III-MMPs if one of the molecules was at the same time part of a Type-II-MMP and therefore had a predicted bioactive conformation. We applied the filter procedure to improve the data quality as we did for Type-II-MMPs.

The chemical environment of a substituent is defined as all non-hydrogen receptor atoms as well as all non-hydrogen ligand atoms located within a 4.5 Å radius around each non-hydrogen atom of the substituent. In case the substituent is a hydrogen atom itself, the radius is taken around its coordinates. A pharmacophore type is thereafter assigned to each atom of the receptor environment (a full listing of the atom types is given in the Supporting Information) and electrotopological state (EState) atom types[20] are assigned to each atom of the ligand environment. The atoms of the substituent itself are also typed using EState atom types. After the chemical environment is defined the LORD_FP is calculated combining ligand–ligand interactions (LLI), protein–ligand interactions (PLI), and protein–protein interactions (PPI). Interactions are defined as atom pairs (bonded or nonbonded) and their distances binned into short distance interactions (≤2 Å), medium distance interactions (>2–3 Å), and long distance interactions (>3–4.5 Å). LLIs and PPIs represent the atom type arrangement within the ligand respectively the protein environment while the PLIs describe the atom type distances between the substituent and receptor atoms. All binned atom type pairs are counted and lead to a numerical descriptor. Taking the example of PLIs, Figure 3 shows a depiction of how the SMART_FP is calculated.

For validation of LORD_FP and within the VAMMPIRE-LORD user interface we use a numeric version of the Tanimoto coefficient to compare two descriptors A and B (minimum similarity = 0, maximum similarity = 1):

$$T(A, B) = \frac{\sum_{i=1}^{n} \mathrm{common}(A_i, B_i)}{\sum_{i=1}^{n} A_i + \sum_{i=1}^{n} B_i - \sum_{i=1}^{n} \mathrm{common}(A_i, B_i)} \quad (1)$$

where common(A_i, B_i) is the number of common atom types in descriptor A and B at position i and A_i and B_i represent the number of atom types in descriptor A and B at position i.

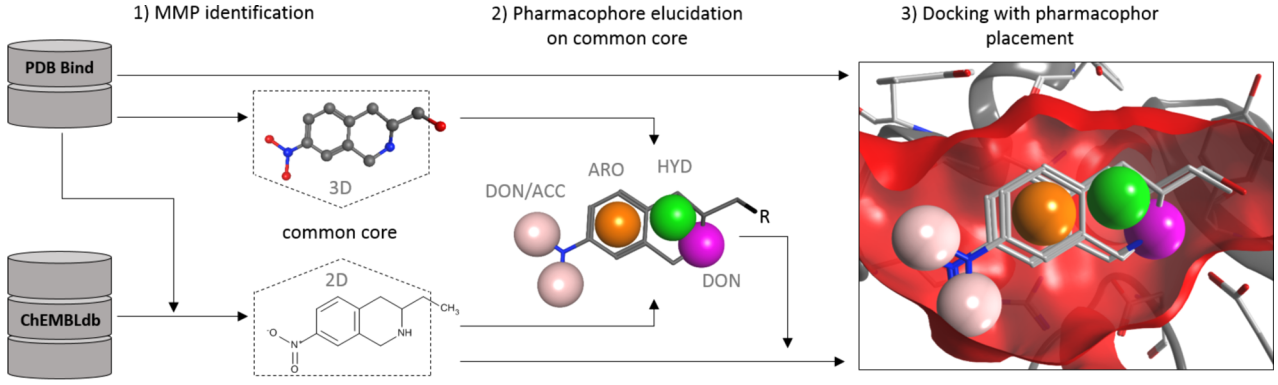


Figure 2. Strategy to generate Type-II-MMPs. (1) Identification of MMPs between ligands from PDBbind and ChEMBL database. (2) Extraction of the molecules common core and pharmacophore annotation on the basis of the 3D conformation of the PDBbind ligand. (3) Docking with pharmacophore placement.

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207–213

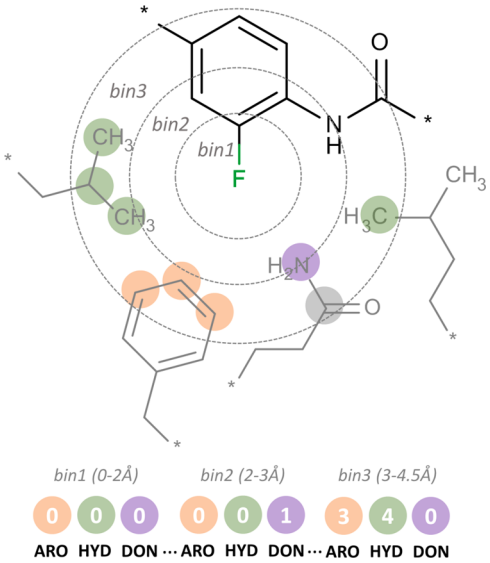


Figure 3. Representation of protein−ligand interactions (PLIs) in LORD_FP. A simplified depiction of a substituent in its protein environment is shown. Three distance bins (2, 3, and 4.5 Å), represented as dotted circles, are formed with their centers on the fluorine substituent (EState atom type: sF). The surrounding amino acids (side chains) are depicted in gray. The pharmacophore types assigned to the side chain atoms are marked as filled circles (hydrophobic: green, aromatic: orange, polar: gray, H-bond donor: violet). An extract of the resulting descriptor is shown below. The number of atom pairs formed between sF and the pharmacophore types within their respective bin form the numerical descriptor.

To provide an interface for medicinal chemists we implemented VAMMPIRE-LORD in the form of a wizard that guides the user step-by-step through the process of lead optimization (Figure 4).

During the steps of the wizard, the user is asked to define the target and the ligand, as well as the substituent of interest. The VAMMPIRE database will subsequently be searched for MMPs containing the desired substituent in a similar chemical environment (SMART_FP similarity ≥ threshold). The results are presented as a table storing the information about the ligands and targets (ChEMBL IDs and PDB codes), the substituents (SMILES), the publications (pubmed IDs), the type of affinity values (K_i, K_d, IC_{50}), the substitution effect, and the descriptor similarity. Additionally a plot, where each data point represents one MMP (Δaffinity plotted against similarity of the substitution environment), gives an overview of the results. Green dots located in the upper right corner represent MMPs with highest positive effects and at the same time highest atom type similarities. By selecting an MMP from the results page the corresponding molecules will appear in the 3D viewer in the context of the receptor environment. Additionally the composition of the receptor atom types and the amino acids surrounding the substituent can be visually inspected.

We were confident that a particular substitution observed for two different targets in a similar chemical environment (in terms of the Tanimoto coefficient between their LORD_FP descriptors) has a similar effect on ligand affinity. Therefore, we calculated the LORD_FP similarity for each target pair with different Enzyme Commission numbers (EC numbers) stored together with the same substitution. The overall distribution of similarity values is shown in Figure 5.

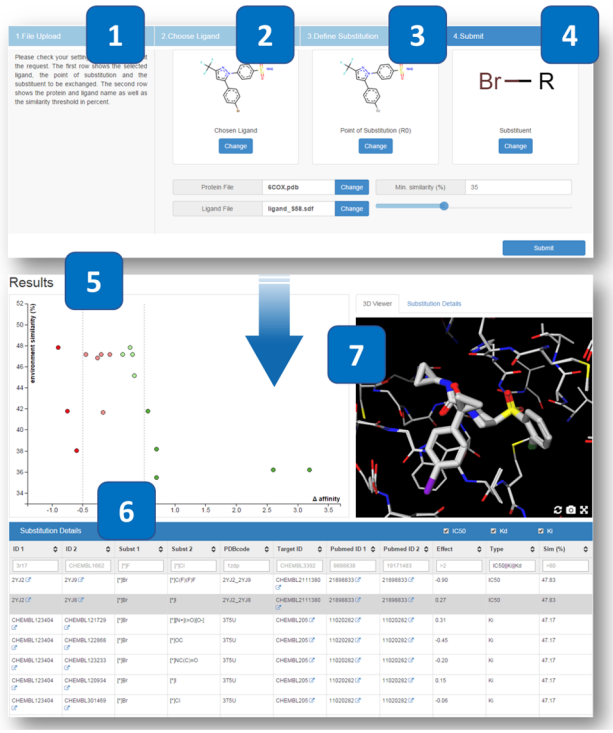


Figure 4. VAMMPIRE-LORD wizard. (1) Input of a PDB code or file upload. (2) Selection of the lead compound. (3) Substituent selection. (4) Summary and definition of a similarity threshold. (5) Plot showing all MMPs that match the query (Δaffinity plotted against similarity in percent). Positive effects are shown in green, and negative effects are shown in in red. (6) Table with details about molecules, substituents, targets, and publications. (7) 3D viewer showing both molecules of an MMP in the context of the receptor environment.

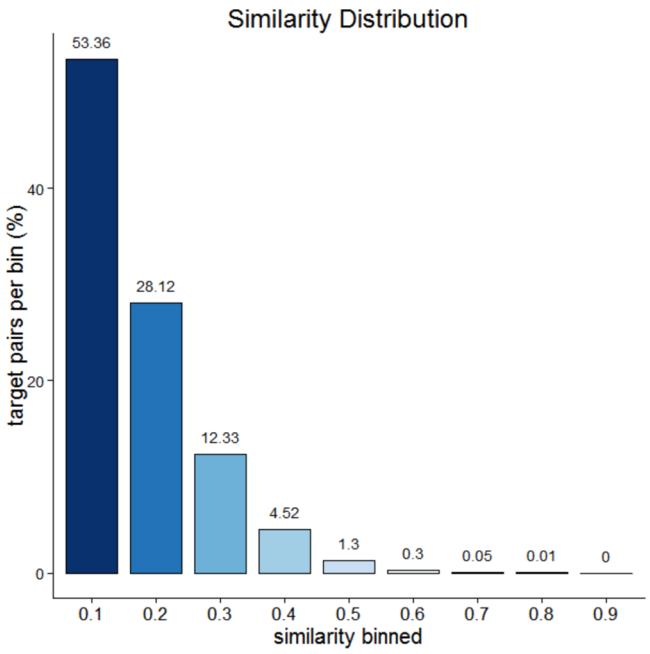


Figure 5. LORD_FP similarity distribution for all target pairs stored in VAMMPIRE database. The data is binned into nine groups representing the number of target pairs with a LORD_FP similarity between 0 and 1 (in steps of 0.1).

To prevent that overrepresented target pairs dominate the statistical evaluation we picked random samples for each target

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207−213

pair and repeated the calculation 20 times. We then plotted the number of matching substitution effects, which is the number of target pairs where both effects are either positive or negative, depending on the LORD_FP similarity (Figure 6). We expected the number of matches to increase with increasing similarity as well as with the strength of the substitution effects on both targets. In order to demonstrate this, we calculated the number of matches for three subsets: one including all target pairs, one including those pairs with an effect of at least 0.5 log units, and one with an effect of at least 1 log unit. Each of these subsets was subsequently analyzed regarding five defined thresholds (0, 0.3, 0.4, 0.5, 0.6) which appeared appropriate after visual inspection of the overall similarity distribution.

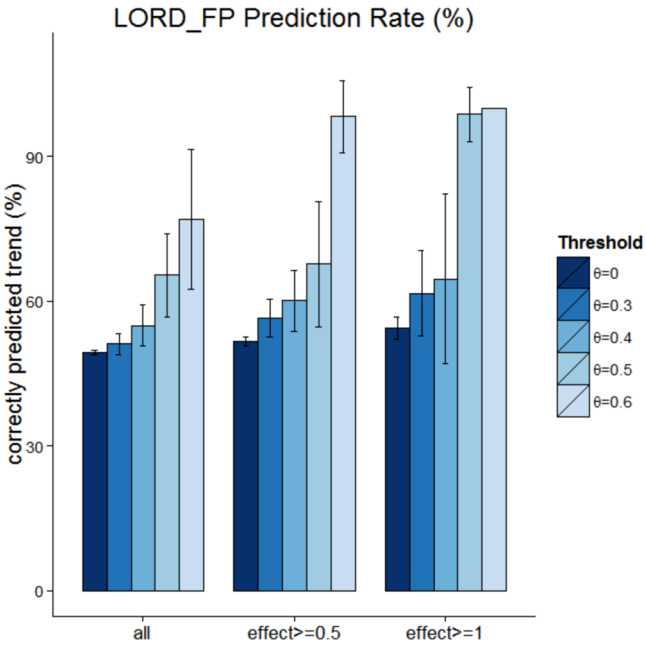


Figure 6. LORD_FP prediction rate. The number of target pairs where both effects are either positive or negative are shown in dependency of the LORD_FP similarity. Three subsets are shown (all: all target pairs, effect ≥ 0.5: target pairs with an effect of at least 0.5 log units, effect ≥ 1: target pairs with an effect of at least 1 log unit). The threshold (θ) gives the minimum similarity of a target pair to be included in the statistics.

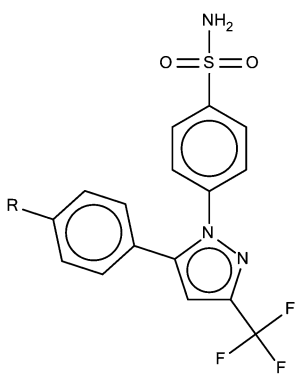
The intrinsic validation of the LORD_FP on VAMMPIRE database shows that the number of correctly predicted trends is highly dependent on the similarity as well as on the strength of the substitution effects. However, it should be noted that the subset of substitutions with effects greater than 1 order of magnitude (for both targets) is comparatively small. In each of the 20 iterations an average of 3655 target pairs are detected of which 21 have a similarity of at least 0.5 and only 5 have a similarity of at least 0.6.

To demonstrate the usefulness of VAMMPIRE-LORD we reproduced an extract of the SAR (structure activity relationship) of cyclooxygenase-2 (COX-2) inhibitors using the SAR of Celecoxib derivatives as a reference[21] (Table 1). This example is especially interesting because the different substituents cover 6 orders of magnitude in inhibitory activity.

All molecules where docked using MOE docking to predict the bioactive conformations (results can be found in the Supporting Information). For each substituent R the LORD_FP was calculated and VAMMPIRE database was then searched for substitutions containing R and one of the other substituents of the COX-2 SAR. Only those substitutions with an effect of at least 0.5 log units and a LORD_FP similarity of at least 0.5 were considered since a mean prediction rate of almost 70% is expected (according to Figure 6). As a result we got a directed substitution network with green and red edges representing matches and mismatches respectively (Figure 7). Eighty-three substitutions with an effect of at least 0.5 log units were found in VAMMPIRE database of which 25 had a LORD_FP similarity of at least 0.5. Eighteen of the 25 substitutions matched the effects of the COX-2 SAR while 7 substitutions were mismatches which corresponds to the expected prediction rate.

The environment with the highest similarity (0.63) was found for the substitution of $-CF_3$ to $-Cl$ together with a positive effect on ligand affinity on both targets (Figure 8). The associated protein is the Endothelial PAS domain-containing protein 1 (EPAS1) and is a result of a Type-III-MMP (CHEMBL2311960 → CHEMBL2311959) placed into the crystal structure with the ligand *N*-(3-chloro-5-fluorophenyl)-4-nitro-2,1,3-benzoxadiazol-5-amine (PDB code: 4GHI). The

Table 1. SAR of Celecoxib Derivatives Forming MMPs



R-group [a]	IC_{50} (μM) [b]	R-group [a]	IC_{50} (μM) [b]
$-NMe_2$	0.005	-NH2	0.34
-OMe	0.008	-OEt	0.64
-SMe	0.009	-Et	0.86
-Cl	0.010	$-NO_2$	2.63
-NHMe	0.016	$-CF_3$	8.23
-H	0.032	$-CO_2H$	11.2
-Me	0.040	$-CH_2OH$	93.3
-F	0.041	-OH	>100

[a]Substitutions at R-group. [b]IC_{50} values determined in a recombinant human COX-2 assay.[21]

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207−213

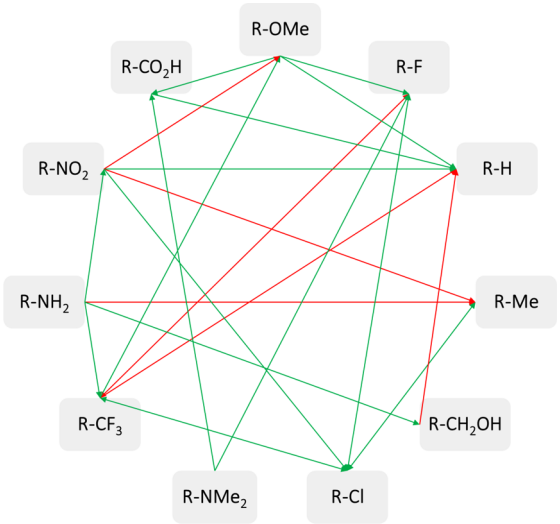


Figure 7. Substitution network representing the results of the VAMMPIRE-LORD search algorithm. All substitutions within the COX-2 SAR (Table 1) were searched in the VAMMPIRE database (effect of at least 0.5 log units and a LORD_FP similarity of at least 0.5). The VAMMPIRE substitutions found are represented as directed edges and colored green when the effect matched the COX-2 SAR and red if the effect did not match.

ligand environment is aromatic in both cases, moreover the amino acids around the substitution are mostly aromatic and hydrophobic. In both cases methionine, phenylalanine, and tyrosine are part of the chemical environment. Although the arrangement and the entire composition of the amino acids is different, the LORD_FP fingerprint similarity is high. Considering the sequence identity of the two proteins (4.3%) this is a good example for the generalizing power of the LORD_FP approach.

DISCUSSION AND CONCLUSION

VAMMPIRE-LORD was implemented as a tool for straightforward lead optimization built up on the idea to use experimental affinity data to create a target-independent model. Based on the assumption that substitutions cause similar effects in similar chemical environments, VAMMPIRE-LORD can be very useful as it is indicating appropriate substitution options. We were able to show that a target independent model to predict a substitution effect (in terms of a trend) is possible by only taking the immediate chemical environment of a substitution into account. With a prediction rate of 70% we could show in the example of the COX-2 SAR that substitutions in similar environments cause similar effects even when the proteins are not closely related. A limiting factor of VAMMPIRE-LORD is certainly the number of MMPs as well as the protein diversity. A simplification of the substitutions like it was done in the recently published Fuzzy matched pairs approach[10] could significantly increase the number of MMPs and may lead to a wider choice of suggestions generated by VAMMPIRE-LORD. However, the permanent increase in the number of protein−ligand complexes in the Protein Data Bank (PDB)[22] as well as the affinity data in ChEMBL database[23] will contribute to future enhancement of the prediction power of this tool.

The confidence of the experimental data deposited in the ChEMBL database is an additional unavoidable problem. Translation or assignment errors of the automatically extracted experimental data are not uncommon.[24,25] Furthermore, the comparison of affinity values determined in different laboratories and assays is far from ideal, especially regarding the small substitution effects. The evaluation of the proposed results by an expert is therefore reasonable to whom VAMMPIRE-LORD can serve as a valuable idea generator for structure-based drug design.

MATRIALS AND METHODS

Database Preparation. We optimized the procedure building VAMMPIRE database to receive a more trustworthy prediction of the 3D conformations. The latest version of VAMMPIRE database was created using the following strategy: The PDBbind v2014, a comprehensive data set of protein−ligand complexes with annotated affinity data obtained from the PDB was the starting point for our data processing. The ChEMBL database v19 expanded the knowledge base by 277 964 compound records with affinity values annotated for the targets deposited in the PDBbind. All procedures were implemented as automatized workflows within the data mining tool KNIME.[26] The MMP identification was carried out using the *Matched Pairs Detector*[7,27,28] available as a node (provided by Erl Wood Cheminformatics) in KNIME. MMPs were detected between the molecules stored in the PDBbind and those molecules stored in the ChEMBL database that have affinity data (K_i, K_d, and IC_{50} values) assigned to one of the PDBbind targets. We implemented a few restriction rules for MMPs including the maximum size of a substituent to be limited to nine non-hydrogen atoms and the molecules common core to be twice as big as the substituents.

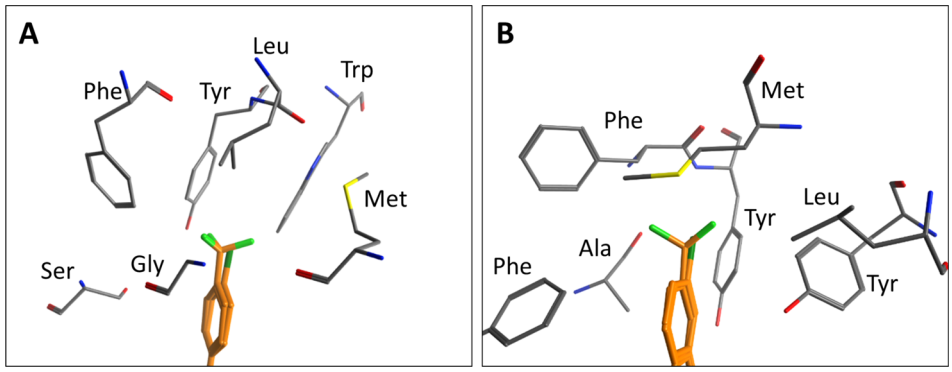


Figure 8. Comparison of the chemical environments of the substitution of $-CF_3$ to −Cl in within COX-2 (A) and EPAS1 (B). All residues within a 4.5 Å radius around the $-CF_3$ group were selected for visualization.

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207−213

Additionally the measured affinity data within an MMP had to be of the same experimental type whereby MMPs with annotated IC_{50} values were only accepted, if they were obtained from the same publication. In case more than one affinity value was stored for a molecule to a specific target the median value was taken. If a value was redundantly represented in the ChEMBL database we took the earliest publication as a reference. The substitution effect was then calculated by subtraction of the $\log_{10}$ affinity values (e.g., an effect of 1.0 represents an increase of ligand affinity by 1 order of magnitude).

3D-MMP Generation. The common core of an MMP is provided by the *Matched Pairs Detector* in terms of a 2D molecule representation. To create a pharmacophore model on the 3D coordinates of the common core, a mapping was performed using the Small Molecule Subgraph Detector (SMSD) toolkit.[29] The resulting 3D-common core was then typed by the ph4 annotation function ("Unified" scheme) provided by MOE as SVL snippet. The annotation points were then translated to a pharmacophore model by adding a 1 Å radius to the annotation points. The created pharmacophore model then was used as a placement method within MOE docking using *Pharmacophore* as a placement method within the *Docking Placement* node implemented in KNIME. Ten poses were created and refined using the *Pose Refinement* node which is also provided by MOE. The grid based minimization with default setup was used for the refinement. The pose with the smallest Root Mean Square Deviation (RMSD) between the common core of the placed molecule and the template was selected as "best pose" for further processing.

Web Server. The web server was implemented in Python using Flask version 0.10.1 (http://flask.pocoo.org/). Client-side 3D visualization is done with a custom GL mol (0.47) build (http://www.glmol.com/). As a database server PostgreSQL 9.3 (http://www.postgresql.org/) with the RDKit Cartridge is used (http://www.rdkit.org). The RDKit is also employed in server-side calculations. Client-side structure input is handled by Marvin4JS (http://www.chemaxon.com/).

■ ASSOCIATED CONTENT

Supporting Information

Assignment of the pharmacophore types and the docking procedure for the COX-2 SAR. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: proschak@pharmchem.uni-frankfurt.de. Mailing address: Institute of Pharmaceutical Chemistry, Goethe University, Max-von-Laue-Str.9, D-60438 Frankfurt am Main.

Funding

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, Sachbeihilfe PR 1405/2-1, SFB 1039 A07) and LOEWE-Schwerpunkt: Anwendungsorientierte Arzneimittelforschung.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.W. thanks the Beilstein-Institut zur Förderung der Chemischen Wissenschaften (Beilstein Institute for the Advancement of Chemical Sciences), J.A., the Else Kröner Fresenius Foundation (EKFS), Research Training Group Translational Research Innovation−Pharma (TRIP), and D.M., the Deutsches Konsortium für Translationale Krebsforschung (DKTK) for a fellowship.

■ REFERENCES

(1) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672−6682.

(2) Griffen, E.; Leach, A.; Robb, G.; Warner, D. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750.

(3) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discovery Today* **2013**, *18*, 724−731.

(4) O'Boyle, N. M.; Boström, J.; Sayle, R. a; Gill, A. Using Matched Molecular Series as a Predictive Tool to Optimize Biological Activity. *J. Med. Chem.* **2014**, *57*, 2704−2713.

(5) Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Bioisosteric Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3*, 425−436.

(6) Posy, S. L.; Claus, B. L.; Pokross, M. E.; Johnson, S. R. 3D Matched Pairs: Integrating Ligand- and Structure-Based Knowledge for Ligand Design and Receptor Annotation. *J. Chem. Inf. Model.* **2013**, *53*, 1576−1588.

(7) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; Macdonald, S. J. F. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of HERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872−1886.

(8) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. B. SwissBioisostere: A Database of Molecular Replacements for Ligand Design. *Nucleic Acids Res.* **2013**, *41*, 1137−1143.

(9) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* **2010**, *50*, 1350−1357.

(10) Geppert, T.; Beck, B. Fuzzy Matched Pairs: A Means to Determine the Pharmacophore Impact on Molecular Interaction. *J. Chem. Inf. Model.* **2014**, *54*, 1093−1102.

(11) Posy, S. L.; Claus, B. L.; Pokross, M. E.; Johnson, S. R. 3D Matched Pairs: Integrating Ligand- and Structure-Based Knowledge for Ligand Design and Receptor Annotation. *J. Chem. Inf. Model.* **2013**, *53*, 1576−1588.

(12) Bradley, A. R.; Wall, I. D.; Green, D. V. S.; Deane, C. M.; Marsden, B. D. OOMMPPAA: A Tool To Aid Directed Synthesis by the Combined Analysis of Activity and Structural Data. *J. Chem. Inf. Model.* **2014**, *54*, 2636−2646.

(13) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization. *J. Med. Chem.* **2013**, *56*, 5203−5207.

(14) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of Correlation Vector Methods for Ligand-Based Similarity Searching. *J. Comput. Aided. Mol. Des.* **2003**, *17*, 687−698.

(15) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 1675−1679.

(16) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(17) Chemical Computing Group Inc. *Molecular Operating Environment (MOE), 2013.08*, 2013.

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207−213

(18) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *J. Chem. Inf. Model.* **2014**, *54*, 1717−1736.
(19) Weber, J.; Rupp, M.; Proschak, E. Impact of X-Ray Structure on Predictivity of Scoring Functions: PPARγ Case Study. *Mol. Inform.* **2012**, *31*, 631−633.
(20) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Model.* **1995**, *35*, 1039−1045.
(21) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1H-Pyrazol-1-Yl]benze Nesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *44*, 1−8.
(24) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K I Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.
(25) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PLoS One* **2013**, *8*, e61007.
(26) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. {KNIME}: The {K}onstanz {I}nformation {M}iner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer: Heidelberg-Berlin, 2007; pp 319−326.
(27) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677−685.
(28) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.
(29) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small Molecule Subgraph Detector (SMSD) Toolkit. *J. Cheminform.* **2009**, *1*, 12−25.

DOI: 10.1021/ci5005256
J. Chem. Inf. Model. 2015, 55, 207−213

Supporting Information

VAMMPIRE-LORD: A Webserver for Straightforward Lead Optimization using Matched Molecular Pairs

*Julia Weber, Janosch Achenbach, Daniel Moser, and Ewgenij Proschak**

Institute of Pharmaceutical Chemistry, Goethe University, Frankfurt 60438, Germany

**Ewgenij Proschak, proschak@pharmchem.uni-frankfurt.de*

Table S1. Assignment of pharmacophore types. The 6 pharmacophore types are characterized by a combination of the PDB atom type and the associated amino acid:

Color	Pharmacophore type
	H-bond acceptor
	aromatic
	H-bond donor or H-bond acceptor
	H-bond donor
	hydrophobic
	polar

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Backbone																				
CA																				
N																				
O																				
C																				
Sidechain																				
CB																				
CD																				
CD1																				
CD2																				
CE																				
CE1																				
CE2																				
CE3																				
CG																				
CG1																				
CG2																				
CH2																				
CZ																				
CZ2																				
CZ3																				
ND1																				
ND2																				
NE																				
NE1																				
NE2																				
NH1																				
NH2																				
NZ																				
OD1																				
OD2																				
OE1																				
OE2																				
OG																				
OG1																				
OH																				
SD																				
SG																				

Docking setup for COX-2 validation

The crystal structure of COX-2 with 1-phenylsulfonamide-3-trifluoromethyl-5-parabromophenylpyrazole (PDB code: 6COX) was used as a template for molecular docking. The structure was protonated by means of the function *Protonate 3D*[1] and energy minimized using the Amber12:EHT forcefield, both implemented in the software MOE.[2] MOE docking is used for ligand placement (*Triangle Matcher*) as well as for scoring (*London dG*), energy minimization (Amber12:EHT) and rescoring (*GBVI/WSA dG*). The conformation with the highest score for each molecule is used for VAMMIRE-LORD predictions.

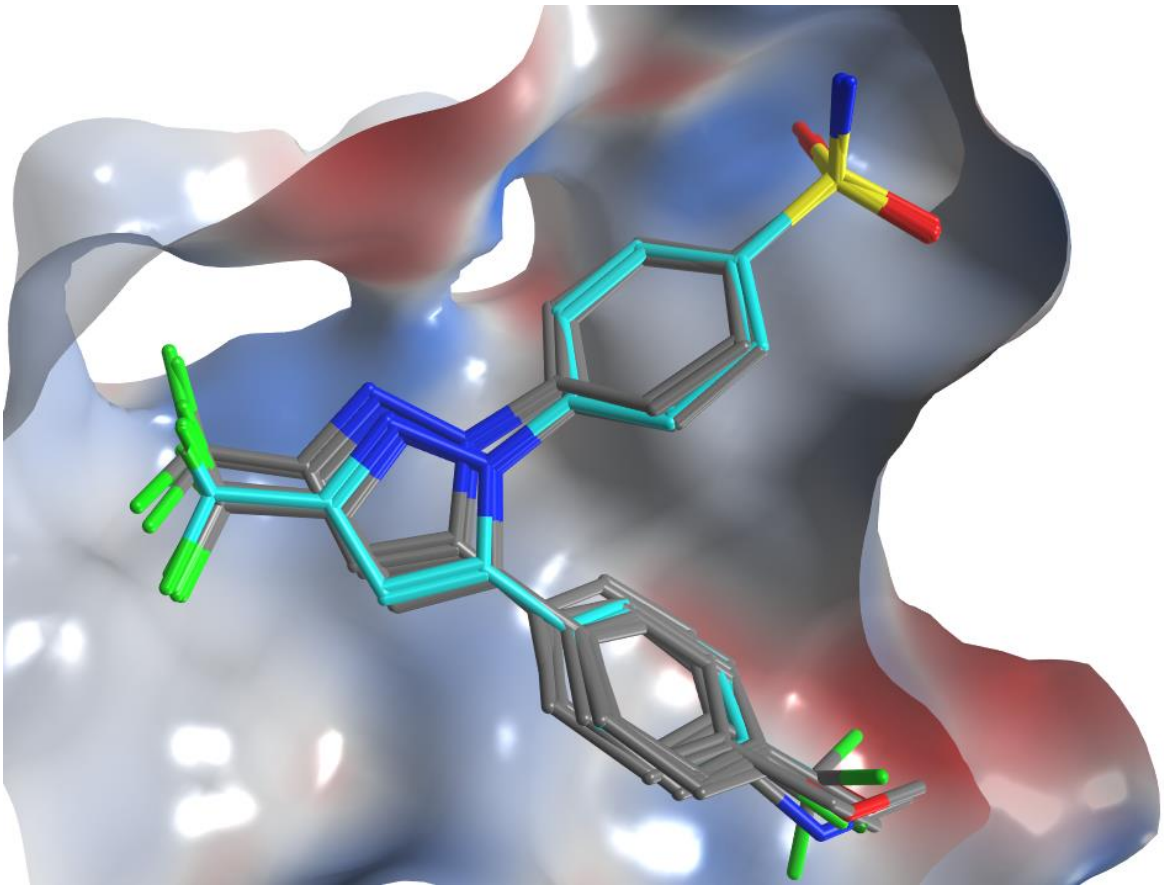


Figure S1. Superposition of the COX-2 inhibitors. Template ligand (light blue) and docked molecules (grey).

References

1. Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins* **2009**, *75*, 187–205.

2. Chemical Computing Group Inc. Molecular Operating Environment (MOE), 2013.08. *Molecular Operating Environment (MOE), 2013.08*, 2013.