

Text-Based Similarity Searching for Hit- and Lead-Candidate Identification

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Biowissenschaften (15)
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Volker Dirk Hähnke
aus Frankfurt am Main

Frankfurt 2010
(D 30)

vom Fachbereich Biowissenschaften (15) der

Johann Wolfgang Goethe-Universität als Dissertation angenommen

Dekan: Prof. Dr. Anna Starzinski-Powitz

Gutachter: Prof. Dr. Gisbert Schneider
Prof. Dr. Ina Koch

Datum der Disputation:

*“Gegenüber der Fähigkeit, die Arbeit
eines einzigen Tages sinnvoll zu ordnen,
ist alles andere im Leben ein
Kinderspiel.”*

Johann Wolfgang von Goethe
(28.8.1749 – 22.3.1832)

Table of Contents

1 - Abbreviations	6
2 - Zusammenfassung	9
3 - Abstract	14
4 - Introduction	16
4.1 - <i>The Drug Development Process</i>	16
4.2 - <i>From High-Throughput Screening to Virtual Screening</i>	17
4.3 - <i>Chemical Similarity</i>	21
4.4 - <i>Line Notations</i>	23
4.4.1 - <i>Wiswesser Line-Formula Notation</i>	24
4.4.2 - <i>Representation of Organic Structures Description Arranged Linearly</i>	25
4.4.3 - <i>Simplified Molecular Input Line Entry System</i>	26
4.4.4 - <i>IUPAC International Chemical Identifier</i>	26
4.5 - <i>Virtual Screening employing Line Notations</i>	29
4.5.1 - <i>LINGO</i>	30
4.5.2 - <i>Comparison by Compression</i>	30
4.5.3 - <i>General String Metrics</i>	31
5 - Study Objective	32
5.1 - <i>Pharmacophore Alignment Search Tool (PhAST)</i>	32
5.2 - <i>Preliminary Parameterization</i>	39
5.2.1 - <i>Scoring System</i>	39
5.2.2 - <i>Alignment Evaluation</i>	40
5.3 - <i>Retrospective Evaluation</i>	41
5.3.1 - <i>Dataset</i>	41
5.3.2 - <i>Performance Measure</i>	41
5.3.3 - <i>Significance Assessment</i>	44
6 - Influence of Canonical Atom Labeling on Similarity Searching	45
6.1 - <i>Motivation</i>	45
6.2 - <i>Discussion</i>	47
7 - Influence of the Third Dimension on Text-based Similarity Searching	49
7.1 - <i>Motivation</i>	49
7.2 - <i>Discussion</i>	50
8 - Influence of Scoring Systems on Text-based Similarity Searching	52
8.1 - <i>Motivation</i>	52
8.2 - <i>Discussion</i>	53
9 - Comparison of Text-Based Virtual Screening Techniques	55
10 - Significance-Assesment in Global Sequence Alignment	57
10.1 - <i>Motivation</i>	57
10.2 - <i>Calculation of p-values</i>	58
10.2.1 - <i>Simple Sampling</i>	59
10.2.2 - <i>Sampling of Rare Events</i>	60
10.3 - <i>Retrospective Evaluation</i>	64
10.3.1 - <i>Parameterization</i>	64
10.3.2 - <i>Results and Discussion</i>	65
10.4 - <i>Calculation of E-values</i>	67
10.5 - <i>Discussion</i>	69

11 - Prospective Application	73
<i>11.1 - Bacterial Thymidinkinase of Staphylococcus aureus</i>	73
<i>11.2 - Application to γ-Secretase</i>	75
12 - Conclusions	79
13 - Outlook	81
14 - List of Publications	84
15 - References	88
16 - Acknowledgements	102
17 - Appendix	103
<i>Appendix A</i>	
<i>Appendix B</i>	
<i>Appendix C</i>	
<i>Appendix D</i>	
18 - Curriculum Vitae	

1 - Abbreviations

2D	two-dimensional
3D	three-dimensional
A β	Amyloid- β
ACE	Angiotensine-converting enzyme
AD	Alzheimer's Disease
APP	Amyloid Precursor Protein
AWLN	Advanced Wiswesser Line-Formula Notation
BEDROC	Boltzmann-enhanced Receiver Operating Characteristic
BLOSUM	Block Substitution Matrix
CANGEN	Canonization and Generation
CbC	Comparison by Compression
COBRA	Collection Of Bioactive Reference Analogues
COX	Cyclooxygenase
CROSSBOW	Computer Retrieval of Organic SubStructures by means of Wiswesser
CSI	Chemical Substructure Index
CUDA	Compute Unified Device Architecture
DDP	Double Dynamic Programming
DHFR	Dihydrofolatreductase
dMTP	Deoxythymidine Monophosphate
EF	Enrichment Factor
ELISA	Enzyme-linked immunosorbent assay
FAST	Fragment Alignment Search Tool
FDA	Food and Drug Administration
FPGA	Field Programmable Gate Array
FSM	Finite State Machine
FXA	Factor Xa
GPU	Graphics Processing Unit
GS	γ -Secretase
GSI	γ -Secretase Inhibitor
GSM	γ -Secretase Modulator

HTS	High Throughput Screening
IC ₅₀	Inhibitory Concentration 50%
InChi	International Union of Pure and Applied Chemistry International Chemical Identifier
InChiKey	International Union of Pure and Applied Chemistry International Chemical Identifier Key
IUPAC	International Union of Pure and Applied Chemistry
LBVS	Ligand-based Virtual Screening
MCMC	Marcov Chain Monte Carlo
MCMCMC	Metropolis-coupled Marcov Chain Monte Carlo
MCS	Maximal Common Subgraph
MIC	Minimal Inhibitory Concentration
MOE	Molecular Operating Environment
MOS	Maximum Overlapping Set
MQL	Molecular Query Language
NID	Normalized Information Distance
NIST	National Institute of Standard and Technology
NP	Non-deterministic Polynomial Time
NSAID	Non-Steroidal Anti-Inflammatory Drug
OpenGL	Open Graphics Library
PAM	Point Accepted Mutations
PhAST	Pharmacophore Alignment Search Tool
PID	Percent Sequence Identity
PPAR	Peroxisome-Proliferator Activated Receptor
PPP	Potential Pharmacophoric Point
PSI-BLAST	Position-Specific Iterated Basic Local Alignment Search Tool
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROCAUC	Receiver Operating Characteristic Area Under Curve
ROSDAL	Representation of Organic Structures Description Arranged Linearly
SBVS	Structure-based Virtual Screening
SHA	Secure Hash Algorithm
SMILES	Simplified Molecular Input Line Entry System
SPP	Similar Property Principle

SSE2	Streaming Single Instruction Multiple Data Streams Extensions 2
SXT	combination of Trimethoprim and Sulfamethoxazole
THR	Thrombine
VEGFR	Vascular Endothelial Growth-Factor Receptor
VS	Virtual Screening
WLN	Wiswesser Line-Formula Notation

2 - Zusammenfassung

Die Entwicklung neuer Wirkstoffe ist ein langwieriger und kostenintensiver Prozess, der bis zu 15 Jahre dauern und 2 Milliarden Dollar kosten kann. Das ‚High Throughput Screening‘ (HTS) hat sich in diesem Prozess als Technik für die Identifizierung vielversprechender Startstrukturen, so genannter ‚Hits‘, etabliert. Während eines HTS werden 50.000 bis 100.000 Substanzen automatisiert in einem Assay auf ihre biologische Aktivität getestet. Setzt man diese Anzahl evaluierter Substanzen in Relation zu vorsichtigen Schätzungen der Gesamtzahl möglicher wirkstoffartiger Verbindungen (10^{60}), wird klar, dass mit HTS allein ein großer Teil dieses ‚Chemischen Raums‘ unerforscht bleibt.

Eine schnellere Alternative bieten computerbasierte Methoden. Ist eine Struktur mit einer gewünschten biologischen Wirkung bekannt, ist es mit diesen Methoden möglich, die Einträge in Molekülsammlungen nach ihrer berechneten Ähnlichkeit zu dieser Referenzstruktur zu sortieren. Diese Technik wird als virtuelles Screening bezeichnet. Die Annahme hierbei ist, dass Substanzen, die als ähnlich zur verwendeten Referenzstruktur bewertet werden auch in ihren biologischen Wirkeigenschaften ähnlich zu dieser sind.

In dieser Arbeit wurde eine neue Methode entwickelt und evaluiert, mit der sich die Ähnlichkeit zweier Moleküle berechnen lässt. Die Bezeichnung dieser Methode ist ‚Pharmacophore Alignment Search Tool‘ (PhAST). In dieser Methode werden Moleküle verglichen durch paarweises globales Sequenzalignment, einer Technik für den Vergleich von Zeichenketten. Sie wurde bisher nur auf Sequenzen aus Aminosäuren oder Nukleotiden angewendet, um Homologe zu identifizieren. In einem Sequenzalignment werden die Symbole zweier Sequenzen einander zugeordnet, wobei die Reihenfolge der Symbole innerhalb jeder Sequenz erhalten bleibt. Das Einfügen von Lücken („Gaps“) in Sequenzen ist erlaubt, wenn es die Gesamtzuordnung verbessert. Werden gleiche Symbole einander zugeordnet, wird dies als ‚Match‘ bezeichnet, bei ungleichen Symbolen wird dies als ‚Mismatch‘ bezeichnet. Jedes dieser Ereignisse wird bewertet. Der Score eines Alignments wird berechnet als die Summe der Einzelbewertungen. Die in dieser Arbeit verwendeten Algorithmen berechnen stets das ‚optimale‘ Alignment, also das, das den höchstmöglichen Alignment Score hat.

Bedingt durch die Unterschiede zwischen Biopolymeren und wirkstoffartigen Molekülen wurde Sequenzalignment auf die Problemstellung des Molekülvergleichs angepasst und neu parametrisiert. Mit allen Parametrisierungen wurde PhAST in

retrospektiven Screenings auf seine Fähigkeit getestet, mit einer aktiven Substanz als Referenz andere aktive Substanzen zu erkennen und für diese höhere Ähnlichkeiten zu berechnen als für inaktive Substanzen. Werden die Einträge einer Molekülsammlung nach den berechneten Ähnlichkeiten absteigend sortiert, konzentrieren sich so die aktiven Moleküle am Beginn der Rangliste, verglichen mit einer uniformen Verteilung über die gesamte Molekülsammlung (Anreicherung). Die Grundlage dieser retrospektiven Experimente war die Wirkstoffsammlung COBRA, die in der verwendeten Version 6.1 insgesamt 8,311 wirkstoffartige Moleküle enthält. Dabei wurden die aktiven Liganden von insgesamt sechs verschiedenen Zielproteinen jeweils einmal als Referenz verwendet.

PhAST berechnet nicht die strukturelle sondern die funktionelle Ähnlichkeit zwischen Molekülen. Um dies zu erreichen, wurde eine Abstraktion jedes Moleküls erstellt, die aus potentiellen Interaktionspunkten besteht. Die Zuweisung dieser Interaktionsmöglichkeiten geschah basierend auf einer Sammlung von Fragmenten, in der jedem nicht Wasserstoff Atom eines Fragments bereits eine Interaktionsmöglichkeit zugewiesen war. Immer, wenn ein Molekül ein Fragment als Substruktur aufwies, wurden die Zuweisungen aus dem Fragment auf die korrespondierenden Atome des Moleküls übertragen. Insgesamt wurde zwischen den folgenden neun Interaktionstypen unterschieden: positive Ladung, negative Ladung, aromatisch, lipophil, Wasserstoffbrücken Akzeptor, Wasserstoffbrücken Akzeptor kombiniert mit Wasserstoffbrücken Donor, Wasserstoffbrücken Akzeptor kombiniert mit Polarität, Wasserstoffbrücken Akzeptor kombiniert mit Wasserstoffbrücken Donor und Polarität sowie keiner möglichen Interaktion. Jeder dieser neun Typen wurde durch ein einziges Symbol repräsentiert.

Sequenzen aus Aminosäuren oder Nukleotiden sind unverzweigt, azyklisch und gerichtet. Wirkstoffartige Moleküle hingegen sind verzweigt, enthalten Ringschlüsse und sind ungerichtet. Um paarweises globales Sequenzalignment zum Vergleich von wirkstoffartigen Molekülen nutzen zu können, mussten diese folglich zunächst in einer linearisierten Form gespeichert werden. Die Notwendigkeit dieses Schritts wurde in dieser Arbeit bewiesen. Die Umwandlung von Molekülen in Zeichenketten muss eindeutig sein in dem Sinn, dass für ein Molekül nur eine einzige Zeichenkette generiert werden kann. Dies ist notwendig, damit identische Moleküle durch die Identität ihrer linearen Repräsentationen erkannt werden können. Um dies sicherzustellen, wurden verschiedene Algorithmen implementiert und evaluiert, die den Atomen in einem Molekül einen eindeutigen Satz von Indizes zuweisen. Die Zuweisung der Indizes zu den Atomen ist eindeutig, es wird also jedem Atom stets derselbe Index zugewiesen, unabhängig davon, in welcher Form das Molekül an den

Algorithmus übergeben wird. Die zugewiesenen Indizes bestimmten die Reihenfolge, in der die mit den Eigenschaften der Atome korrespondierenden Symbole zu einer Zeichenkette zusammengesetzt wurden, beginnend beim niedrigsten Index. Die evaluierten Methoden lassen sich in zwei Klassen einteilen: Algorithmen die für die kanonische Indizierung von Molekülgraphen und Methoden zur Dimensionsreduktion. Die Methode, mit der PhAST in den retrospektiven Studien am besten abschnitt, war ‚Minimum Volume Embedding‘. Dies ist eine Methode zur nichtlinearen Dimensionsreduktion, die in dieser Arbeit mit topologischen Distanzen gemessen über einen Diffusionskernel kombiniert wurde.

Für die Berechnung von Sequenzalignments ist ein Bewertungssystem nötig, das das wechselseitige Zuweisen gleicher oder ungleicher Symbole bewertet. Solche Bewertungssysteme existierten bisher nur für Aminosäuren und Nukleotide. Im Rahmen dieser Arbeit wurden eine stochastische sowie zwei systematische Methoden entwickelt, mit denen solche Bewertungsschemata berechnet werden können. In den systematischen Varianten wurden die Ereignisse bewertet in Abhängigkeit ihrer Häufigkeit in paarweisen Alignments beziehungsweise durch eine Kernelfunktion berechneter Atomzuweisungen, die in einem Referenzdatensatz berechnet und zu den Gesamthäufigkeiten der beteiligten Symbole in Relation gesetzt wurden. Die resultierenden Bewertungssysteme wurden untereinander verglichen sowie mit zwei weiteren Bewertungsmöglichkeiten. In einer wurden alle Matches sowie alle Mismatches gleich bewertet. Im letzten Bewertungssystem wurden die verschiedenen Ereignisse bewertet basierend auf den relativen Häufigkeiten der beteiligten Symbole und dem Grad, zu dem sich die durch sie repräsentierten Funktionalitäten entsprechen. Mit dem zuletzt vorgestellten Bewertungsschema erzielte PhAST in retrospektiven Experimenten die höchste Anreicherung. Der beobachtete Unterschied war signifikant. Das einheitliche Bewertungsschema erzielte signifikant schlechtere Anreicherung verglichen mit den übrigen Schemata.

Sequenzalignment als Methode für den Vergleich von Zeichenketten ist umfassend parametrisierbar. Dadurch konnte das Bewertungsschema weitergehend modifiziert werden. So war es möglich, Symbole in einer Zeichenkette stärker zu gewichten, die Interaktionsmöglichkeiten entsprachen, von denen bekannt war, dass sie essentiell für die Rezeptor-Ligand-Interaktion sind. Am Beispiel des Peroxisom-Proliferator-aktivierten Rezeptors wurde demonstriert, dass mit einer sinnvoll gewählten Gewichtung signifikant erhöhte Anreicherung erzielt werden kann. Es wurde gezeigt, dass die systematische Anwendung von Gewichten auf alle Positionen in retrospektiven Experimenten dazu geeignet ist, essentielle Interaktionspunkte zu identifizieren. Dafür ist es allerdings notwendig, dass ein

entsprechender Datensatz mit einer ausreichenden Anzahl von Strukturen vorhanden ist. Es konnte gezeigt werden, dass Sequenzalignment auch für die Berechnung struktureller Ähnlichkeiten benutzt werden kann.

Es wurden verschiedene Algorithmen für die Berechnung von globalen Sequenzalignment veröffentlicht. Die Standardlösung dieses Problems ist der Algorithmus von Needleman und Wunsch, der in seiner generalisierten Form eine Laufzeit von $O(n^3)$ hat. In dieser Arbeit wurde zunächst eine angepasste Version dieses Algorithmus verwendet mit einer Laufzeit von $O(n^2)$. Ein weiterer Algorithmus wurde implementiert und evaluiert, der zwar die gleiche asymptotische Laufzeit hat, in der Praxis jedoch nur 40% der Zeit benötigt, um die gleiche Menge von Sequenzen zu alignieren. Dies wird durch die Vereinfachung erreicht, dass im Alignment ein Gap in einer Sequenz nicht auf einen Gap in der anderen Sequenz folgen darf. Dies reduziert die Anzahl der Rechenoperationen, die zur Berechnung eines Alignment nötig sind. In einigen Fällen wurden so jedoch Sequenzalignments berechnet, die von denen des Needleman Wunsch Algorithmus abwichen. Es konnte aber gezeigt werden, dass diese Abweichungen auf die von PhAST berechneten Sortierungen von Molekülen nur geringen Einfluss hatten. Die entstehenden Ranglisten waren nahezu identisch, was sich in einer hohen und als signifikant berechneten Rangkorrelation widerspiegelte. Daher wurde für PhAST der schnellere Algorithmus verwendet.

Um die Ähnlichkeit von Zeichenketten aus deren Alignment zu berechnen, müssen die Alignments bewertet werden. Für die Alignments von Aminosäuresequenzen wurden bereits verschiedene Maße entwickelt: die Sequenzidentität, der Alignment Score und die Signifikanz des Alignment Scores. Alle drei Ansätze wurden in verschiedenen Varianten implementiert und evaluiert. Es konnte gezeigt werden, dass wie auch für Aminosäuresequenzen, der Score eines Alignments besser geeignet ist um Ähnlichkeiten zu identifizieren. Mit dem Alignment Score als Bewertungskriterium erzielte PhAST signifikant höhere Anreicherung verglichen mit der Sequenzidentität. Zur Bewertung der Signifikanz des Scores eines Alignments wurden p-Werte berechnet. Die mit ihnen erzielte Anreicherung war vergleichbar mit der, die mit dem Alignment Score erzielt wurde. Über die Rangkorrelation der zugehörigen Ranglisten konnte dennoch gezeigt werden, dass die berechneten Molekülsortierungen nicht identisch sind. Zur Berechnung von p-Werten war es zwingend erforderlich, die Verteilung von Alignment Scores zu kennen für die jeweiligen Paare von Sequenzlängen. Auch für Aminosäuresequenzen ist die Verteilung der Scores globaler Alignments nicht bekannt. Folglich mussten für die Berechnung von p-Werten in PhAST die Verteilungen von Alignment Scores simuliert werden. Dies geschah mit einer Kombination aus Markov Chain

Monte Carlo Simulationen und Importance Sampling. Nachdem die Verteilungen bestimmt waren, wurde für jeden Alignment Score das Integral der zugehörigen Verteilung oberhalb dieses Wertes als p-Wert berechnet. Für die Berechnung von E-Werten wurden die berechneten p-Werte einer Bonferroni Korrektur unterzogen, so dass sie die Gesamtzahl der Einträge in der Molekülsammlung berücksichtigen. Als Ergebnis dieser Arbeit wurde für die Signifikanz mit PhAST berechneter Ähnlichkeiten ein Grenzwert von $1 \cdot 10^{-5}$ vorgeschlagen: Alignments mit einem E-Wert unterhalb dieses Grenzwerts werden als signifikant angesehen.

PhAST wurde in retrospektiven Experimenten mit anderen Methoden zum virtuellen Screening verglichen, die bereits in der Wirkstoffentwicklung eingesetzt werden. Es konnte gezeigt werden, dass die mit PhAST erzielte Anreicherung vergleichbar oder höher war. Allerdings waren die von PhAST berechneten Ranglisten sehr unähnlich zu denen anderer Methoden. Folglich ist es mit PhAST möglich, in einem Screening auf den frühen Rängen der berechneten Ranglisten eine ähnliche Anzahl von aktiven Substanzen anzureichern, die sich jedoch von den mit anderen Methoden identifizierten Hits unterscheiden. Das macht PhAST zu einem wertvollen neuen Bestandteil der frühesten Phase der Wirkstoffentwicklung, da mit dieser neuen Methode Hits identifiziert werden können, die mit anderen Methoden nicht gefunden werden. Die Anwendung von PhAST auf dreidimensionale statt zweidimensionale Molekülrepräsentationen erzeugte nur leichte Änderungen in der beobachteten Anreicherung, wenn auch die erzeugten Ranglisten von einander abwichen.

PhAST wurde erfolgreich in zwei prospektiven Anwendungen eingesetzt. Bei der Suche nach nicht von Nukleosiden abgeleiteten Inhibitoren der bakteriellen Thymidinkinase wurde ein Hit identifiziert. Er zeigte eine deutliche strukturelle Abweichung von der verwendeten Referenzstruktur, war jedoch nur schwach aktiv. In einem Screening nach neuen Modulatoren der γ -Sekretase wurde ein potentes Molekül identifiziert. Es zeigt deutliche Unterschiede zur verwendeten Referenzstruktur. Eine im selben Screening identifizierte inaktive Substanz ermöglichte einen ersten Eindruck der zugehörigen Struktur-Aktivitäts-Beziehung, da es sich lediglich durch den Austausch eines einzigen Atoms von der aktiven Struktur unterschied, jedoch komplett inaktiv war.

PhAST unterscheidet sich von anderen Methoden für das virtuelle Screening durch die Möglichkeit die Signifikanz der berechneten chemischen Ähnlichkeit zu bestimmen, bekannte essentielle Interaktionspunkte höher zu gewichten, solche essentiellen Interaktionsmöglichkeiten zu identifizieren und durch die berechneten Ranglisten von Molekülen. Die gezeigten Beispiele für eine erfolgreiche prospektive Anwendungen haben deutlich gemacht, dass PhAST eine Bereicherung für die Wirkstoffentwicklung ist.

3 - Abstract

This work investigated the applicability of global pairwise sequence alignment to the detection of functional analogues in virtual screening. This variant of sequence comparison was developed for the identification of homologue proteins based on amino acid or nucleotide sequences. Because of the significant differences between biopolymers and small molecules several aspects of this approach for sequence comparison had to be adapted. All proposed concepts were implemented as the ‘Pharmacophore Alignment Search Tool’ (PhAST) and evaluated in retrospective experiments on the COBRA dataset in version 6.1.

The aim to identify functional analogues raised the necessity for identification and classification of functional properties in molecular structures. This was realized by fragment-based atom-typing, where one out of nine functional properties was assigned to each non-hydrogen atom in a structure. These properties were pre-assigned to atoms in the fragments. Whenever a fragment matched a substructure in a molecule, the assigned properties were transferred from fragment atoms to structure atoms. Each functional property was represented by exactly one symbol.

Unlike amino acid or nucleotide sequences, small drug-like molecules contain branches and cycles. This was a major obstacle in the application of sequence alignment to virtual screening, since this technique can only be applied to linear sequences of symbols. As a consequence, molecules and their properties had to be encoded as linear representations. To ensure the detection of identical molecules and close analogues, these representations had to be unambiguous, meaning that one molecule can only be encoded to exactly one sequence. This problem was solved by canonical vertex labeling, where an index is assigned to each vertex in a molecular graph, and the assignment of indices to vertices is identical each time the same molecular graph is handled. This canonical set of indices defines the order of vertices in the linear representation of molecules. Several algorithms for canonical vertex labeling were investigated. They belonged to two classes: Algorithms developed for canonical atom labeling and techniques for dimensionality reduction. To the best of knowledge, this work represents the first application of dimensionality reduction to graph linearization.

Sequence alignment relies on a scoring system that rates symbol equivalences (matches) and differences (mismatches) based on functional properties that correspond to rated symbols. Existing scoring schemes are applicable only to amino acids and nucleotides. In this work, scoring schemes for functional properties in drug-like molecules were developed

based on property frequencies and isofunctionality judged from chemical experience, pairwise sequence alignments, pairwise kernel-based assignments and stochastic optimization. The scoring system based on property frequencies and isofunctionality proved to be the most powerful (measured in enrichment capability). All developed scoring systems performed superior compared to simple scoring approaches that rate matches and mismatches uniformly. The frameworks proposed for score calculations can be used to guide modifications to the atom-typing in promising directions.

The scoring system was further modified to allow for emphasis on particular symbols in a sequence. It was proven that the application of weights to symbols that correspond to key interaction points important to receptor-ligand-interaction significantly improves screening capabilities of PhAST. It was demonstrated that the systematic application of weights to all sequence positions in retrospective experiments can be used for pharmacophore elucidation. A scoring system based on structural instead of functional similarity was investigated and found to be suitable for similarity searches in shape-constrained datasets.

Three methods for similarity assessment based on alignments were evaluated: Sequence identity, alignment score and significance. PhAST achieved significantly higher enrichment with alignment scores compared to sequence identity. p-values as significance estimates were calculated in a combination of Markov Chain Monte Carlo Simulation and Importance Sampling. p-values were adapted to library size in a Bonferroni correction, yielding E-values. A significance threshold of an E-value of $1 \cdot 10^{-5}$ was proposed for the application in prospective screenings.

PhAST was compared to state-of-the-art methods for virtual screening. The unweighted version was shown to exhibit comparable enrichment capabilities. Compound rankings obtained with PhAST were proven to be complementary to those of other methods. The application to three-dimensional instead of two-dimensional molecular representations resulted in altered compound rankings without increased enrichment.

PhAST was employed in two prospective applications. A screening for non-nucleoside analogue inhibitors of bacterial thymidin kinase yielded a hit with a distinct structural framework but only weak activity. The search for drugs not member of the NSAID (non-steroidal anti-inflammatory drug) class as modulators of γ -secretase resulted in a potent modulator with clear structural distinction from the reference compound.

The calculation of significance estimates, emphasizing on key interactions, the pharmacophore elucidation capabilities and the unique compound rankings set PhAST apart from other screening techniques.

4 - Introduction

4.1 - The Drug Development Process

Discovery and development of new drugs is a lengthy and cost-intensive process: Analysis of drug design campaigns leading to approved drugs between 1989 and 2002 resulted in estimated costs (measured in time and money) of 15 years and up to 2 billion dollar per successful campaign depending on therapy and developing firm.^{1,3} The first step in target-based drug discovery (Figure 1) is the identification and validation of a drug target and the ascertainment of its role in the disease process. After assays, which are capable of measuring activity modulating effects of proposed small organic molecules, are developed, the next challenge is the identification of 'hits': non-promiscuous binding compounds with known structure that exhibit reproducible activity above a certain threshold value.^{4,5} If their activity and selectivity is confirmed and they exhibit novel pharmacological features, they are optimized to 'leads' with respect to pharmacodynamics and pharmacokinetics. At this stage, compounds with unwanted groups responsible for fast metabolization or toxicity are weeded out. Each remaining lead is subject to further development into a lead series: compounds exhibiting the same molecular frame ('scaffold') coupled with variations in one or several positions. These are further optimized regarding their activity, bioavailability, toxicity, metabolization and off-target activity. Pre-clinical development involves in vitro and in vivo tests. The conducted studies test for effectiveness and especially safety for further testing in humans. The following clinical trials are separated into three steps: Phase I is an initial testing

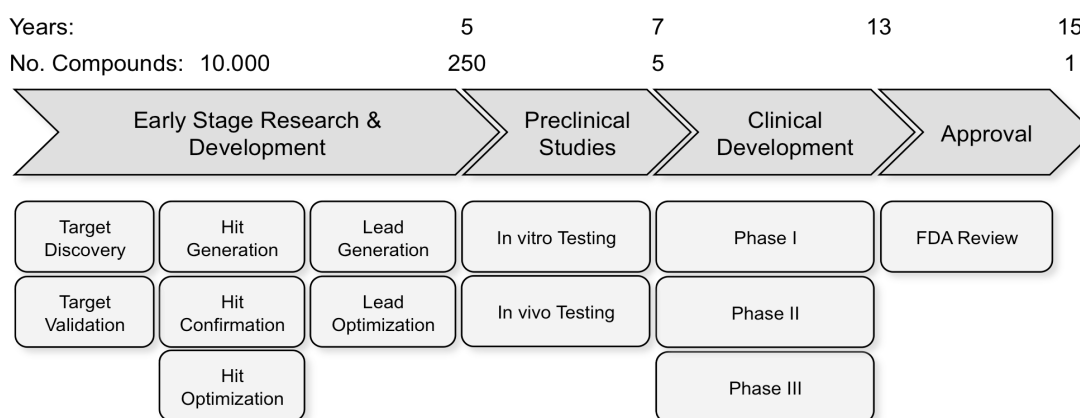


Figure 1. The drug development process. Development time and number of compounds according to diMasi *et al.* 2003,¹ Rankovic & Morphy 2010,² and Adams & Brantner 2006.³

on up to 100 healthy volunteers. The main goal of this clinical study is a first assessment of safety in humans and determination of safe dosing ranges. Phase 2 trials involve up to 500 patients and investigate the candidate drug's effectiveness. They as well examine short-term side effects. In phase 3 drug candidates are studied in a larger number of patients (up to 5,000). These trials generate statistically significant data regarding efficacy and safety. If a drug candidate completes all clinical trials, the developing company files a new drug application with the Food and Drug Administration (FDA). There, the complete data generated during the drug development process is reviewed. If the FDA concludes the drug is safe and effective enough, it is approved. After approval, the production process has to be up-scaled for large-scale manufacturing before the drug can be marketed.

The identification of hits and leads is a major milestone in drug development, since by lack of active compounds every drug discovery campaign is on hold. The identification of a large number of diverse hits and leads is essential. According to the Pharmaceutical Research and Manufacturers of America, 10,000 hits are necessary on average to get one drug to the market.⁶

4.2 - From High-Throughput Screening to Virtual Screening

Since the early 1990s 'High Throughput Screening' (HTS) has dominated hit generation by systematically testing compound libraries containing between 50,000 and 100,000 molecules in automated systems. Combinatorial chemistry helped maximizing library size by taking advantage of miniaturization and parallel synthesis.⁷ The systematic combination of building blocks allows the generation of more than 100,000 compounds within several months.⁸ So far, 10^7 small organic compounds have been synthesized by man or were encountered in nature.⁹ But even cautious estimates of the total number of synthesizable organic molecules (also known as 'chemical space') exceed values of 10^{60} .¹⁰ These estimations are heavily constrained, only considering molecules

- with up to 30 non-hydrogen atoms
- built solely from the elements carbon, nitrogen, oxygen and sulfur
- containing a maximum of four rings
- containing up to 10 branch points

Collections of known bioactive compounds contain structures with more than 30 non-hydrogen atoms: Figure 2 displays the distribution of non-hydrogen atom counts in the

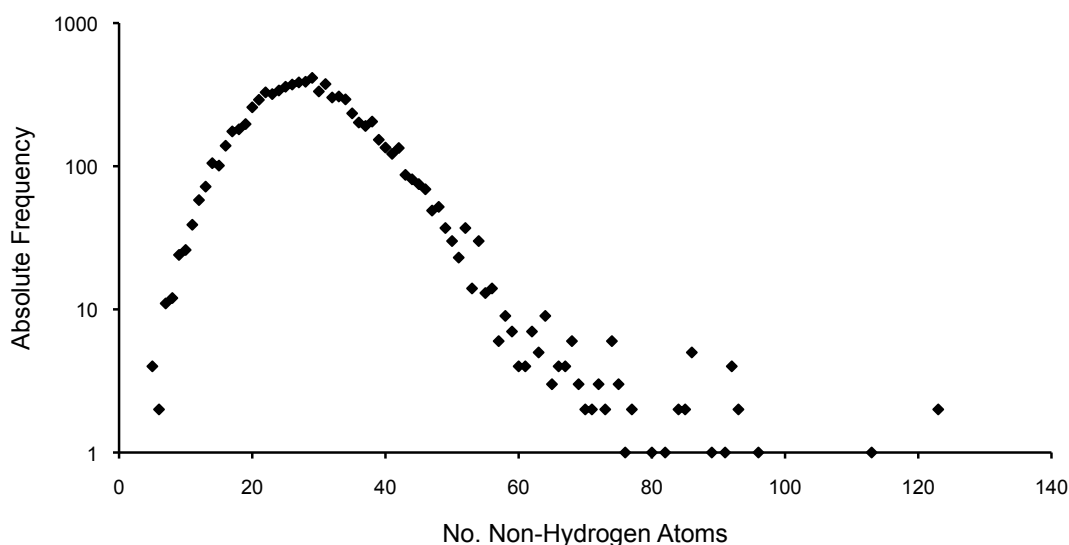


Figure 2. Distribution of the number of non-hydrogen atoms per structure in the COBRA collection of bioactive reference compounds (version 6.1, 8,311 compounds). Logarithmic Y-axis for better visualization of low absolute frequencies.

COBRA¹¹ collection of bioactive reference compounds. As a consequence, the true size of chemical space has to be even higher. Implied by these numbers, it remains unexplored to a vast majority.

HTS is a suitable method for evaluating existing compounds that have been synthesized and stored. But it is limited to this repository due to the fact that it relies on the availability of the actual compound. That way, explorations in chemical space by HTS have to be preceded by costly and time-consuming syntheses. Speed and cost could be optimized, if non-promising candidates were identified and excluded (‘negative design’) from a set of possible structures, or if efforts could be focused on the most promising candidates (‘positive design’). Advances in computer sciences lead to the emerging field of chemoinformatics: “*The application of informatics methods to solve chemical problems*”.¹² It combines aspects of computer sciences, chemistry, biology, medicine and pharmacology. These methods allow for the computer-based evaluation of chemical compounds with regard to various properties. They are based on virtual libraries with the advantages of being independent of synthesized compounds and tremendously faster evaluation compared to HTS, leading to methods for compound prioritization known as ‘Virtual Screening’ (VS). VS has been described in literature as “*the computational equivalent of high-throughput screening, wherein a large number of samples are quickly assayed to discriminate active samples from inactive samples*”.¹³ In the beginning, computational methods were used for negative design.¹⁴ This,

Table 1. Properties of drug-like¹⁴ and lead-like¹⁸ molecules. Hydrogen-bond donors: nitrogen or oxygen atoms with one or more hydrogen atoms, Hydrogen-bond acceptors: nitrogen or oxygen atoms. The original description of the rule of 5 for drug-like compounds does not restrain the number of rotatable bonds.

Property	Drug-Like Compounds	Lead-Like Compounds
Molecular Mass [Dalton]	< 500	< 300
No. Hydrogen-Bond Donors	< 5	< 3
No. Hydrogen-Bond Acceptors	< 10	< 3
Octanol-Water Partition Coefficient (logP)	< 5	< 3
No. Rotatable Bonds	-	< 3

for example, was done by eliminating molecules that judged by their properties appeared non ‘drug-like’. An example for such a set of criteria is Lipinski’s ‘rule of five’,¹⁵ deduced from statistics of known drugs. It describes constraints for molecules with high probability for sufficient oral bioavailability. Properties are listed in Table 1. These guidelines have been updated ever since their original proposition resulting in a larger number of constraints.¹⁶ As the rule of five was compiled based on analysis of properties observed in drugs, it might not be ideally suited for the filtering of promising candidates at the beginning of the drug design process.¹⁷ As a result, a set of criteria for ‘lead-likeness’ was proposed.¹⁸ The corresponding compound properties are also listed in Table 1. Virtual screening methods have evolved and are now also used for the selection of promising candidates (positive design). Based on the origin of the starting point, virtual screening methods can be distinguished in two concepts: structure-based and ligand-based methods.

Structure-based (also: receptor-based¹⁹) virtual screening (SBVS) relies on an available model of the target obtained by X-ray crystallography or nuclear magnetic resonance spectroscopy. The most prominent technique in SBVS is ‘docking’ that aims at predicting the most likely binding pose and the corresponding binding energy. But the available model might display low resolution or there might actually be no model due to induced-fit effects or protein size. In these cases, a homology model for a protein can be built with a close homolog as template, and this model will be treated as the actual receptor structure. In the absence of an actual structure or appropriate template as well as in parallel to structure-based approaches, ligand-based virtual screening (LBVS) is a promising alternative. This screening concept relies on the availability of a known active compound as reference (‘query’). Screening compounds are ranked according to their calculated similarity to the query based on the expectation that compounds with activity against the same target are

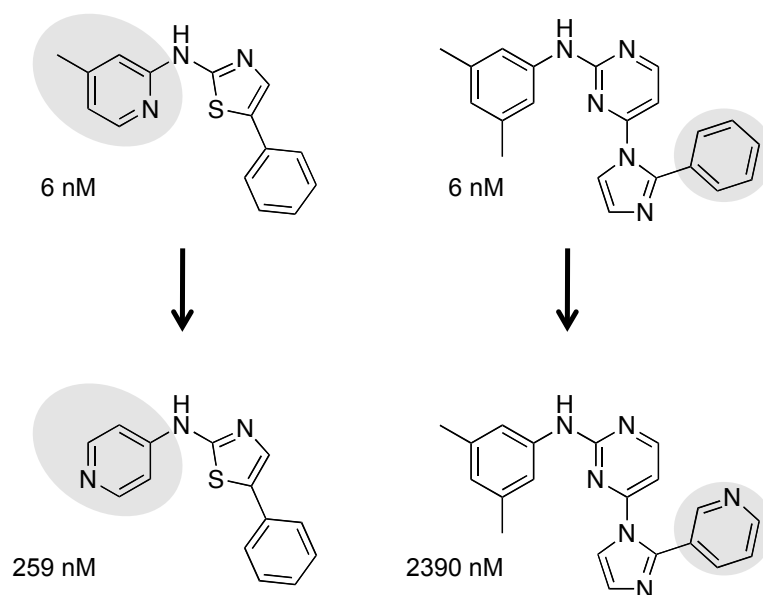


Figure 3. Activity cliffs. Shown are four vascular endothelial growth-factor receptor (VEGFR-2) tyrosine kinase inhibitors with different structures and potencies. The two inhibitors at the top are potent and bind with IC_{50} values of 6 nM. Slight structural modifications result in a decrease in activity by two to three orders of magnitude. Example adapted from Eckert *et al.* 2007.²²

enriched on early ranks. These methods are based on the ‘similar property principle’ (SPP), which assumes a close relationship between structure and activity. It states that similar compounds exhibit similar properties.²⁰ The SPP describes the ideal case, where small modifications to molecules lead to slight alterations in biological activity. It does not account for so called ‘activity cliffs’,²¹ where small modifications cause drastic activity changes (examples in Figure 3). Such effects may occur as a consequence of small modification eliminating essential interactions or other constraints such as the necessity to coordinate metal ions.²³ Methods have been proposed for the characterization of structure-activity relationships in order to assess the validity of the SPP for a particular target.^{23,24} Ligand-based methods can even be applied in the absence of active compounds. In this case, actives for close homologs of the target can be used as queries. This strategy started under the term ‘ligand transfer’ and is now fully exploited as ‘chemogenomics’.²⁵ The idea behind these approaches is that proteins of the same family most likely exhibit similarities in the overall structure, especially concerning the binding pocket. Therefore corresponding ligands should likewise display structural and functional similarities. Successful applications to protein kinases and G-protein-coupled receptors have been reported.²⁶

Even for computational methods, the prediction of activity values for each molecule in chemical space or their classification into sets of active and inactives for a particular target is not feasible. But these techniques can help to navigate in chemical space, *i.e.* guiding drug design campaigns to interesting regions and limiting the number of compound tested in biological assays only to the most promising candidates. This way these techniques help to reduce development costs for new drugs and speed up the drug design process: It was shown that the combination of VS and HTS reduced the number of compounds that had to be tested to find an active to 100 – 1000 instead of $10^4 - 10^6$ with HTS alone.²⁷

4.3 - Chemical Similarity

The key element of ligand-based virtual screening is the assessment of chemical similarity that is more an abstract concept than a calculable property.²⁵ Methods for similarity assessment are based on calculations performed on the ‘molecular graph‘ that is defined as a “*connected undirected graph one-to-one corresponded to the structural formula of a chemical compound so that vertices of the graph correspond to atoms of the molecule and edges of the graph correspond to chemical bonds between them*”.²⁸ In this definition, vertices are labeled with an element symbol derived from the periodic table, edges are labeled with an integer indicating their bond order. The molecular graph can be used for direct similarity assessment through substructure searching. Given a set of active molecules, their maximal common subgraph (MCS) can be calculated and used as query in substructure searches.²⁹ Methods for the identification of a disconnected MCS (maximum overlapping set, MOS) identify a set of substructures common to all known actives. The similarity of structures matching the MCS or MOS usually also depends on parts of the screening compound not matching the query and is located in a value range of 0 (not similar) to 1 (exact match).³⁰

The molecular graph can also be used for the calculation of ‘descriptors’, where a descriptor is “*the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*”.³¹ Descriptors can be distinguished by the dimensionality of the molecular representation that they are calculated from. Table 2 represents a short overview on these differences. ‘Fingerprints’ are the combination of several descriptors in a vector. As bitstrings they can indicate the absence and presence of certain features. Holographic fingerprints on the other hand count the occurrences of certain substructures and / or properties. Fingerprints can be compared using several distance metrics

Table 2. Classification of descriptors by the dimensionality of their molecular representation. Adapted from Gasteiger & Engel 2003.³²

Dimension	Description	Examples
0	Properties independent from atom connectivity and spatial arrangement	Atom Counts, bond counts, molecular mass, sum of van der Waals volumes
1	Properties depending on local atom neighborhoods	Fragment counts like primary, secondary, tertiary and quaternary sp ³ hybridized carbon atoms
2	Properties depending on atom connectivity but invariant to spatial arrangement	Zagreb index, Wiener path index, molecule radius, molecule diameter
3	Properties depending on the spatial arrangement of atoms	Radius of gyration, solvent-accessible surface volume

(Manhattan distance, Euclidean distance) or similarity measures (Tanimoto coefficient, Dice coefficient, Cosine coefficient), of which many can be applied to binary as well as holographic fingerprints.³³

An abstraction from the molecular graph that is used for similarity assessment as well is the ‘pharmacophore’. The term was first used by Lemont Kier³⁴ and according to the International Union of Pure and Applied Chemistry (IUPAC) is “*the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response*”.³⁵ It is a purely abstract concept and does not describe a real molecule. The pharmacophore can be considered as the set of interactions in the correct spatial configuration that is necessary for the activity of a molecule. As a consequence of this definition, not all functional groups and the corresponding interaction possibilities present in a particular molecule might be part of the pharmacophore. Those are named ‘potential pharmacophoric points’ (PPPs) because it is unknown *a priori* which of them actually contribute to the ligand-receptor interaction.³⁶ Typical interaction types considered potential pharmacophoric points are hydrogen-bond donors, hydrogen-bond acceptors, positive and negative charges and lipophilic as well as aromatic features. The first interaction pattern published under the term pharmacophore is shown in Figure 4.

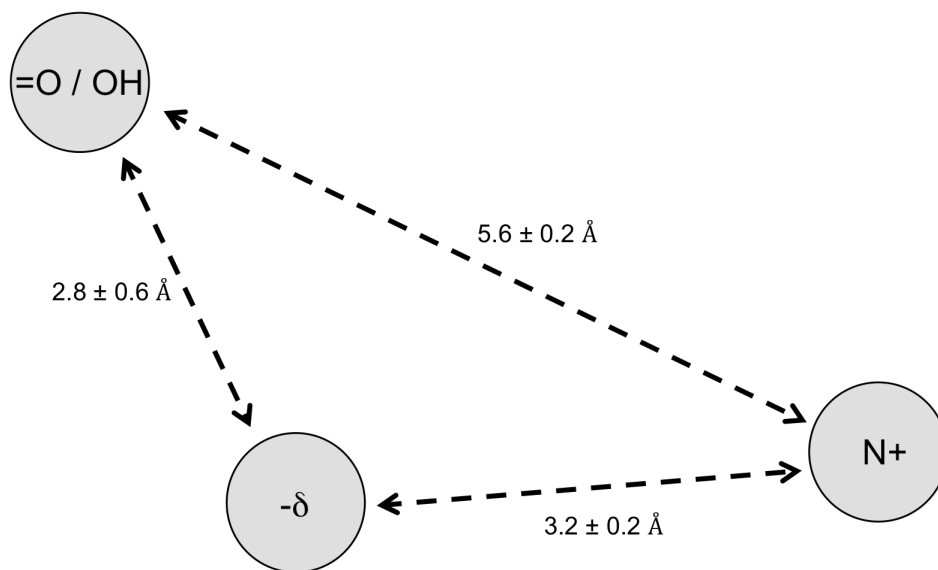


Figure 4. The first published pharmacophore model, by L. B. Kier, for muscarinic agonists. Adapted from Kier 1971.³⁴

An abstraction describing a molecule employing these six interaction types or combinations between them is a blurred characterization, as different functional groups can be responsible for the same interaction possibilities. Hence, molecules with different structures can exhibit the same pattern of potential pharmacophoric points. This potential for ‘Scaffold Hopps’ is extremely valuable in the hit identification phase because it generates a more diverse set of unrelated starting points compared to methods measuring structural similarity. The same concepts of descriptors and fingerprints calculated from the molecular graph can be applied to the pattern of potential pharmacophoric points as well. Several successful applications of virtual screening methods have been reported.^{37,38}

4.4 - Line Notations

Besides descriptors and fingerprints, the molecular graph can be described as an alphanumeric sequence. Compared to descriptors and fingerprints they do not only describe the presence and absence of structural features but can also mirror their topological or spatial arrangement if the respective property influenced the linearization procedure. An example every chemist is familiar with is the systematic IUPAC name of a compound. But several other line notations were developed for compound storage, substructure search and duplicate detection.

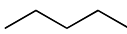

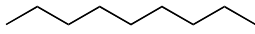



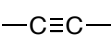





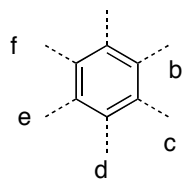

4.4.1 - Wiswesser Line-Formula Notation

A first version of the ‘Wiswesser Line-Formula Notation’ (WLN) was published in 1954.³⁹ The WLN assigns symbols to atoms or substructures of a molecular graph. For many atoms the WLN symbol is equal to the atomic symbol in the periodic table. Functional groups, ring systems, positions of ring substituents and positions of condensed rings are assigned to symbols or combinations of symbols. Definitions are chosen in a way that ensures frequently met substructures are encoded with only one symbol to keep linear representations short.⁴⁰ This way the WLN facilitates searches for particular substructures and functional groups. Symbols used for WLN coding of chemical structures are listed in Table 3.⁴¹ Examples for

Table 3. Symbols used for WLN coding of chemical structures. Symbols are listed with increasing priority used for unambiguity.

Symbol	Usage
Capital letters A-Z	Elements, substructures, branches, bonds, ring positions
Numbers 0-9	Length of alkyl chains, ring number
“&”, “/”, “-”, “ “	Rings and substitution positions

Table 4. Examples for substructure codes used in WLN.

Structure	WLN	Structure	WLN
	5		Q
	9		O
	U		V
	UU		VH
	X		VQ
	Y		VO
	R substituent positions B C D E F		N

coded fragments are shown in Table 4. Starting from different atoms in the molecular graph, many representations of the same molecule in WLN are possible. Unambiguity of WLN representations is achieved through a simple rule-based prioritization of possible starting points. Coding begins at the element with highest priority based on prioritization order shown in Table 3.⁴¹ After the starting point is selected, the WLN notation is determined by the topology of the molecular graph. Linear representations of molecules in WLN are very compact, since complete substructures are condensed to only one symbol. Examples of WLN usage are the application to indexing the Chemical Substructure Index (CSI) at the Institute for Scientific Information and the CROSSBOW (Computer Retrieval of Organic SubStructures by means of Wiswesser) System of Imperial Chemical Industries.^{42,43} In 1982 an advanced version of the WLN (AWLN) was published that utilizes more than the 40 WLN symbols and has an extended rule set.⁴¹

4.4.2 - Representation of Organic Structures Description Arranged Linearly

The ‘Representation of Organic Structures Description Arranged Linearly’ (ROSDAL) syntax was developed at the Beilstein institute in 1985 for the Beilstein DIALOG system.⁴⁴ The ROSDAL generation process is straightforward: Integer numbers beginning from 1 are assigned to all non-hydrogen atoms randomly, and paths through the molecular graph are written in linear order. Atoms are written as their index followed by the element symbol. Carbon atoms are an exception to this rule. They are represented just by digits because of their high frequency in drug-like compounds. Bonds are represented by symbols corresponding to their bond order. Symbols used for ROSDAL coding of chemical structures are listed in Table 5. ROSDAL is used as data exchange format in the Beilstein DIALOG-system.

Table 5. Symbols used for ROSDAL coding of chemical structures.

Symbol	Usage
Integer numbers	Indicate paths through the labelled graph
Element symbols	Element types, carbon atoms are only referenced to by their index
- / = / # / ?	Single / double / triple / any bond
,	Delimiter for sequences generated from separated branches

4.4.3 - Simplified Molecular Input Line Entry System

David Weininger developed the ‘Simplified Molecular Input Line Entry System’ (SMILES) while working at the United States Environmental Research Laboratory in 1986.⁴⁵ SMILES are human understandable and very compact. A molecule graph is transformed into a line notation following six simple rules:

- Hydrogen atoms are omitted, they automatically saturate free valences
- Atoms are represented by their corresponding atomic symbols
- Neighboring atoms in the molecular graph stand next to each other in the line notation
- Single bonds are represented by “-“, double bonds by “=” and triple bonds by “#”
- Branches are represented by parentheses
- Rings are indicated by identical digits following the element symbol of the atoms closing the ring

SMILES generation can start at any vertex in the molecular graph, resulting in a large number of possible valid SMILES describing the same molecule. An unambiguous line notation named ‘canonical SMILES’ can be created using a two-step algorithm proposed under the name CANGEN:⁴⁶ First, canonical labels are assigned to the vertices in the molecule graph. The second step is a depth-first search visiting the vertices with low indices with highest priority that concatenates symbols of atoms and bonds and inserts symbols for branching and ring closures. The SMILES line notation was subject to several extensions designed for special purposes, such as substructure description and reaction notation.

4.4.4 - IUPAC International Chemical Identifier

The ‘IUPAC International Chemical Identifier’ (InChi, originally ‘IChI’ for ‘IUPAC Chemical Identifier’) was developed from 2000 to 2005 as a project of the IUPAC chemical nomenclature and structure representation division and the National Institute of Standards and Technology (NIST).⁴⁷ The objective of the project was the development of a non-proprietary identifier for chemical substances than can be used in print media and electronic data sources.

Every InChi starts with the fragment “InChi=” followed by the version number (currently 1). Structural information is organized in six layers and sub-layers, describing different aspects of a molecule.⁴⁸ InChi layers are listed and described in Table 6. The InChi generation process has three steps:

- Normalization: removes redundant information, disconnects salts and metals, eliminates radicals if possible
- Canonization: creates canonical labels for atoms, ensures unambiguity
- Serialization: generates the actual InChi string

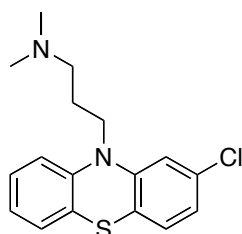
A special form of line notation as structure representation is InChiKey. It is a condensed version created from InChi through hashing using the Secure Hash Algorithm (SHA-256).^{48,49} InChiKey has a fixed length of 27 characters: The first 14 symbols result

Table 6. InChi layers, their identification characters and meaning.

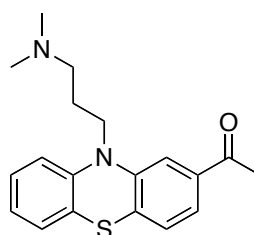
Layer	Sublayer Symbol	Sublayer	Description
Main	/	chemical formula	Specifies bonds separately for non-hydrogen and hydrogen-atoms
	/c	connectivity	
	/h	hydrogen	
Charge	/q	charge	Specifies absolute charge and protonation alterations necessary for representation without regard to protonation
	/p	proton balance	
Stereochemical	/b	double bond	Specifies E/Z- and tetrahedral stereochemical properties of a molecule
	/t	sp3	
	/m	sp3	
	/s	sp3	
Isotopic	/i	isotopic	Specifies aberrations from the majoritarian isotopes
Fixed-H	/f	fixed-H	Hydrogen-atoms mobile due to tautomerism can be bound to specific atoms in the original structure molecule; if these changes affect earlier layers, appropriate changes are added to this layer
Reconnected	/r	reconnected	Used for the handling of organometallic compounds; represents such as one large structure instead of two individual components; if these changes affect earlier layers, appropriate changes are added to this layer

from a hash of the connectivity information, followed by a hyphen and 8 characters representing the remaining layers (except charge), 1 for the InChi version and a checksum character. Separated by a hyphen the last character describing the protonation layer. They were developed to facilitate easy searching. InChi and InChiKey are currently used by several public and commercial databases (for example the Pubchem project, the United States National Cancer Institute Database and the Chemical Entities of Biological Interest database of the European Bioinformatics Institute) as well as scientific journals like Nature Chemical Biology and the Beilstein Journal of Organic Chemistry.

Examples for the described line notations are shown in Figure 5.



Trivial Name: Chlorpromazine
 IUPAC Name: 3-(2-chlorophenothiazin-10-yl)-N,N-dimethylpropan-1-amine
 WLN: T C666 BN ISJ B3N1&1 EG
 ROSDAL: 1-2N-3-4-5-6N-7=8-9=10-11=12-13S-14-15=16-17=18-19-6, 7-12, 14=19, 2-20, 17-21Cl
 SMILES: CN(C)CCCN1C2=CC=CC=C2SC3=C1C=C(C=C3)Cl
 InChi: InChI=1S/C17H19ClN2S/c1-19(2)10-5-11-20-14-6-3-4-7-16(14)21-17-9-8-13(18)12-15(17)20/h3-4,6-9,12H,5,10-11H2,1-2H3
 InChiKey: ZPEIMTDSQAKGNT-UHFFFAOYSA-N



Trivial Name: Acetylpromazine
 IUPAC Name: 1-[10-[3-(dimethylamino)propyl]phenothiazin-2-yl]ethanone
 WLN: T C666 BN ISJ B3N1&1 EV1
 ROSDAL: 1-2N-3-4-5-6N-7=8-9=10-11=12-13S-14-15=16-17=18-19-6, 7-12, 14=19, 2-20,17-21-22, 21=23O
 SMILES: CC(=O)C1=CC2=C(C=C1)SC3=CC=CC=C3N2CCCN(C)C
 InChi: InChI=1S/C19H22N2OS/c1-14(22)15-9-10-19-17(13-15)21(12-6-11-20(2)3)16-7-4-5-8-18(16)23-19/h4-5,7-10,13H,6,11-12H2,1-3H3
 InChiKey: NOSIYYJFMPDDSA-UHFFFAOYSA-N

Figure 5. Structure diagrams and corresponding line notations of chlorpromazine and acetylpromazine. IUPAC name generated by ChemBioDraw Ultra (v12.0, CambridgeSoft, Cambridge, USA), SMILES generated by MOE (Molecular Operating Environment v2009.10, Chemical Computing Group, Montreal, QC, Canada), InChi and InChiKey generated by IUPAC InChi generator v1.02.

4.5 - Virtual Screening employing Line Notations

None of the described line notations was developed for virtual screening. That is why most of them are not suitable for this purpose.

ROSDAL has the major disadvantage that it is not unambiguous. The same molecule can be represented by several ROSDAL sequences without any string similarity measure being able to detect even identical structures. That makes it unsuitable for virtual screening, as even identical molecules could not be recognized as such.

WLN is unambiguous, but the sequence generation process is complicated. Computer programs for the automated input and output of molecules in WLN were developed,^{50,-52} but the full set of rules could not be implemented and the automated sequence generation process was prone to errors.³² Furthermore the substructure encoding system of WLN has the disadvantage that insertions and deletions of one vertex to the molecular graph are not necessarily reflected by additions or deletions of only one symbol in WLN notation. As a consequence, there is no one-to-one correspondence in the severity of molecule differences between their molecular graph and WLN sequence. This is a clear disadvantage for a virtual screening method.

Due to its layered structure, there are different possible InChi representations of the same structure despite the fact that InChi is unambiguous. And depending on the way a molecule is drawn, in some cases the generation of certain layers is not possible, for example the stereo layer from a structure diagram without specified stereochemistry. InChiKey is a hashed version of InChi with fixed length. Different structures yield different InChiKey representations. But comparing structures by their InChiKey strings is unreasonable, as small modifications cause drastic differences at least in the first part of InChiKey because of the hash operation. The only meaningful result from the comparison of InChiKey representations of molecules is the identification of identical molecules.

SMILES are an unambiguous description of the molecular graph. Except for changes in branching and ring closures they mirror modifications to the molecular graph by the addition or deletion of the same number of symbols as non-hydrogen atoms are inserted or deleted in the graph. These properties make them the best choice of the described line notations for the development of virtual screening methods. But as well as the other line notations, SMILES is a description of the molecular graph, not of its interaction possibilities. The following sections describe the two known virtual screening approaches based on SMILES.

4.5.1 - LINGO

LINGO is based on the comparison of absolute word frequencies calculated from SMILES representations of molecules.^{53,54} The term ‘LINGO’ refers to a SMILES substring of length q . To compare two molecules M_1 and M_2 with lengths m_1 and m_2 their corresponding canonical SMILES are generated and the following preprocessing steps are applied to ensure each feature of the molecule is represented by only one symbol: Cl is altered to L, Br is altered to R, all numbers indicating ring closures are replaced by 0. Then, all $(m_1 - (q - 1))$ substrings of M_1 and $(m_2 - (q - 1))$ substrings of M_2 in modified SMILES representation are collected and condensed to a unique set of size l while counting the occurrences of each unique LINGO in each SMILES. For similarity assessment LINGO frequencies are used to calculate the integral Tanimoto coefficient of the form (Equation 1)

$$T_C = \frac{\sum_{i=1}^l 1 - \frac{|N_{M_1,i} - N_{M_2,i}|}{N_{M_1,i} + N_{M_2,i}}}{l} \quad (1)$$

where $N_{M_1,i}$ is the frequency of LINGO i in M_1 and $N_{M_2,i}$ is the frequency of LINGO i in M_2 . Calculated similarities are bound between 0 and 1. Besides virtual screening,⁵⁴ LINGO has been successfully applied to the calculation of biophysical properties.⁵³

4.5.2 - Comparison by Compression

The Kolmogorov complexity of a sequence X ($K(X)$), is defined as the shortest binary program that computes X on a computer.⁵⁵ The conditional Kolmogorov complexity $K(X|Y)$ is the shortest binary program that computes X from Y . These definitions can be used for the calculation of a distance between X and Y , the ‘Normalized Information Distance’ (NID) defined in Equation 2.⁵⁶ It returns values between 0 and 1.

$$NID(X,Y) = \frac{\max\{K(X|Y), K(Y|X)\}}{\max\{K(X), K(Y)\}} \quad (2)$$

The Kolmogorov complexity is noncomputable. But it can be approximated by the size of compressed representations $C(X)$ and $C(Y)$ of X and Y . The DEFLATE algorithm (a

combination of LZ77 compression⁵⁷ and Huffman coding⁵⁸) has been successfully used in a virtual screening approach to compress SMILES representations of molecules and calculate their similarity as given in Equation 3⁵⁹

$$S(X,Y) = 1 - \frac{\min\{C(XY),C(YX)\} - \min\{C(X),C(Y)\}}{\max\{C(X),C(Y)\}} \quad (3)$$

where $C(XY)$ ($C(YX)$) is the size of the compressed representation of X and Y (Y and X) concatenated. The only necessary preprocessing step identified was the duplication of SMILES to overcome storage overhead effects of the compression algorithm.

4.5.3 - General String Metrics

There are other string metrics that could be applied to the comparison of line notation representations of molecules. The Levenshtein distance between two sequences is defined as the minimum number of edit operations necessary to transform one sequence into the other with insertion, deletion and substitution of a single symbol being the allowed edit operations.⁶⁰ The Damerau-Levenshtein distance uses an additional edit operation: transpositions of neighboring symbols.⁶¹ Dice's coefficient can be applied to strings based on bigram (substrings of length 2) counts as shown in Equation 4⁶²

$$S(X,Y) = \frac{2n_t}{n_x + n_y} \quad (4)$$

where n_t is the number of bigrams common to both strings, n_x is the number of bigrams found only in X and n_y is the number of bigrams found only in Y .

5 - Study Objective

The described line notations are linear representations of the molecular graph, and as a consequence capture only structural properties. But the hit generation phase of a drug discovery campaign relies on a preferably diverse set of hits representing independent starting points. With ligand-based approaches in virtual screening, these can be generated using pharmacophore methods, whereas techniques comparing molecule structures are more suitable during the development of lead-series. That is why existing line notations are not ideal for the early stage of the drug development process and a way to encode the pattern of potential pharmacophoric points as alphanumeric sequence has to be developed, preferably with a one-to-one correspondence in the numbers of vertices in the molecular graph and symbols in line notation representation.

The mentioned similarity measures applicable to strings only count symbol identities. Not identical symbols are not further distinguished. A line notation representing potential pharmacophoric points as symbols would clearly benefit from a similarity measure that is able to differentiate between several cases of dissimilarity, because the exchange of certain pairs of potential pharmacophoric points is more or less severe than others. A sequence comparison method that is sensitive to different cases of not identical symbols is sequence alignment used in biology and bioinformatics for the identification of homologue amino acid or nucleic acid sequences. So far sequence alignment has been parameterized for such biopolymers. Given the existence of a line notation describing functional properties of molecules, a new parameterization of global pairwise sequence alignment has to be undertaken before it can be used in virtual screening.

5.1 - Pharmacophore Alignment Search Tool (PhAST)

The combination of a line notation describing a linear form of the pattern of potential pharmacophoric points of a molecule and global pairwise sequence alignment as similarity measure between these sequences was developed under the name 'Pharmacophore Alignment Search Tool' (PhAST). It represents molecules as unambiguous sequences of symbols describing their pattern of potential pharmacophoric points (PhAST-sequence), meaning that the program creates exactly the same sequence for the same molecule at every time. But this molecule might not be the only one leading to this sequence. Each symbol in a PhAST-

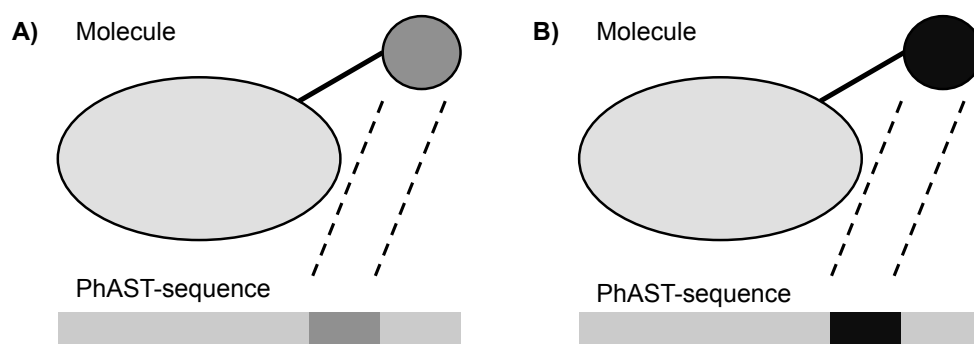


Figure 6. Outline of the ideal correspondence between molecule and line notation. Exchanges of potential pharmacophoric points from A to B (indicated by different shades of grey) only influence the PPP types in the line notation, but not the position where this information is coded.

sequence describes the interaction possibilities of a non-hydrogen atom in the original molecule, thus corresponds to a potential pharmacophoric point. The ideal correspondence between the pattern of PPPs and the PhAST-sequence of a molecule is illustrated in Figure 6: The position of each symbol in the sequence should only depend on graph topology, not on PPP type. That way molecules with topologically identical PPP patterns diverging only in one PPP type yield identical PhAST-sequences except for the symbols representing the diverging PPP types.

The textual representation of a molecule is created in three steps:

- 1) **Categorization:** A graph of potential pharmacophoric points is created, in which each non-hydrogen atom of the molecular graph is represented by a vertex. Each vertex is colored with a symbol describing the possible interaction of the original atom. This atom-typing is fragment-based, employing a set of substructures with pre-defined assignments of potential pharmacophoric points to non-hydrogen atoms. Substructure searches are carried out using the Molecular Query Language (MQL).⁶³ Types of potential pharmacophoric points used in PhAST are listed in Table 7. MQL queries as representations of molecular fragments used for atom-typing and their corresponding assignments of potential pharmacophoric points are shown in Table 8.
- 2) **Canonization:** To ensure unambiguity of PhAST-sequences, vertices in the graph of potential pharmacophoric points have to get assigned a unique set of indices (called canonical labels) in the integer range from 1 to n , where n equals the number of vertices in the graph.

Table 7. Potential pharmacophoric points employed in PhAST and their corresponding symbols used in the line notation.

Possible Interactions	Symbol
hydrogen bond acceptor	A
charge positive	P
charge negative	N
lipophilic	L
aromatic	R
hydrogen bond acceptor, hydrogen bond donor	E
hydrogen bond acceptor, polar	Q
hydrogen bond acceptor, hydrogen bond donor, polar	U
no possible interactions	O

Table 8. MQL⁶³ queries defining pharmacophoric points in PhAST. Symbols are assigned to atoms used in the queries from left to right. Queries are used in the given order from top to bottom.

MQL Query	PPP Symbols
c	R
n	R
*[charge<0]	N
*[charge>0]	P
C(=O)-O-H	O;N;E
P(=O)-O-H	O;N;E
S(=O)-O-H	O;N;E
N[allHydrogens=0&totalConnections=3]	Q
N[allHydrogens=1&totalConnections=3](-C')-C'	U
N[allHydrogens=2&totalConnections=3]-C'	U
N[allHydrogens=1&totalConnections=2]=C'	E
N[allHydrogens=0&totalConnections=2](=C')-C'	A
O-H	E
C=O	O;A
C[!bound(~N)&!bound(~O)]~*[C F Cl Br I S]	L
Cl	L
Br	L
I	L
S[!bound(~N)&!bound(~O)]~*[C H]	L

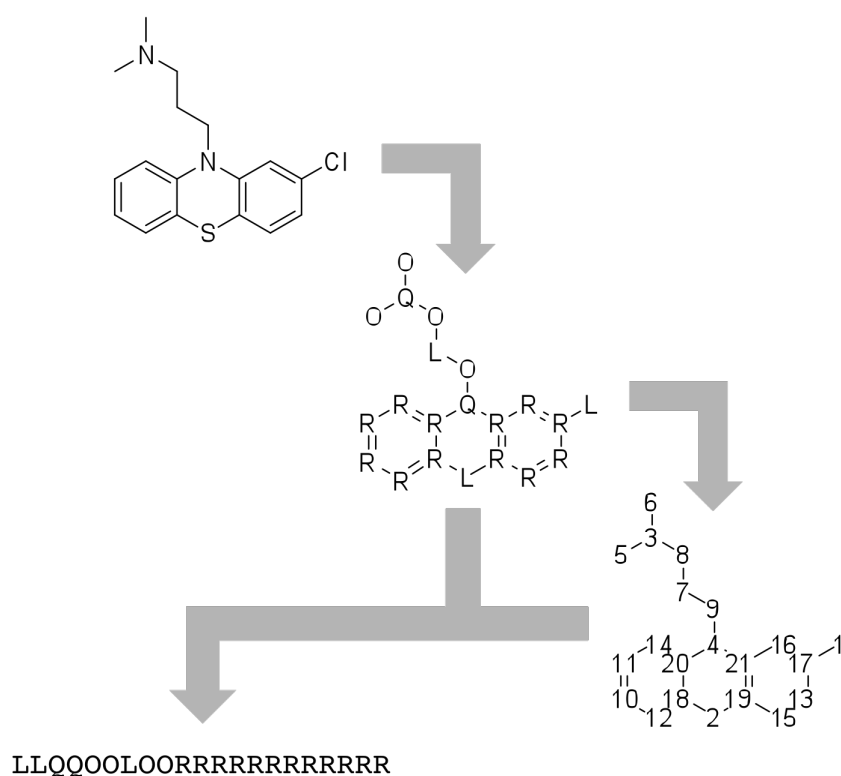


Figure 7. Outline of the sequence generation process of PhAST. After all atoms are typed, vertices in the created graph of potential pharmacophoric points are canonically labeled. The generated indices dictate the order of symbol concatenation.

- 3) Concatenation: To finally create the PhAST-sequence as representation of a molecule, the symbols corresponding to the vertices in the graph of potential pharmacophoric points created in step (1) are concatenated in the order determined by the canonical labels generated in step (2).

The workflow of PhAST-sequence generation is illustrated in Figure 7 using chlorpromazine as example. It is noteworthy that, unlike SMILES, a PhAST-sequence lacks any explicit description of branching and ring closures in a molecule. This information is only implicitly encoded if it was used in the canonization process.

Pairwise sequence alignment was developed as sequence comparison method to answer the question whether two amino acid sequences are related.⁶⁴ To create the alignment of two sequences $X = x_1, x_2, \dots, x_n$ and $Y = y_1, y_2, \dots, y_n$, their symbols are matched. In this, the symbol order is retained and gaps may be inserted to improve the matching (insertion of paired gaps is forbidden). Three cases exist: (i) x_i is aligned to y_j and $x_i = y_j$ (match), (ii) x_i is aligned

to y_j and $x_i \neq y_j$ (mismatch), (iii) x_i is aligned to a gap in Y or y_j is aligned to a gap in X . In protein sequence alignment, matches represent conserved residues; mismatches may arise from mutations, and gaps from insertions or deletions in an assumed evolutionary process of the compared sequences. Consequently, matches are rewarded with a positive score, mismatches are, depending on the exact case, either rewarded with a positive score or penalized with a negative score, and gaps are penalized with a negative score. The score of the complete alignment is calculated as the sum of scores for matches, mismatches and gap penalties. The optimal alignment of two sequences is the alignment with the maximum score. This way, the alignment identifies similar regions of both sequences, and from these similarities the decision can be made whether significant homology exists or whether the observed similarities could have occurred by chance.

There are two types of sequence alignments: global and local. Global alignments span the entire length of both sequences, aligning every symbol in each one with a symbol from the other sequence or a gap. Local alignments on the other hand identify only the most similar region between two sequences that can be highly divergent overall. If the compared sequences are very similar, global and local alignments are identical. PhAST-sequences represent whole molecules. As a consequence, in order to obtain similarity scores for the original molecules, PhAST employs global alignment. The standard technique for this purpose is the Needleman Wunsch algorithm based on dynamic programming.^{64,65} The algorithm depends on a scoring function $s(x_i, y_j)$ that returns scores for symbol matches and mismatches. The original algorithm applicable to any gap cost function has complexity of $O(\max(n, m)^3)$, but for the special case of affine gap penalties there is a simplified algorithm with complexity $O(nm)$ ($\approx O(n^2)$).⁶⁶ Affine means that the opening of a gap is penalized by a gap open penalty d , the extension of a gap by one position with a gap extension penalty e , with $d > e$. The penalty P of a gap with length g can be calculated by Equation 5.

$$P(g) = -(d + (e(g - 1))) \quad (5)$$

The algorithm calculates a two-dimensional matrix F ('alignment graph') with three entries in each cell $F_{i,j}$ representing the score of the optimal alignment of initial subsequences (x_1, \dots, x_i) and (y_1, \dots, y_j) ending with the alignment of x_i and y_j ($F_{i,j}^D$), x_i aligned to a gap position in Y ($F_{i,j}^H$) and y_j aligned to a gap position in X ($F_{i,j}^V$). The algorithm is divided into two phases: Initialization of border cells according to Equations 6 – 12 and the recursive

$$F_{1,1}^D = s(x_1, y_1) \quad (6) \quad F_{i+1,j+1}^D = s(x_{i+1}, y_{j+1}) + \max\{F_{i,j}^D, F_{i,j}^H, F_{i,j}^V\} \quad (13)$$

$$F_{1,j}^D = s(x_1, y_j) - (d + e(j-1)) \quad (7) \quad F_{i+1,j}^H = \max\{F_{i,j}^D - d, F_{i,j}^H - e, F_{i,j}^V - d\} \quad (14)$$

$$F_{i,1}^D = s(x_i, y_1) - (d + e(i-1)) \quad (8) \quad F_{i,j+1}^V = \max\{F_{i,j}^D - d, F_{i,j}^H - d, F_{i,j}^V - e\} \quad (15)$$

$$F_{i,0}^H = -(d + e(i-1)) \quad (9)$$

$$F_{1,j}^H = -(2d + e(j-1)) \quad (10)$$

$$F_{0,j}^V = -(d + e(j-1)) \quad (11)$$

$$F_{i,1}^V = -(2d + e(i-1)) \quad (12)$$

calculation of remaining matrix elements according to Equations 13 –15. After all matrix elements are calculated, the optimal alignment score equals $\max\{F_{n,m}^H, F_{n,m}^D, F_{n,m}^V\}$.

The actual sequence alignment can be assembled in a traceback procedure starting from the traceback step T_0 that equals the alignment possibility of sequence ends x_n and y_m with the maximum score. The next step (T_1) back in the alignment graph towards sequence beginnings is determined by the value in the alignment graph that was used in the maximization for the current value. The traceback procedure with corresponding maximization possibilities is outlined in Table 9. There might be different optimal alignments

Table 9. Sequence alignment traceback. The traceback procedure is explained by the step from T_0 to T_1 . Besides index adjustments, the procedure is the same for all following steps. From each possible T_0 there are three possibilities for T_1 . T_1 is chosen as the possibility corresponding to the maximum of the three possible maxima listed for each T_0 .

T_0	T_1 Possibilities	Necessary Maximum
$F_{n,m}^H$	$F_{n-1,m}^H$	$F_{n-1,m}^H - e$
	$F_{n-1,m}^D$	$F_{n-1,m}^D - d$
	$F_{n-1,m}^V$	$F_{n-1,m}^V - d$
$F_{n,m}^D$	$F_{n-1,m-1}^H$	$F_{n-1,m-1}^H + s(x_n, y_m)$
	$F_{n-1,m-1}^D$	$F_{n-1,m-1}^D + s(x_n, y_m)$
	$F_{n-1,m-1}^V$	$F_{n-1,m-1}^V + s(x_n, y_m)$
$F_{n,m}^V$	$F_{n,m-1}^H$	$F_{n,m-1}^H - d$
	$F_{n,m-1}^D$	$F_{n,m-1}^D - d$
	$F_{n,m-1}^V$	$F_{n,m-1}^V - e$

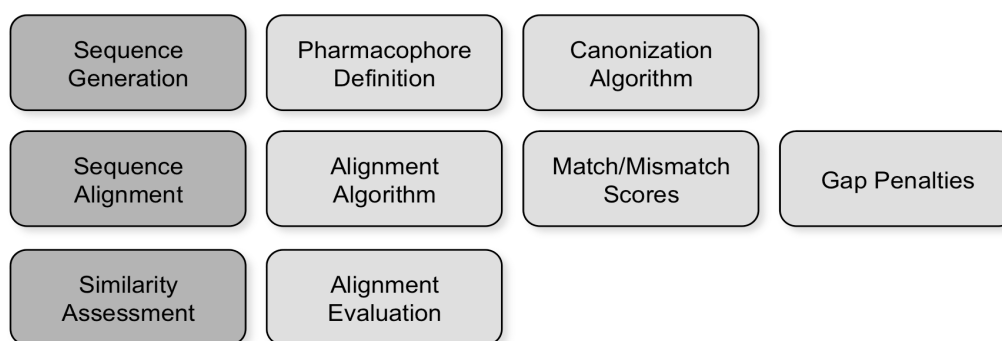


Figure 8. Major steps (dark grey) of PhAST and their variable parameters (light grey).

all yielding maximum score. To ensure that the algorithm always returns the same alignment for the same pair of sequences the three possibilities for the next step are always checked in the order $F_{i,j}^H$, $F_{i,j}^D$ and $F_{i,j}^V$, and a possibility is only accepted as next step if it has higher score. If only the alignment score is of interest, performing the traceback procedure is not necessary.

Sequence alignments were developed for the comparison of amino acid sequences but are applied to nucleotide sequences (deoxyribonucleic acid and ribonucleic acid) as well. With an appropriate scoring scheme, each type of sequence can be compared using this technique. For protein sequence alignments score matrices like PAM⁶⁷ (point accepted mutations) or BLOSUM⁶⁸ (block substitution matrix) are a common choice as scoring systems for matches and mismatches. These matrices were calculated from multiple reference alignments of sequences and the observed substitution frequencies of amino acids. Affine gap penalties described in Equation 5 are the standard choice in most applications of global and local sequence alignments. But other choices are possible, for example:

- constant gap penalty: any gap is penalized with penalty p
- linear gap penalty: analogue to the affine gap penalty with $d = e$
- logarithmic gap penalty: a gap of length g is penalized with $\log(g)$

The steps and variable parameters of PhAST are described in Figure 8.

5.2 - Preliminary Parameterization

In a preliminary study a first parameterization of PhAST was identified that exhibited basic screening capabilities.⁶⁹ In this version the Weininger algorithm for graph canonization⁴⁶ was used for canonical labeling of the graph of potential pharmacophoric points. PhAST-sequences were aligned by the Needleman-Wunsch algorithm.⁶⁴

5.2.1 - Scoring System

The scoring system used for matches and mismatches is shown Table 10. It is based on the idea of entropy scoring, assigning less frequent events higher scores or penalties, that way giving them higher influence on the calculated sequence alignment.⁷⁰ PPP frequencies determined from the COBRA collection of bioactive reference compounds are: A = 4.95%, E = 1.44%, L = 19.95%, N = 1.22%, O = 24.63%, P = 1.8%, Q = 1.58%, R = 41.61%, U = 3.11%. The idea of entropy scoring in PhAST is motivated by the fact that frequent PPP types like R, L and O form main parts of molecules, with their only function being to ensure other PPPs have the correct spatial arrangement. Matches and mismatches between these frequent PPP types could occur in an alignment just because of their overall frequency, not because they represent the same interaction at the same position in different molecules.

PPP types can be separated in two groups: single-interaction and multiple-interaction PPPs. Scores for single-interaction PPPs were chosen to resemble the corresponding frequencies in drug-like molecules determined from the COBRA¹¹ collection of bioactive

Table 10. Scoring scheme for matches and mismatches of potential pharmacophoric points. For a description of the interaction possibilities associated with a symbol see Table 7.

	A	E	L	N	O	P	Q	R	U
A	8	2	2	-1	-2	-4	4	-4	-2
E		12	-4	-9	-4	-6	-4	-9	0
L			2	-2	-2	-2	-4	1	-6
N				10	-2	-6	-7	-4	-10
O					2	-2	-4	-4	-6
P						10	6	-5	4
Q							14	-9	6
R								3	-13
U									16

reference compounds, so that the influence of less frequent PPPs increased. PPP types O and L are the most frequent ones. As a consequence the corresponding matches get only a reward of 2. The less common type A gets a match score of 8. Least frequent types E, N and P receive highest score of 10. These scores were chosen arbitrarily.

Mismatch scores are based on isofunctionality of mismatching types. P and N represent opposite interaction possibilities and this mismatch is scored with -6, just like A and D. Any mismatch involving O represents a loss of function and is scored with -2, so are mismatches of any type with L. N can partly act the same way A does, so this mismatch receives minimal penalty of -1. D on the other hand as opposite of A is scores with a higher penalty of -3 for the same reason. As A is negatively polarized, the mismatch with P is penalized with -4 – stronger than a mismatch with O, but weaker than the PN event. Aligning P to D is penalized by -2.

Events involving multifunctional PPPs are scored following a straightforward scheme: All possible unique pairs of single-functional PPPs included in the multifunctional atom types are considered and their scores are added up to the final match or mismatch score. A special case is type R, as this PPP represents aromatic features. Those represent electron-rich regions in a molecule but behave close to lipophilic PPPs. Because of this relation between R and L the corresponding mismatch receives a small reward of 1 instead of a penalty despite the fact that lipophilic regions do not exhibit this electron-richness.

Gap penalties for the affine penalty model were determined in a grid search with gap open penalty between 2 and 15 and gap extension penalty equal to any value lower than the gap open penalty. The combination of gap open penalty -3 and gap extension penalty -1 showed best results.

5.2.2 - Alignment Evaluation

The global alignment of two sequences identifies similar parts as matched regions. But the alignment of symbols between sequences alone does not indicate homology. For this purpose alignment evaluation methods have to be used that calculate a similarity based on the alignment. A simple and intuitive method is the calculation of ‘percent sequence identity’ (PID) between sequences. This method has been used successfully for the identification of homologue proteins.^{71,72} In the preliminary parameterization PhAST assesses similarity between two molecules as the PID calculated from the alignment of their corresponding PhAST-sequences as shown in Equation 16

$$PID(X,Y) = \frac{M(A(X,Y))}{L(A(X,Y))} \quad (16)$$

where $A(X,Y)$ is the global alignment of sequences X and Y , $M(A(X,Y))$ is the number of matches in $A(X,Y)$ and $L(A(X,Y))$ is the length of the $A(X,Y)$, that is the length of either sequence including gapped positions. This definition is based on a measure proposed by Doolittle,⁷¹ but in contrast to that first idea of sequence identity calculated from alignments it does not exclude terminal gaps.⁷³ This modification is necessary as PhAST is supposed to calculate the similarity between complete molecules. The exclusion of regions aligned to gapped positions would confine similarity assessment only to fractions of molecules and as a consequence calculated similarities would no longer describe relationships between complete structures.

5.3 - Retrospective Evaluation

5.3.1 - Dataset

Screening performance of different PhAST configurations and other virtual screening methods was assessed through retrospective virtual screenings employing the COBRA¹¹ collection of bioactive compounds as reference dataset that contains active and inactive compounds for different targets. The COBRA library was used in version 6.1 containing 8,311 molecules with target receptor information compiled from selected scientific journals. Retrospective screenings were performed on six targets listed in Table 11.

5.3.2 - Performance Measure

In the scenario retrospective virtual screenings it is known which molecules are active against the same biological target as the query. The enrichment capabilities of a virtual screening method, meaning the assignment of better ranks to actives than to inactives, can be assessed by the ranks of actives assigned in the screening. Several metrics for this purpose have been proposed. Virtual screening aims at enrichment of active compounds on early ranks in the ranked screening library. As a consequence, performance on the first part of the ranked

Table 11. Targets in the COBRA library version 6.1 used for retrospective virtual screenings. Shown are abbreviations used in this study as well as the number of active compounds. The total number of molecules in the COBRA library is 8,311.

Target	Abbreviation	No. Actives
Angiotensine-converting enzyme	ACE	34
Cyclooxygenase 2	COX2	136
Dihydrofolat-reductase	DHFR	64
Factor Xa	FXA	228
Peroxisome-proliferator activated receptor γ	PPAR γ	44
Thrombin	THR	183
Total		689

library is of special interest and a good performance measure should be able to detect such ‘early enrichment’.

The ‘enrichment factor’ (EF) is one of the simplest performance measures for virtual screening.⁷⁴ It compares the ratio of actives to inactives within the first n ranked samples to uniform distribution. Typical values for n are 1% and 5% of the library size. The enrichment factor is calculated according to Equation 17.

$$EF(n) = \frac{n^+ / n}{N^+ / N} \quad (17)$$

where n^+ is the number of actives in the first n ranked samples, N^+ is the total number of actives in the library and N is the total number of compound in the library. The enrichment factor is easy to calculate and to interpret but has several drawbacks. It depends on the number of actives. For the same n and N , increasing N^+ lowers the range of possible enrichment factors. Choosing n is critical for the enrichment factor. It is calculated based on the number of high ranked compounds instead of the number of different ranks, so it does not consider ties. Furthermore the order of actives within the first n compounds does not matter, but cases where the n^+ samples are ranked on the first n^+ positions should clearly be preferred.

The receiver ‘operating characteristic’ (ROC) curve plots the true positive rate over the false positive rate.⁷⁵ But the visual comparison of a multitude of ROC curves would not be feasible. The corresponding area under the curve (ROCAUC) is easy to calculate and became an established performance measure for ranking methods. ROCAUC can be calculated according to Equation 18.

$$ROCAUC = \frac{1}{(nN)} \sum_{k=2}^N F_a(k) [F_i(k) - F_i(k-1)] \quad (18)$$

where N is the library size, n is the number of actives, $F_i(k)$ is the cumulative count of inactives at rank k and $F_a(k)$ is the cumulative count of actives at rank k . It equals the probability of ranking a randomly chosen positive sample better than a randomly chosen negative sample. Early enrichment is visible in ROC curves, but ROCAUC fails in the detection of such behavior. The ROCAUC score of a random model that has to be outperformed, corresponding to uniform distribution of actives in the screening library is, 0.5.

The ‘Boltzmann-enhanced discrimination of receiver operating characteristic’ (BEDROC) is based on the idea of an exponential weighting according to rank.¹³ It emphasizes the beginning of the ranked list, giving more weight to early ranked samples. The exponential weighting function can be influenced through an ‘early recognition parameter’ α determining the range from the beginning of the ranked list with most influence: higher values for α correspond to fewer early ranks dominating the BEDROC score. 20 is suggested as default value for α .¹³ BEDROC is calculated according to Equation 19

$$BEDROC = \frac{\sum_{i=1}^n e^{-\alpha r_i/N}}{R_a \left(\frac{1 - e^{-\alpha}}{e^{\alpha/N} - 1} \right)} \times \frac{R_a \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_a)} + \frac{1}{1 - e^{\alpha(1-R_a)}} \quad (19)$$

where N is the number of compounds in the library, r_i is the rank of the i -th active and R_a is the ratio of actives. Because of its ability to detect early recognition, in this work BEDROC was used for performance evaluation in retrospective screenings, with $\alpha = 20$. The BEDROC score of a random model that has to be outperformed, corresponding to uniform distribution of actives in the screening library, is $1/\alpha = 0.05$ for $\alpha = 20$.

5.3.3 - Significance Assessment

When two virtual screening methods or different parameterizations of the same method perform different in retrospective virtual screenings, indicated by different scores received from the performance measure of choice, the question is whether this difference is significant. The most powerful solution to the problem of significance assessment found so far is a paired permutation test.^{76,77} It has the null hypothesis that virtual screening method P performs significantly better than method Q . Assuming p and q are rank lists of actives resulting from the virtual screening methods, the null hypothesis requires that – with BEDROC as example – $BEDROC(p) > BEDROC(q)$. As each active has two ranks, one in p and one in q , new rank lists p^* and q^* can be created by swapping its rank in p with its rank in q for each active with probability $1/2$. This was repeated 10^4 times and the frequency of the event that $BEDROC(p) - BEDROC(q)$ is less than $BEDROC(p^*) - BEDROC(q^*)$, the type I error rate for the null hypothesis, was used as p-value for significance estimation. 0.05 and 0.01 were used as significance levels.

6 - Influence of Canonical Atom Labeling on Similarity Searching

This section discusses the publication listed as Appendix A.

6.1 - Motivation

The preliminary parameterization of PhAST employed the Needleman-Wunsch algorithm for global sequence alignment.⁶⁴ Other alignment algorithms have been proposed. A faster algorithm described in Durbin *et al.* 1998 (referred to as FSM algorithm) gains computational speed through simplifications made for the alignment process: The introduction of subsequent gaps in both sequences is forbidden, they have to be separated by at least one match or mismatch.⁷⁸ The assumption of two insertions or deletions not following directly onto each other is biologically valid and reduces the number of necessary computations. The alignments obtained with this algorithm are guaranteed to be identical to those calculated by the Needleman Wunsch algorithm if the sum of both gap penalties is lower than the lowest mismatch score.⁷⁸ If that is not the case, resulting alignments might differ slightly. The FSM and Needleman-Wunsch algorithm were compared in this study in a set of 689 virtual screenings (*vide supra*). Similarity of results was assessed by the respective retrospective performance measured using BEDROC and the correlation of rankings obtained with each algorithm for the same screening. Both algorithms were compared in the same preliminary parameterization (*vide supra*).

The preliminary parameterization of PhAST had two drawbacks that conflicted with the ideal case of line notation and sequence comparison.

First, it employed the Weininger algorithm for graph canonization⁴⁶ as linearization method applied to the pattern of potential pharmacophoric points (PPPs) associated with a molecule. Many properties used for vertex prioritization in this algorithm depend on atom types in the original molecule: atomic symbol, number of neighbors, number of connected vertices and the number of connected vertices that are non-hydrogen atoms. Because this conflicts with the idea of a line notation that depends only on graph topology (*cf.* Figure 6), several other methods for graph canonization were investigated. In addition to the Weininger algorithm two other algorithms developed for canonical atom labeling of chemical structures

were included in this survey: The Jochum-Gasteiger algorithm⁷⁹ and the method developed by Prabhakar and Balasubramanian.⁸⁰ All three algorithms were evaluated in an implementation adapted to the problem of indexing vertices in a graph of potential pharmacophoric points instead of a molecular graph. In a second implementation, each algorithm was modified in a way that prioritization criteria depending on the atom type or potential pharmacophoric point were excluded from the prioritization process. Besides these algorithms originating from molecular graphs, methods developed for dimensionality reduction were investigated for graph linearization. They can be applied to this problem by embedding the graph of potential pharmacophoric points in one dimension, yielding a linear sequence of vertices from which canonical labels are deduced. Methods investigated were Principal Component Analysis,⁸¹ Laplacian Eigenmaps,⁸² Isomap⁸³ and Minimum Volume Embedding.⁸⁴ The necessity of graph canonization was proven by comparison to a random model for symbol concatenation.

Second, percent sequence identity (PID)⁷¹ is not sensitive to different types of mismatches. Sequence alignment itself is sensitive to these differences as they are scored differently. But calculating PID for alignment evaluation, these differences are neglected. The alignment score on the other hand depends directly on these scores. Furthermore it was shown for protein sequence alignments that the alignment score is more suitable to the detection of homologues compared to PID.⁸⁵ Due to these reasons the alignment score as method for evaluation of global alignments of PhAST-sequences was investigated, comparing two variants of PID and three measures derived from the alignment score. All combinations of canonization algorithm and alignment evaluation measure were evaluated in a set of 689 retrospective virtual screenings, optimizing gap penalties in a grid-search.

Algorithms used for graph canonization compared in this study employed diverging principles. The Jochum-Gasteiger algorithm assigns vertices based on buriedness, the Weininger algorithm creates equivalence classes based on topological properties and the Prabhakar-Balasubramanian method labels subsequent vertices on uninterrupted paths through a graph. Methods for dimensionality reduction applied to graph linearization included linear as well as non-linear methods. An ideal canonization algorithm would mirror modifications to the molecular graph in the corresponding PhAST-sequence without changes to subsequences originating from unchanged subgraphs. In order to assess to what degree the compared algorithms comply with this ideal case, PhAST-sequences generated from molecules with small topological and functional alterations were compared with regard to neighborhood relations between vertices and symbols.

6.2 - Discussion

Changing the alignment algorithm from Needleman-Wunsch to FSM was a significant improvement for PhAST as virtual screening method. With FSM, calculated alignments were not the exact optimal alignments in some cases, but high correlation between ranked lists obtained in screenings indicated that the method still was suitable for virtual screening. The comparison speed more than doubled. Both, retrospective and prospective applications benefit from this speed-up: The number of parameterizations that can be evaluated in retrospective experiments increases as well as the size of libraries that can be screened for hits in the same time.

Graph canonization was shown to be a necessary step in PhAST-sequence generation. Retrospective performance significantly increased if the Weininger algorithm for canonical labeling was employed instead of randomized symbol concatenation. But differences between canonization algorithms measured by their retrospective performance were dominated by those caused by the alignment score as alignment evaluation measure. This outcome was not expected. PID and the alignment score evaluate the same alignments, but the alignments themselves are altered if PhAST-sequences are generated using different canonization algorithms. So the expectation was for the variable influencing generation of PhAST-sequences to have higher influence.

Although differences between alignment evaluation methods were more severe, the canonization algorithm influenced retrospective performance as well. Highest retrospective performance was observed with Minimum Volume Embedding employing a Diffusion Kernel. This technique of non-linear dimensionality reduction relies only on distances between vertices in the graph of potential pharmacophoric points measured based on topological distances. This way it is independent from atom types in the molecular graph and corresponding types of PPPs. This way, it is consistent with an the concept of an ideal canonization algorithm.

The analysis of canonization robustness against modifications to the molecular graph showed that none of the compared algorithms retained PhAST-sequence subsequences corresponding to identical subgraphs unaltered. Nevertheless all canonization algorithms succeed in generating PhAST-sequences yielding enrichment. This indicated that global sequence alignment as sequence comparison method was flexible enough to compensate for these

deficiencies. Comparison of canonization algorithms by their ability to retain symbol order revealed similarities between methods. Dimensionality reduction methods exhibited a diverging behavior compared to methods specifically designed for canonical atom labeling. But computational cost for this analysis was higher than actually performing a complete retrospective comparison due to the large number of altered molecular graphs. As a consequence, this kind analysis is suitable for the comparison of algorithms. But for assessment of screening performance, retrospective experiments remain the method of choice.

Findings of this study helped to point further investigations and development of PhAST into promising directions. Findings for the alignment of amino acid sequences have been affirmed for PhAST-sequences as well: The alignment score was more suitable for the detection of similarities (amino acid sequences: homologues) than PID. Differences between PID and the alignment score were significant. In combination these results encourage the implementation and evaluation of further alignment evaluation measures. For proteins, it has been shown that measures based on significance estimation are more suitable for reliable homologue identification than both, PID and alignment score.⁸⁵ Several approaches for the calculation of p-values have been published with regard to local alignments only.⁸⁶⁻⁸⁹ But only methods applicable to global alignments can be investigated in PhAST.

Concluding, PhAST was significantly improved in this study. The parameterization with highest observed retrospective performance was:

- Canonization: Minimum Volume Embedding, Diffusion Kernel (diffusion parameter 0.4), covalent connectivity
- Gap Open Penalty: 5
- Gap Extension Penalty: 1
- Alignment Evaluation: Alignment score normalized to alignment length

This parameterization was used as basis of all further investigations.

7 - Influence of the Third Dimension on Text-based Similarity Searching

This section discusses the publication listed as Appendix B.

7.1 - Motivation

Descriptors calculated from two-dimensional (2D) molecular representations have been successfully applied in virtual screening campaigns.^{37,38} But bioactive conformations of compounds are three-dimensional, as receptor structures are three-dimensional as well. Because of this fact, virtual screening methods handling three-dimensional (3D) conformations should yield better results, manifesting in higher enrichment observable in retrospective experiments. This should at least be true if the bioactive conformation of molecules is known. For most compounds this ideal case does not apply, but computational methods for the generation of low-energy 3D conformations have been developed. Until today there is no agreement whether 2D or 3D descriptors should generally be preferred.^{90,91}

In order to investigate whether PhAST can benefit from the application to three-dimensional conformations of molecules, the canonization process was modified. All methods for dimensionality reduction employed as canonization algorithms were applied to 3D conformations and systematic 2D structure diagrams in order to assess differences in screening performance. In case of Minimum Volume Embedding, additional kernel functions were investigated for this purpose. Because computational cost increases with increasing number of compared molecules, only the best performing version identified in the application to single 3D conformations was evaluated in retrospective experiments with multiple conformations per molecule.

Another step in PhAST that can be modified to exploit 3D information is the alignment algorithm itself. For protein sequence alignment scoring systems have been proposed that reward structural instead of functional similarity. These methods succeed in the identification of similar proteins by shape comparison. Six different parameterizations of PhAST employing such a technique named ‘Double Dynamic Programming’ (DDP)⁹² were compared for the evaluation of the applicability of this method to the comparison of small organic molecules.

There are a multitude of available virtual screening techniques. New methods are only useful if they yield new chemical entities on early ranks. The novelty of PhAST in this context was assessed in comparison to common methods. For this comparison, retrospective performance (measured by BEDROC) including the significance of differences in enrichment capability and ranks of actives measured by rank correlation⁹³ were used as similarity measures. Data fusion, the combination of complementary screening methods, has been reported to enhance screening performance.⁹⁴ But to the best of knowledge, so far no criterion for the selection of such fusion candidates was proposed. This study investigated rank correlation for this purpose by selecting a suitable method that was then combined with PhAST.

7.2 - Discussion

Retrospective comparison of different canonization algorithms employed in PhAST applied to 2D and 3D representations of molecules demonstrated that the dimensionality of molecular representations influences screening behavior. But differences were observed mostly in compound ranking, not in overall enrichment. The application of Minimum Volume Embedding employing a Diffusion Kernel in combination with covalent connectivity to 2D structure layouts still displayed best screening performance. Screening performance was increased by usage of multiple 3D conformations. But compared to the increase in computational cost, the increase was too small to justify this rise in cost.

PhAST was shown to have screening performance comparable or superior to other screening approaches. But the comparison of ranked lists obtained with different screening methods revealed that rankings calculated with PhAST were dissimilar to those of other methods. As a consequence, PhAST ranks active compounds on early ranks that are missed by other methods. This underlined the novelty of the PhAST concept.

Descriptions of data fusion methods for virtual screenings mostly describe fusion rules or results from combined descriptors that were chosen by intuition. But they do not suggest selection criteria for methods selection. In this study, rank correlation was shown to successfully identify promising fusion candidates. ‘Pseudoreceptor Point Similarity’ (PRPS)⁹⁵ was selected as candidate for data fusion with PhAST. The observed increase in retrospective performance was assessed as significant.

The application of structural instead of functional scoring systems to the alignment of PhAST-sequences for similarity assessment of small molecules yielded enrichment better than random active distribution in most cases. But only for cyclooxygenase-2, functional scoring systems was outperformed. Due to a small binding pocket, most of the actives in the screening dataset are small and possess a common structural element. Only the query molecule sharing this element achieved high enrichment. This indicated that structural similarity assessment through sequence alignment is intolerant to structural variations. Further investigations of structural scoring schemes and combinations with functional score matrices should concentrate on this observation by comparing retrospective results for targets with similar constraints for active molecules. Furthermore, distances are measured without a directional component in the evaluated approach. For amino acid sequence alignments, improvements have been developed that place coordinate systems on every residue, enabling comparison of directions as well as distances. These methods improved structural scoring for proteins and should be evaluated for small molecules as well.

The comparison of PhAST with other methods clearly showed that there is no method for virtual screening that performs best on each target used in the comparison. These findings suggest that in prospective applications methods should be chosen, evaluated and fine-tuned for the application to a particular target to maximize screening success.

Using 3D molecular representations could not increase screening performance in a reasonable manner. That is why after this study the recommended parameterization of PhAST for prospective application remained:

- Canonization: Minimum Volume Embedding, Diffusion Kernel (diffusion parameter 0.4), covalent connectivity
- Gap Open Penalty: 5
- Gap Extension Penalty: 1
- Alignment Evaluation: Alignment score normalized to alignment length

8 - Influence of Scoring Systems on Text-based Similarity Searching

This section discusses the publication listed as Appendix C.

8.1 - Motivation

Scoring system for protein sequence alignments are based on mutation rates manifesting in observed amino acid exchanges.^{67,68} Given a set of multiple sequence alignments, scores can be calculated systematically using an established and well-founded framework. Scores in the PhAST score matrix for potential pharmacophoric points (PPPs) on the other hand were chosen only from conceptual isofunctionality of types and observed frequencies in drug-like molecules. A modification of the underlying atom-typing would change those frequencies and could affect the applicability of the existing scoring scheme. Minor changes in the assignments of PPPs to substructures would reflect in altered PPP frequencies. Major changes like the introduction of completely new interaction types (for example: so far no purely hydrogen-bond donor functionality has been assigned) would require an extended scoring scheme including all new types of matches and mismatches. Systematic approaches to score calculations are preferable to intuition-based scoring because they enable rapid computation of new score matrices and their evaluation. Because of these reasons principles from score calculations in amino acid score matrices were evaluated for PPPs of drug-like molecules in this study based on the six compound classes in COBRA used for retrospective evaluation.

The only concession necessary was the limitation to pairwise comparisons instead of multiple sequence alignment.⁹⁶ This was due to the sequence generation process including Minimum Volume Embedding: Small modifications can cause the one-dimensional coordinate system to invert. As a consequence, the correct comparison direction of a sequence pair is not known, and both possible combinations have to be evaluated. For n sequences there are 2^n possible combinations of orientations. Calculating this amount of multiple sequence alignments was not feasible, and it is unclear whether the correct one, in which all sequences have the correct orientation, can be identified at all. In addition to sequence alignments, a kernel-based assignment method⁹⁷ was evaluated. It operates on graphs and circumvents mismatches occurring in sequence alignments due to positioning compromises during linearization.

Besides systematic approaches for score calculation, stochastic optimization was applied to match and mismatch scores. This optimization iteratively proposed and evaluated new solutions and generated derivatives of the most promising one.

The application of the same score matrix to each position in a PhAST-sequence gives equal weight to all symbols of the same type. But as already implied by the definition of the term pharmacophore, this is not correct: The pharmacophore comprises only the interactions necessary for activity, not all possible interactions. The logical consequence was the application of weights to symbols in PhAST-sequences corresponding with interactions known to be essential for ligand-receptor-interaction, leading to position-specific scoring. The other way around, if screening performance increases with weights at particular symbols, their corresponding interactions are common to other actives and as a consequence most likely essential for activity. Both variants of weight application were evaluated in retrospective screenings.

8.2 - Discussion

All three methods for score determination yielded score matrices that performed superior to a simple uniform scoring scheme. The original scoring system remains significantly superior, but the two non-stochastic approaches generated reproducible and comprehensible scores. Therefore they are ideal methods for the standardized generation of score matrices that can be used for the evaluation of modifications to the interaction types and assignments employed in the atom-typing step of PhAST in future studies. Independent runs of stochastic optimizations returned similar but not identical matrices that perform not as good as those generated with the systematic approaches.

Two datasets were used in the stochastic optimization of the score matrix employed in PhAST. Optimization was stopped after a fixed number of iterations. A better solution would have been to use three datasets: Training, test and validation data. The training dataset that is subject to the optimization. Retrospective performance on this compound collections serves as fitness function for matrix evaluation. Screening performance on a test dataset can serve as stop criterion for the optimization: If performance further increases on the training data but decreases on the test data, scores are overfitted to the training dataset and optimization should be stopped. This way the test data influences optimization, and a third dataset is needed for objective evaluation. Retrospective performance on a validation dataset that at no point is used for optimization serves as unbiased performance measure. But publicly available and

well-curated datasets for ligand-based virtual screening methods rare. The maximum unbiased validation dataset (MUV)⁹⁸ was designed to eliminate analogue bias (high structural similarity between actives) and artificial enrichment (actives structurally too dissimilar from inactives), both leading to too optimistic estimations of screening performance. The resulting dataset turned out to be too hard for well-established screening methods.⁹⁹ Furthermore the binding modes of compounds are unknown and inactivity of presumed inactive compounds is not confirmed. That way MUV is not suitable as any one of test, training or validation dataset. The COBRA¹¹ and the Krier¹⁰⁰ dataset used in this study have unequal numbers of actives (and as a consequence: of presumed inactives). For some targets the number of compounds is high enough for saturation effects to occur.¹³ For others it is too low for cross-validation approaches to represent generalized performance estimation in a fold. Facing these facts, the performed optimization was a knowingly chosen compromise that yielded acceptable results. COBRA was chosen as test dataset because it has been used for performance assessment of previous parameterizations of PhAST, providing a large set of results that allow a comparative interpretation of new results. This demonstrates the need for well-built and curated datasets.

The application of weights to certain PPPs increased screening performance of PhAST significantly. So far general screening performance of PhAST was comparable to other established methods. But with the possibility to incorporate target- and query-specific knowledge about receptor-ligand interaction PhAST becomes superior to these methods. But even more important, the reversed application of weights allows the identification of essential features. This way PhAST cannot only be used as virtual screening method but also for pharmacophore elucidation and information mining in the elucidation of ligand-receptor interaction.

The recommended parameterization of PhAST remained:

- Canonization: Minimum Volume Embedding, Diffusion Kernel (diffusion parameter 0.4), covalent connectivity
- Gap Open Penalty: 5
- Gap Extension Penalty: 1
- Alignment Evaluation: Alignment score normalized to alignment length

Weights have to be chosen based on the particular target and query structure.

9 - Comparison of Text-Based Virtual Screening Techniques

The known text-based virtual screening techniques LINGO^{53,54} and ‘Comparison by Compression’ (referred to as CbC)⁵⁹ are based on SMILES^{45,46} representations of molecules. In order to show the novelty of the PhAST concept compared to these two methods, they were compared in retrospective studies on the COBRA¹¹ collection of bioactive reference compounds. Criteria were retrospective screening performance measured by BEDROC scores (calculated with $\alpha = 20$)¹³ and Kendall’s rank correlation coefficient⁹³ calculated from ranked lists condensed to active compounds. LINGO and CbC were evaluated as described earlier. SMILES were generated with MOE (Molecular Operating Environment, v2010.06, Chemical Computing Group Inc., Montreal, Canada). PhAST was used in the identified parameterization with best screening performance: Canonization with Minimum Volume Embedding, Diffusion Kernel (diffusion parameter 0.4) and covalent connectivity, the original scoring scheme for PPPs, gap open penalty = 5, gap extension penalty = 1 and alignment evaluation by alignment score normalized to alignment length. Results are presented in Table 12 to Table 14.

PhAST and LINGO both exhibited higher averaged retrospective performance compared to CbC (Table 12). For both, these differences were significant in more than 50% of all screenings at both tested significance levels (Table 13). LINGO displayed higher retrospective performance than PhAST. But whereas the superiority of LINGO to CbC was significant in 78% (76%) of all screenings at 0.05 (0.01) significance level, the difference to PhAST was significant only in 53% (51%). The other way around, CbC performed significantly better than LINGO in 17% (15%) of all screenings, but PhAST significantly excelled LINGO in 43% (42%). As a consequence, the superiority of LINGO to CbC was more prominent as that to PhAST. LINGO and PhAST outperformed each other in nearly

Table 12. Retrospective performance of PhAST, LINGO and CbC. Screening performance was measured as averaged BEDROC score per target calculated with $\alpha = 20$. The first column presents the averaged result for all targets.

	∅	ACE	COX2	DHFR	FXA	PPAR γ	THR
PhAST	0.40	0.40	0.40	0.57	0.42	0.25	0.36
LINGO	0.41	0.59	0.47	0.36	0.39	0.25	0.38
CbC	0.35	0.50	0.43	0.27	0.31	0.23	0.35

Table 13. Significance of difference in retrospective performance between PhAST, LINGO and CbC. For each method pair and target the percentage of screenings is presented one method performs significantly better than the other at 0.05 (0.01) significance level. The first column reports the averaged result for all targets.

	∅	ACE	COX2	DHFR	FXA	PPAR γ	THR
PhAST	61 (59)	15 (6)	38 (36)	97 (97)	66 (65)	52 (43)	72 (69)
CbC	33 (31)	71 (62)	58 (58)	2 (2)	29 (27)	27 (23)	24 (23)
PhAST	43 (42)	3 (3)	23 (23)	92 (92)	32 (31)	43 (39)	60 (58)
LINGO	53 (51)	94 (85)	74 (71)	5 (3)	65 (63)	43 (36)	35 (34)
LINGO	78 (76)	97 (97)	68 (68)	81 (81)	86 (85)	55 (50)	76 (74)
CbC	17 (15)	0 (0)	27 (26)	13 (9)	11 (8)	18 (7)	21 (21)

Table 14. Rank correlation between PhAST, LINGO and CbC. Rankings were purged of inactive compounds before calculating Kendall's rank correlation coefficient. The first column shows the averaged result for all targets.

	∅	ACE	COX2	DHFR	FXA	PPAR γ	THR
PhAST / CbC	0.37	0.44	0.30	0.41	0.35	0.36	0.37
PhAST / LINGO	0.41	0.43	0.35	0.48	0.38	0.41	0.40
CbC / LINGO	0.57	0.71	0.60	0.50	0.54	0.54	0.56

identical percentages of screenings. The averaged rank correlation between methods (Table 14) indicated that rankings created by PhAST were complementary to those of the other methods. The rank correlation to CbC was 0.37, that to LINGO was 0.41. Rankings created by CbC and LINGO were more similar, with an averaged rank correlation of 0.57. This seems reasonable, as both methods use the same representation of molecules (SMILES).

In conclusion, the comparison of PhAST to LINGO and CbC proved the usefulness of PhAST. It has comparable retrospective to the better performing one of both other methods (LINGO). At the same time the created rankings of active compounds diverge. In a prospective application this effect would manifest in ranking novel chemotypes at early ranks, generating new ideas in the early stage of a drug design campaign.

10 - Significance-Assessment in Global Sequence Alignment

10.1 - Motivation

Virtual screening yields a ranked list of molecules, with those ranked best that are most likely to have desired properties. Ranks of compounds are determined by their similarity to the query structure. This poses a perturbing problem: Every VS method will rank any collection of molecules, even if the dataset does not contain a single molecule with the desired biological activity.¹⁰¹ The similarity scores of most methods are a poor measure of significance, as analysis of multiple HTS runs revealed: Even molecules with a Tanimoto score¹⁰² calculated from Daylight binary fingerprints¹⁰³ above 0.85 to an active compound have only a 30% chance to be active against the same target.¹⁰⁴ Similarity measures fail to predict that there are no actives in the dataset. Only for structural fingerprints (binary vectors coding the presence / absence of structural features) the problem of significance was addressed recently,¹⁰⁵ yielding a framework for significance estimation.

PhAST employs sequence alignment for molecule comparison. In the original application to amino acid sequences, the significance of local alignments is expressed through E-values (the expected number of hits with a score equal or higher to the observed score under a random sequence model) or p-values (the probability of a hit with score at least as high as the observed score under a random sequence model) for a score s . Efficient methods for the calculation of p-values for local alignments have been proposed.^{86-89,106}

The calculation of significance estimates was investigated under the assumption that it might improve PhAST in two ways: First, it was shown that for PhAST the effect of alignment evaluation on screening performance even excels that of the canonization algorithm.¹⁰⁷ Screening performance of PhAST employing sequence identity was inferior compared to the normalized alignment score. Significance as ranking criterion (with highest significance ranked best) might be beneficial to PhAST as ranking method, even though for protein alignments it has been empirically found that ranked lists generated by alignment score and significance estimates are fairly similar. Second, significance estimates could identify screening libraries containing only molecules that most likely do not possess the desired biological activity, thus saving assay capacity in a screening campaign only for significant screening hits. For alignments of amino acid sequences, this effect already improved the identification of homologue proteins.⁸⁵

10.2 - Calculation of p-values

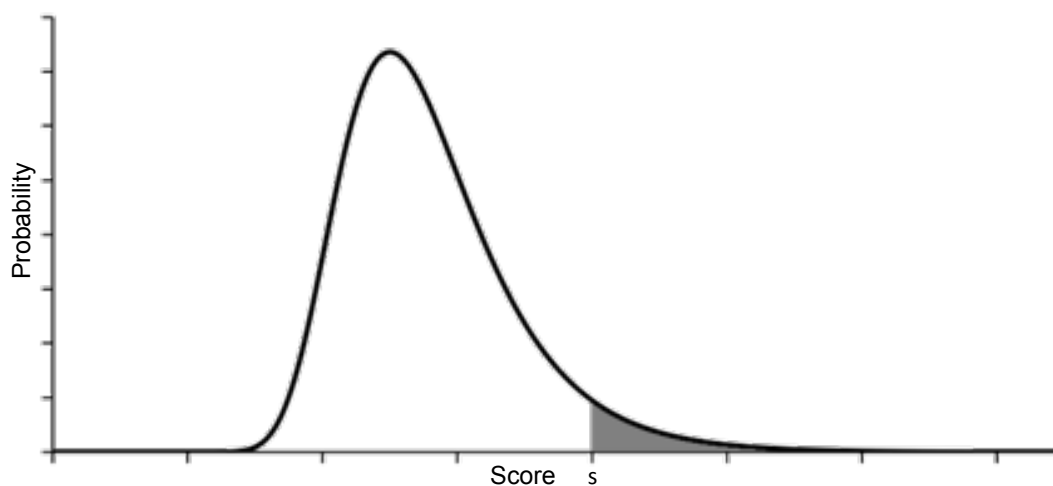


Figure 9. Visualization of p-value calculation. If the distribution of scores is known, the p-value of score s equals the area under curve above s (highlighted in dark grey).

The idea behind the calculation of p-values for sequence alignments is illustrated in Figure 9: If two sequences X and Y of lengths m and n yield alignment score s , the significance of s can be determined if the statistical distribution of alignment scores from the alignment of X with random sequences of length n is known. The p-value for score s equals the area under the curve above s . For protein sequence alignments, in the case of gapless local alignments of long sequences, empirical studies suggest a Gumbel distribution.¹⁰⁸⁻¹¹¹ Approximations for the more realistic scenario of gapped local alignments have been developed.^{112,113} Unfortunately, only little is known about the random distribution of optimal global alignment scores.¹¹⁴ But score distribution of gapless local alignments, gapped local alignments and global alignments are all accessible through sampling in a random sequence model.

All investigated sampling approaches utilize symbol frequencies f_i determined from the COBRA library of reference compounds in version 6.1, containing 8,311 compounds. These are: A = 4.95%, E = 1.44%, L = 19.65%, N = 1.22%, O = 24.63%, P = 1.80%, Q = 1.58%, R = 41.61% and U = 3.11%. Significance estimates in form of p-values were calculated for not-normalized alignment scores that performed best with gap open penalty = 7 and gap extension penalty = 1. All sampling approaches are explained using chlor- and acetylpromazine displayed in Figure 10 as example.

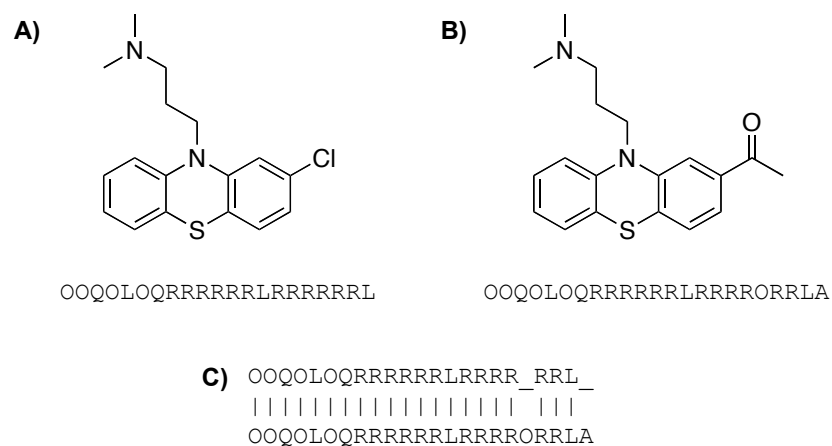


Figure 10. Comparison of chlor- and acetylpromazine using PhAST. A) chlorpromazine with corresponding PhAST-sequence. B) acetylpromazine with corresponding PhAST-sequence. C) alignment of PhAST-sequences shown in (A) and (B). PhAST-sequences were generated using Minimum Volume Embedding with Diffusion Kernel (diffusion parameter 0.4) and covalent connectivity. Sequence alignment was calculated by the FSM algorithm using the original PhASTscoring system for potential pharmacophoric points, gap open penalty = 7 and gap extension penalty = 1. Alignment score is 64.

10.2.1 - Simple Sampling

In simple sampling the query sequence X is aligned to a large number (e.g. 10^5) of random sequences of length n . The resulting score histogram can be used for the calculation of p-values. The problem with this approach is that only the region of alignment scores with high probabilities is sampled. The rare-event tail (high scores, low probability) is not accessible with this technique. Figure 11 shows score distributions obtained with simple sampling: The PhAST-sequence generated from chlorpromazine was aligned to random sequences of length 23, corresponding to the length of the PhAST-sequence of acetylpromazine. The numbers of generated scores are 10^4 , 10^5 , 10^6 and 10^7 . This practical example illustrates the drawbacks of simple sampling: The alignment of PhAST-sequences of chlor- and acetylpromazine has score 64. The highest score with simple sampling (59) was generated with 10^7 samples. As this maximum sampled score is below the actual alignment score, the p-value calculated for the alignment of chlor- and acetylpromazine would be 0. A Lilliefors test¹¹⁵ with the null hypothesis of the data coming from a normal distribution was performed with each sampled distribution. In each test the null hypothesis was rejected with a p-value below $2.2 \cdot 10^{-16}$. This is a strong argument that the distribution of scores obtained from global alignments of PhAST-sequences is not of the family of normal distributions.

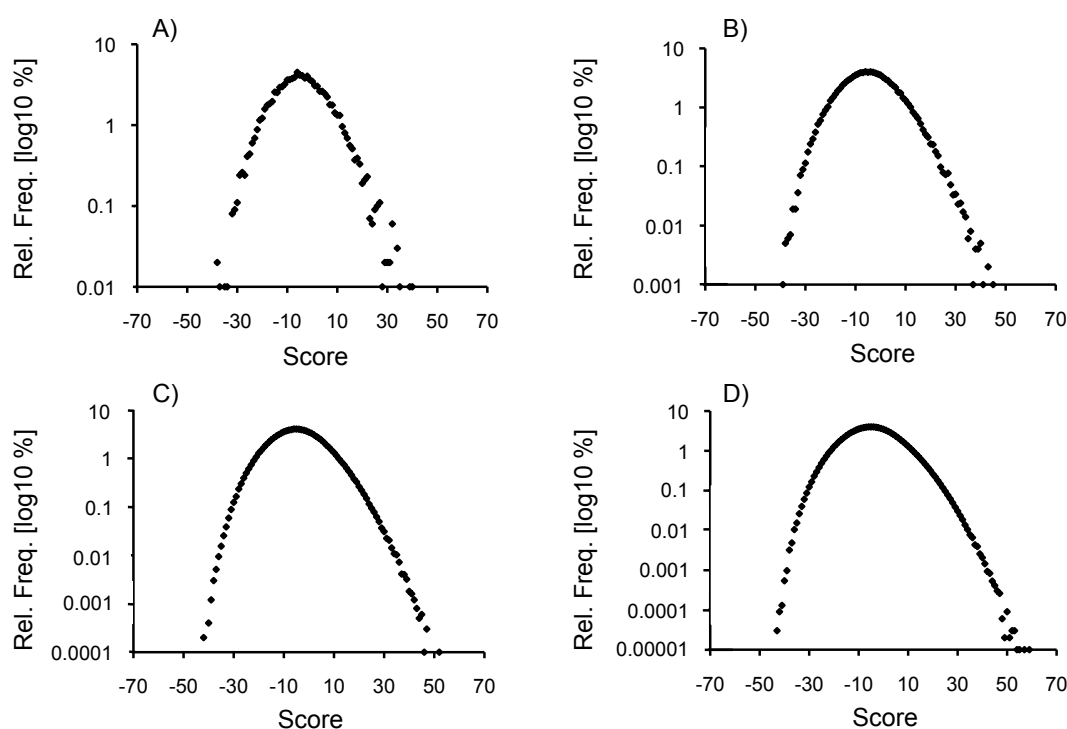


Figure 11. Simple sampling alignment score distributions. PhAST-sequence generated from chlorpromazine was aligned with 10^4 (A), 10^5 (B), 10^6 (C) and 10^7 (D) random sequences of length 23. This length corresponds to the PhAST-sequence of acetylpromazine. The alignment of PhAST-sequences generated for chlor- and acetylpromazine has score 64. Logarithmic Y-axis for improved visualization of low relative frequencies.

10.2.2 - Sampling of Rare Events

The rare-event tail of the distribution of alignment scores is accessible using the Metropolis-Hastings algorithm.^{116,117} A method implementing this concept in a Markov Chain Monte Carlo (MCMC) Simulation, that will be referred to as ‘Rare Event Sampling’, was proposed by Hartmann.^{89,118} In this work, it was evaluated for the calculation of p-values in PhAST. The original method has successfully been applied to significance estimations of amino acid sequence alignments. It uses the idea of Importance Sampling¹¹⁹ to estimate a particular distribution while having samples generated from another one. The probability distribution from which scores are sampled is altered in a way such that the region of interest is sampled with high probability.

The algorithm is based on a Markov Chain with states C_i and transition probabilities p_{C_i, C_j} between them. The Markov Chain has to be ergodic, meaning that from one state of the chain each other state is accessible through other states in finite time. In this application each state of the Markov Chain represents a global alignment of the query sequence X and a

random sequence Y_i with associated score s_i . If the Markov Chain is in C_t with score s_t at time t , a new state C^* with score s^* is proposed. It is accepted as state C_{t+1} with the Metropolis probability $\min[1, \exp(\Delta s/T)]$, where $\Delta s = s^* - s_t$. If C^* is accepted, s^* is counted as sampled score. If C^* is rejected, C_t is accepted as C_{t+1} and s_t is counted as sampled score s_{t+1} . Random sequences Y^* in states C^* are generated from sequence Y_i as follows: One position in Y_i is chosen and deleted by random with all positions being equiprobable. A new symbol at this position is chosen according to symbol frequencies f_i .

Choice of T has consequences for the region that is sampled: With high T the sampling tends towards low scores, with low T towards high scores. To describe the distribution of scores over a wide range, simulations must be carried out with several choices of T . For each run, the unbiased distribution of scores is obtained by scaling relative frequencies of scores with $\exp(-s/T)$. Histograms calculated with different temperatures can be patched together empirically, yielding the final distribution: Simulations with different T have overlapping regions. Starting with simple sampling and the simulation with maximum T for each pair of distributions a rescaling factor can be calculated from overlapping regions such that the difference between shared scores in the distributions is minimal.

Equilibration of Markov Chains can be determined empirically: Two simulations are started simultaneously with each combination of sequence length of the random sequence and temperature. The first simulation starts from high alignment scores, the second from low alignment scores. Equilibration is reached when for the last $t/2$ steps of both simulations averaged scores agree within error bars and this is true for the rest of the simulation.

The original method employs a Markov Chain with two random sequences instead of one fixed sequence (the query) and one random sequence. The described modification allows sampling from a more realistic scenario: During searches in databases and libraries the query remains always identical. As a consequence the score distribution of this particular query sequence and random sequences should be used for significance assessment (Hartmann, personal communication).

Rare event sampling was illustrated using the exemplary case of chlor- and acetylpromazine, results are displayed in Figure 12. With $T = 2.5$, equilibration as defined above was reached for the first time after 30 steps (Figure 12A), where the interval between sampled steps equals 1000 generated alignment scores. But both chains diverged again after ten more steps. Equilibration was reached again 8 steps later (step 48) and maintained for the remaining samples. Figure 12B illustrates the effect of different choices for T . With low

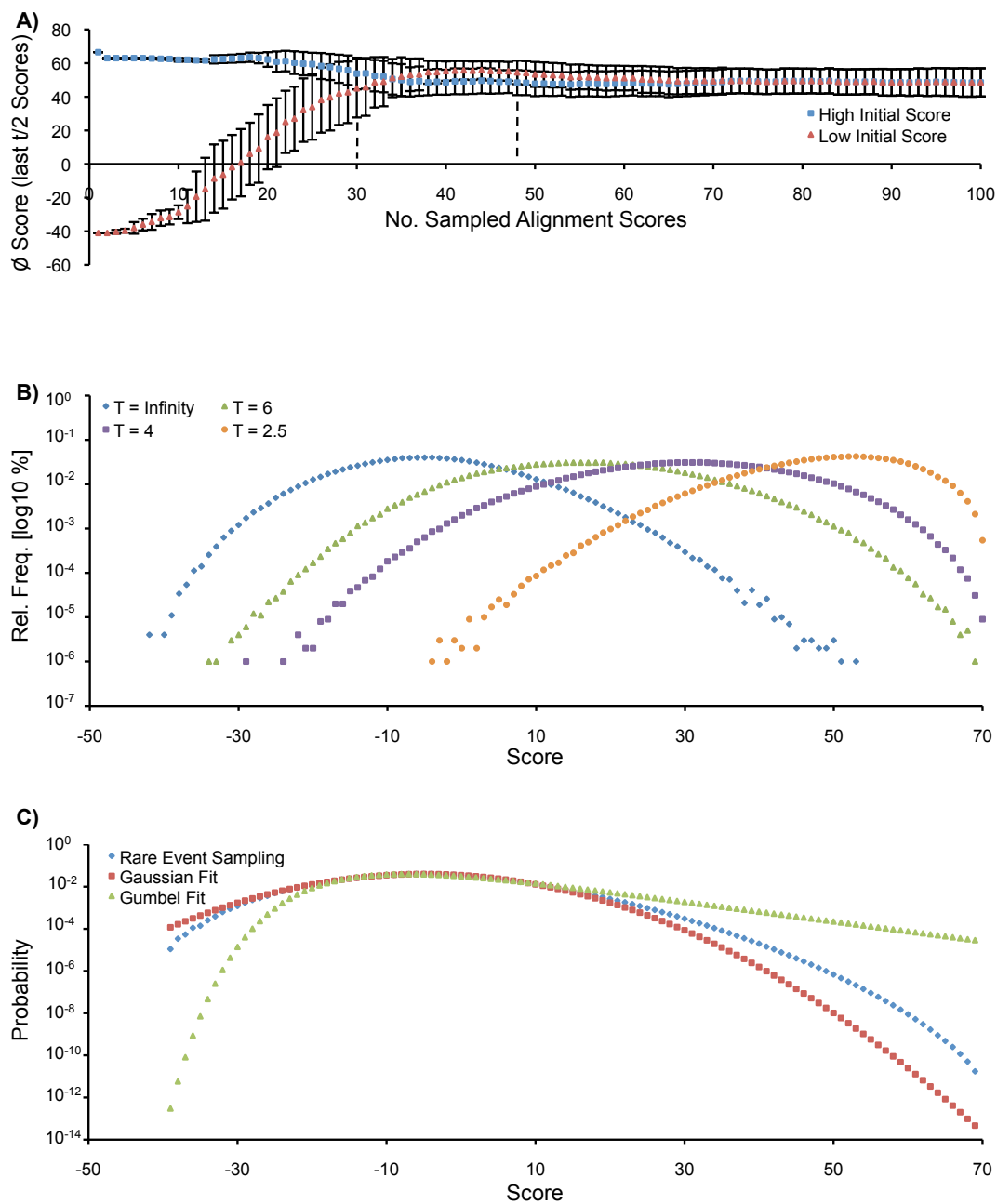


Figure 12. Rare Event Sampling. Illustrated is the example of the PhAST-sequence of chlorpromazine and random sequences of length 23, corresponding to the PhAST-sequence of acetylpromazine. A) Equilibration time determination for two Markov Chains sampling with $T = 2.5$, one starting from high alignment scores, the other starting from low alignment scores. Number of alignment scores in steps of size 1000. Dashed lines indicate steps where equilibration was reached. After step 48, equilibration was maintained for the complete sampling process. B) Sampled distributions with different choices for T . C) Distribution of alignment scores sampled with Rare Event Sampling and fits of Gaussian and Gumbel distributions. Gaussian parameterization: $a = 0.0398$, $b = -4.963$, $c = 14.11$; Gumbel parameterization: $\mu = -7.396$, $\sigma = 9.2920$.

values for T , high scores are preferred. Figure 12C presents the final distribution of alignment scores obtained for the PhAST-sequence of chlorpromazine and random sequences of length 23. A Gaussian distribution as shown in Equation 20

$$f(x) = a * e^{-\left(\frac{x-b}{c}\right)^2} \quad (20)$$

and a Gumbel distribution as defined in Equation 21

$$f(x) = \frac{1}{\sigma} * e^{-\frac{x-\mu}{\sigma}} - e^{-\frac{x-\mu}{\sigma}} \quad (21)$$

were fitted to this distribution using the Levenberg-Marquart algorithm.¹²⁰ Fits were evaluated by the root mean squared error (RMSE) calculated according to Equation 22

$$RMSE(G,H) = \sqrt{\frac{\sum_{i=1}^n (g_i - h_i)^2}{n}} \quad (22)$$

where G and H are two distributions and n is the respective number of samples. The region where probabilities are high agreed well in all three distributions, resulting in an RMSE of $3.9*10^{-4}$ for Rare Event Sampling and the fitted Gaussian distribution and $2*10^{-3}$ for Rare Event Sampling and the fitted Gumbel distribution. Both values indicate good fits. But for the calculation of p-values, the low probability region is of particular interest, so the RMSEs were recalculated based on logarithmic values. The recalculated RMSEs were 1.2 for Rare Event Sampling and Gaussian and 2.4 for Rare Event Sampling and Gumbel, indicating increased divergence in regions with low probability. As illustrated in Figure 12C, the p-value for the highest sampled alignment score 69 that was determined as $1.7*10^{-11}$ using Rare Event Sampling would be $4.6*10^{-14}$ in the fitted Gaussian distribution (overestimation by factor 10^3) and $2.9*10^{-5}$ in the fitted Gumbel distribution (underestimation by factor 10^6). These findings prove that the alignment scores obtained from global alignment of PhAST-sequences follow neither a Gaussian nor a Gumbel distribution, and as a consequence, that efficient sampling approaches like Rare Event Sampling are necessary for significance determination in PhAST. The p-value calculated for the original alignment of PhAST-sequences corresponding to chlor- and acetylpromazin with score 64 using Rare Event Sampling was $1.8*10^{-9}$.

10.3 - Retrospective Evaluation

Rare Event Sampling for p-value calculation was evaluated in a retrospective screening on the COBRA dataset with lisinopril (Figure 13) as query that is known to be active against the angiotensin-converting enzyme (ACE).¹²¹ Calculated p-values were used as ranking criterion with those molecules ranked best that receive lowest p-values. The final ranking was evaluated using the BEDROC metric and compared to alignment evaluation methods described earlier (PID1, PID2, S1, S2 and S3; see section Appendix A for detailed descriptions). Significance of differences in retrospective performance was assessed in a paired permutation test with 10^6 permutations. Ranked lists were compared by their rank correlation.

10.3.1 Parameterization

Simulations were carried out for the PhAST-sequence generated from lisinopril and each of the 85 sequence lengths encountered in the COBRA library (*cf.* Figure 2) with 12 temperatures (0.4, 0.5, 0.7, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 6.0, 8.0, infinity). With each parameterization, one simulation starting from high scores and one starting from low scores was performed as follows: The first simulation started from high alignment scores. If $m > n$, Y_0 equaled the first n symbols of X ; if $m < n$, Y_0 consisted of a copy of X , the remaining $n - m$ symbols were all of type L to minimize mismatch penalties. The second simulation started from low alignment scores. If $m > n$, Y_0 equaled a ‘negative copy’ of the first n symbols of X , where for each symbol in X the symbol with maximum mismatch penalty was chosen according to the scoring scheme. If $m < n$, Y_0 consisted of a negative copy of X , the remaining $n - m$ symbols were all of type L.

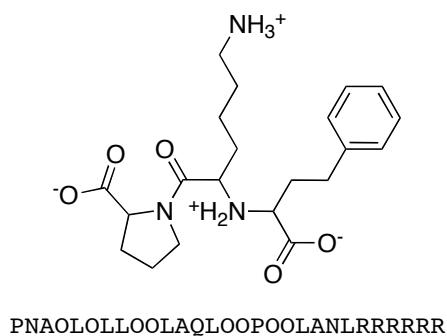


Figure 13. Molecular structure diagram and PhAST-sequence of lisinopril. This compound is a known active for the acetyl-converting enzyme.

Each simulation sampled 10^9 scores, but only each 10^3 -th score was used to avoid correlations between scores that occur in Markov Chain Monte Carlo sampling in contrast to fully random sequences. This resulted in 2,040 simulations, each returning 10^6 scores. Equilibration was assessed after termination, and both simulations with the same combination of sequence lengths and T were combined after equilibration was reached.

10.3.2 - Results and Discussion

Equilibration was reached in all simulations after at most 149,316 steps, with an averaged equilibration time of 868 steps. As a consequence, all distributions were calculated from at least $1.7 \cdot 10^6$ scores.

BEDROC scores calculated with different alignment evaluation methods as similarity measure are reported in Table 15. Retrospective performance of p-values was comparable to the other alignment evaluation methods with exception of PID2 that had significantly lower enrichment. Highest retrospective performance resulted from S2 as alignment evaluation method. These results show that p-values used for similarity assessment do not perform better than the so far best alignment evaluation method (S2), but yield comparable enrichment. But the observed difference was not significant at 0.01 or 0.05 significance level.

Ranked lists obtained from the same virtual screening setup with different alignment

Table 15. Retrospective comparison of p-value and other alignment evaluation methods. Reported are BEDROC scores and p-values assessing significance of differences in retrospective performance. BEDROC scores were calculated with $\alpha = 20$. p-values were calculated with 10^6 permutations in a paired permutation test.

	PID1	PID2	S1	S2	S3	p-value
BEDROC	0.4413	0.1239	0.4587	0.4708	0.4468	0.4579
PID1	-	$< 10^{-6}$	0.1175	0.0136	0.3807	0.1090
PID2		-	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
Score1			-	0.0732	0.1386	0.4864
Score2				-	0.0013	0.0917
Score3					-	0.1226
p-value						-

Table 16. Comparison of alignment evaluation methods based on rank correlation coefficients. Rank correlation was calculated from ranked lists obtained from virtual screenings of lisinopril against the COBRA collection of bioactive compounds with the respective alignment evaluation methods.

	PID1	PID2	S1	S2	S3	p-value
PID1	-	0.1221	0.5350	0.5467	0.5650	0.4894
PID2		-	0.1430	0.1183	0.0381	0.1662
S1			-	0.8732	0.7484	0.8862
S2				-	0.7653	0.8243
S3					-	0.6624
p-value						-

evaluation methods were compared by their rank correlation. The calculated correlation matrix is shown in Table 16. Rank correlation coefficients indicated closest similarity between p-value and the three variant of alignment scores. Within these, the not-normalized alignment score has closest relation to p-value. This is not surprising, as p-values are calculated based on this score.

The p-value calculated for the top-ranked compound in the performed retrospective screening was $1.05 \cdot 10^{-14}$. In order to obtain the same value through simple sampling, at least 10^{14} alignments would have had to be calculated – with no guarantee that at least one of the sampled scores would be equal or higher to the original score. For the calculation of all necessary p-values for retrospective evaluation, $85 \cdot 12 \cdot 2 \cdot 10^9 = 2.04 \cdot 10^{12}$ alignments were calculated. So the evaluation of all alignments with Rare Event Sampling was 100 times faster than the calculation of only one single p-value with simple sampling. This real-life example emphasizes the usefulness of Rare Event Sampling.

Significance estimates in form of p-values can be used for the estimation of thresholds that maximize hitrates (ratio of active to inactive compounds) and reduce costs in prospective screenings. As shown in Figure 14, the early region of the ranked list, where p-values are low, was mostly populated with active compounds. Figure 14C presents retrospective hitrates: If 10^{-9} would be chosen as threshold to indicate promising candidates for a prospective screening, this would result in a hitrate of 50% (7 actives in 14 compounds). The determination of a threshold depends upon the cost associated with making a mistake. A threshold of 0.01 means that an error occurs with 1% probability. Whether this is stringent enough depends on the actual cost of mistakes, for example the evaluation of a screening compound from a prospective application in an assay.

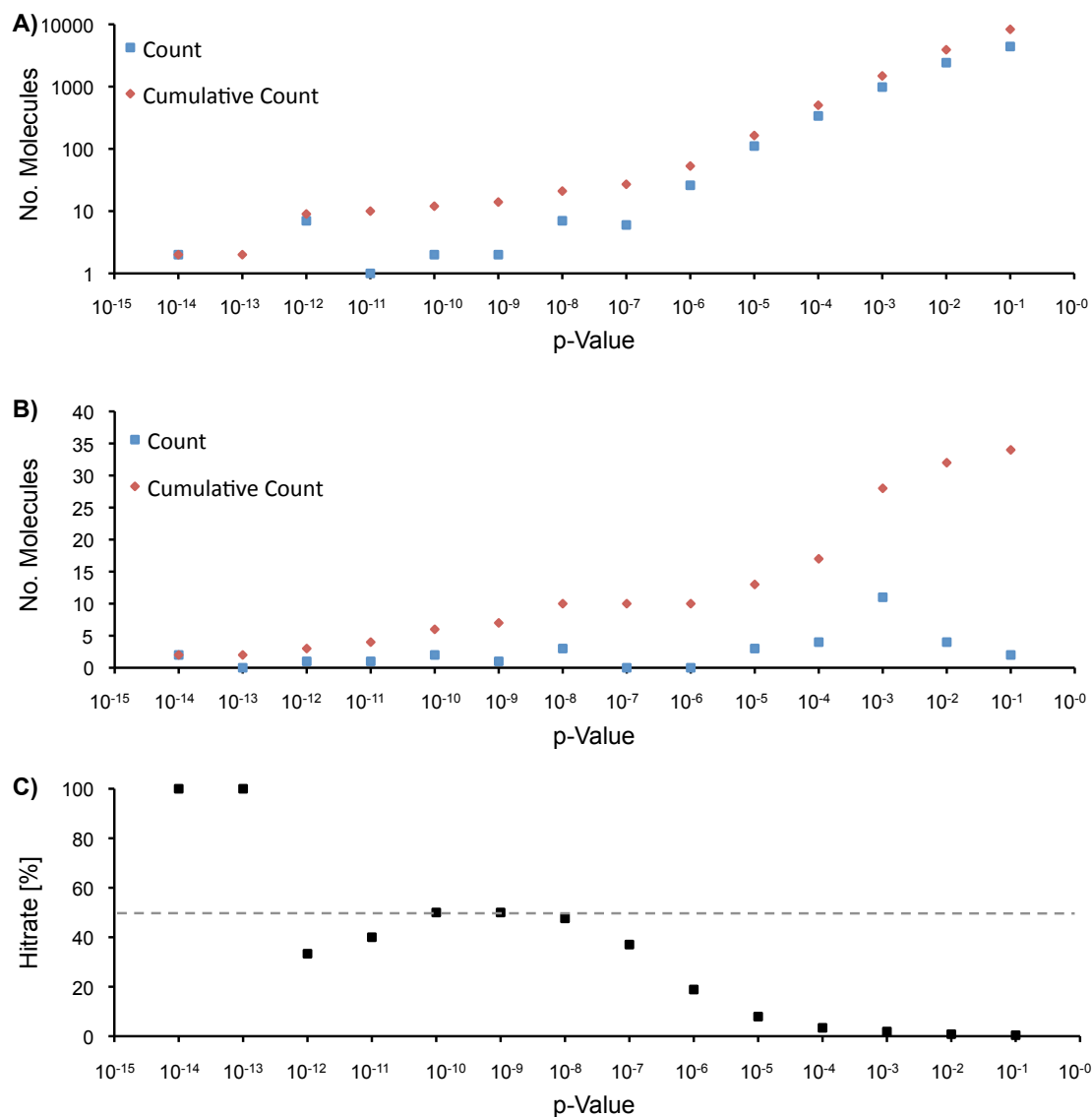


Figure 14. Distribution of p-values resulting from screening lisinopril against the COBRA dataset. A) Number of compounds with a certain p-value. B) Number of active compounds with a certain p-value. C) Retrospective hitrates calculated for certain p-value thresholds.

10.4 - Calculation of E-values

The number of low p-values that has to be expected by chance increases with the number of comparisons. This variable cannot be included in the calculation of p-values. As a consequence, the p-value for a score s has to be modified to account for the number of comparisons, yielding an E-value. In statistics, the Bonferroni correction can be used to address the problem of multiple comparisons:¹²² If a statistical test is performed n times at significance level α , this level may be appropriate for each individual comparison, but not for

the set of all comparisons. Due to the increased number of observations, the chance of a significant event occurring by chance increases. According to Bonferroni, the actual significance level can be calculated as $1 - (1 - \alpha)^n$. Another possibility to account for the number of comparisons being performed is to lower the alpha value for each test. This can be achieved by dividing the significance level by the number of tests performed, hence $\alpha_{actual} = \alpha/n$. The converse of this approach is the correction of the p-value by multiplication with the number of tests and has already been used in adapted forms for the correction of p-values in sequence alignment.¹²³ This operation yields an E-value. The relation of p-value and E-value of score s is described in Equation 23.

$$E(s) = N \times p(s) \quad (23)$$

Several choices for N have been reported or seem reasonable: i) library size n (the original Bonferroni correction), ii) $sc(L)$ where sc is the count of symbols and L is the library (suggested if variation in sequence lengths is large)¹²³, iii) $sc(L)/sc(q)$ where q is the query sequence, iv) $k \times sc(q) \times sc(L)$, ($0 < k < 1$). Figure 15 presents the same statistic as Figure 14C but with several choices for rescaling. Corrections (i) and (iii) behave nearly identical. Which variant of p-value correction is most suitable for prospective application and whether a choice $k \neq 1$ for correction (iv) is necessary can only be determined in additional retrospective and prospective applications. For the time being, there is no reason in evidence to deviate from the original Bonferroni correction. Consequently, a general and reusable E-value threshold of 1×10^{-5} seems reasonable. With this threshold, 15 compounds would have to be evaluated for their activity in the given example. The retrospective hitrate would be 50%. Both are acceptable values.

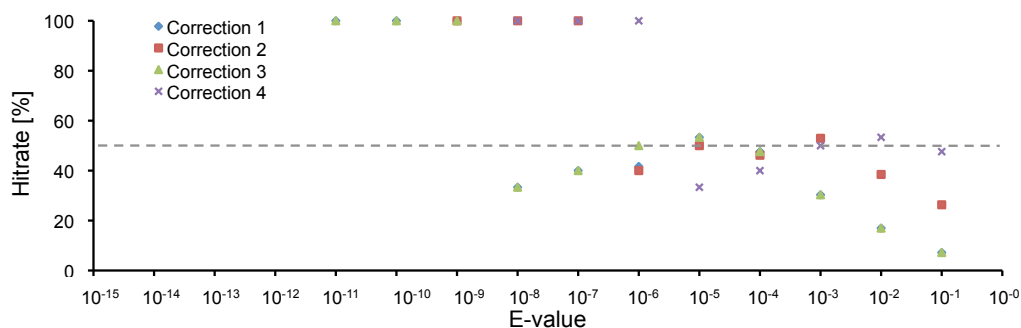


Figure 15. Retrospective hitrates for certain E-value thresholds resulting from screening lisinopril against the COBRA dataset.

10.5 - Discussion

The local alteration of sequences based on symbol frequencies during Markov Chain Monte Carlo simulations is an adequate model for protein sequences. But it does not account for local constraints due to protein structure or function, which could influence exchange frequencies in certain parts of a protein. Drug-like molecules are subject to even more severe constraints that make some symbol combinations in PhAST-sequences impossible. Examples are a number of R symbols (aromatic) that does not fulfill Hückel's '4n + 2' rule^{124,125} and high quantity of positive (P) or negative (N) charges: In the COBRA collection of bioactive reference compounds the respective maximum numbers were 8 Ps in a sequence of length 36 and 8 Ns in a sequence of length 84. The assignment of O symbols in most cases depends on the interaction types of the adjacent vertices, which is also ignored in the current approach. For application with PhAST, a more realistic model would be to use PhAST-sequences generated from random molecules with a fixed number of non-hydrogen atoms. This would require the development of a framework for randomized and chemically meaningful alterations to molecules of fixed size.

Equilibration time could be reduced by the application of parallel tempering, yielding a Metropolis-coupled Markov Chain Monte Carlo simulation (MCMCMC).¹²⁶ There, simulations are performed with different T in parallel ($T_1 < T_2 < \dots < T_{\max}$). After each simulation step a pair (T_i, T_{i+1}) is chosen by random and sequence configurations are exchanged with probability p_e according to Equation 24. But given the observed equilibration times this does not seem necessary.

$$p_e = \min\left(1, \exp\left(-\left(\frac{1}{T_i} - \frac{1}{T_{i+1}}\right) \times (S_i - S_{i+1})\right)\right) \quad (24)$$

The published version of Rare Event Sampling involves two randomized sequences, whereas the implementation proposed in this work leaves the query sequence fixed. Both variants have certain advantages and disadvantages.

With a fixed query sequence, score distributions have to be calculated in each screening. On first sight, this is a clear disadvantage because of the time necessary for simulations. But in this scenario it is possible to adapt the sampling process to a particular screening setup. First, the application of position weights remains possible. Second, symbol frequencies used in the generation of random PhAST-sequences can be adjusted to the

Table 17. Relative symbol frequencies in six molecule repositories. Collections of drug-like compounds: COBRA¹¹ collection of bioactive reference compounds, Specs vendor catalogue for small compounds in version 08/2010 (Specs, Delft, the Netherlands), ZINC¹²⁷ drug-like subset, ZINC¹²⁷ lead-like subset. Collections of natural products: Specs vendor catalogue for natural products in version 08/2010, Analyticon purified natural products of microbial origin release 100915, Analyticon purified natural products of plant origin release 100915 (Analyticon, Potsdam, Germany). Absolute numbers of PPPs are: COBRA = 244,505; Specs SC = 4,828,971; ZINC drug-like = 261,917,160; ZINC lead-like = 30,348,012; Specs NP = 13,460; MEGXm = 33,780; MEGXp = 615,769.

	COBRA 6.1	Specs SC	ZINC drug-like	ZINC lead-like	Specs N	MEGXm	MEGXp
A	4.95	4.84	6.30	6.52	4.99	7.83	4.05
E	1.44	0.49	0.14	0.50	3.47	7.45	12.60
L	19.65	15.53	15.50	16.56	38.57	36.28	27.59
N	1.22	0.48	0.13	0.50	0.51	1.58	0.63
O	24.63	20.83	25.90	24.66	27.42	29.37	37.29
P	1.80	0.34	0.16	0.23	0.65	0.10	0.06
Q	1.58	1.69	2.26	1.92	0.30	0.82	0.04
R	41.61	53.08	46.16	44.88	23.73	14.83	17.66
U	3.11	2.71	3.46	4.22	0.35	1.73	0.08

screening library. Table 17 presents relative PPP frequencies determined from seven different molecule repositories, three for small drug-like compounds, one for lead-like compounds and three collections of natural products. The corresponding correlation matrix is shown in Table 18. Low and in some cases not significant correlation (at 0.05 significance level) is observed between members of the two different groups of molecule collections, whereas correlation is high and significant at 0.05 and 0.01 significance levels within these groups. Despite the argued differences between drug-like and lead-like compounds (*vide supra*), the distributions of PPPs observed in the corresponding subsets of ZINC are fairly similar. Pairwise Chi² tests¹²⁸ performed on absolute frequencies indicate significant differences between each dataset pair with a p-value below $2.2 \cdot 10^{-16}$ as well as a test performed on all datasets at once. In case of the Chi² test, calculated significance is most likely due to the large number of samples. But these findings point to large and relevant deviations of symbol frequencies especially between repositories for drug-like molecules and natural products. These differences can be incorporated into significance estimations if score distributions for the random model are calculated for each screening. Related to this topic is the atom-typing in general: If the atom-typing is changed in the future, new symbols and symbol frequencies can be used in the random model.

With two randomized sequences on the other hand, score distributions can not be adapted to symbol frequencies, new symbol types and position weights. But they can be pre-calculated. Molecules from the COBRA dataset yield PhAST-sequences with 85 different lengths. During retrospective evaluations in this dataset, 3,655 combinations of sequence

Table 18. Correlation of PPP frequencies between six molecule repositories. Shown is Pearson's correlation coefficient and the corresponding p-value, the latter in parentheses. Collections of drug-like compounds: COBRA¹¹ collection of bioactive reference compounds, Specs vendor catalogue for small compounds in version 08/2010 (Sepcs, Delft, the Netherlands), ZINC¹²⁷ drug-like subset, ZINC¹²⁷ lead-like subset. Collections of natural products: Specs vendor catalogue for natural products in version 08/2010, Analyticon purified natural products of microbial origin release 100915, Analyticon purified natural products of plant origin release 100915 (Analyticon, Potsdam, Germany).

	Specs SC	ZINC druglike	ZINC leadlike	Specs N	MEGXm	MEGXp
COBRA 6.1	0.98 (1.03*10 ⁻⁰⁵)	0.99 (3.52*10 ⁻⁰⁷)	0.99 (1.13*10 ⁻⁰⁷)	0.79 (1.13*10 ⁻⁰²)	0.65 (4.70*10 ⁻⁰²)	0.71 (3.05*10 ⁻⁰²)
Specs SC	-	0.99 (1.20*10 ⁻⁰⁶)	0.99 (9.52*10 ⁻⁰⁷)	0.66 (4.55*10 ⁻⁰²)	0.49 (1.26*10 ⁻⁰¹)	0.55 (9.18*10 ⁻⁰²)
ZINC drug-like		-	1.00 (5.93*10 ⁻¹¹)	0.72 (2.79*10 ⁻⁰²)	0.57 (8.23*10 ⁻⁰²)	0.65 (5.03*10 ⁻⁰²)
ZINC lead-like			-	0.73 (2.42*10 ⁻⁰²)	0.58 (7.57*10 ⁻⁰²)	0.65 (4.90*10 ⁻⁰²)
Specs N				-	0.97 (2.77*10 ⁻⁰⁵)	0.90 (1.32*10 ⁻⁰³)
MEGXm					-	0.94 (2.91*10 ⁻⁰⁴)

lengths in alignments are possible. With the corresponding pre-calculated score distributions, significance assessment would only require the look-up of the correct distribution for a certain combination of sequence lengths. The calculation of all distributions would be costly in terms of time, but with an increasing number of performed retrospective and prospective screenings this initial investment would amortize. The pre-calculation of score distributions would enable the large number of screenings that is necessary to establish reliable E-value thresholds.

In this work, choices for T employed in Rare Event Sampling were based on preliminary experiments that determined values suitable for the application of this method to PhAST-sequences. But nevertheless, different T were chosen more or less arbitrarily. Wang and Landau proposed a Monte Carlo algorithm that is independent from choices of T .^{129,130} There, acceptance of a proposed step depends on the inverse density of states starting from a uniform distribution: The more times a state was visited in the past, the less likely is its acceptance in the future. The complete distribution is sampled within defined minimum and maximum values in one single simulation. This method can be further enhanced by performing multiple random walks in parallel with overlapping minimum and maximum values distributed in the interval of the minimum and maximum value of the complete distribution. This sampling method would render significance estimations independent from fixed choices of T .

The Bonferroni correction used for p-value adjustment distributes the significance level equally on all tests (in this case: alignments). As a consequence it is very conservative, because a p-value has to be very low to be still significant after the correction. It controls the probability of false positives but increases the probability for false negatives. More recent methods for p-value adjustments try to overcome this drawback. Examples are the Bonferroni-Holm procedure¹³¹ and the Benjamini-Hochberg method.¹³² Which method is best for significance estimation of chemical similarity with PhAST has to be determined in future studies.

11 - Prospective Application

During its development, PhAST was parameterized and compared to other state-of-the-art virtual screening methods in retrospective experiments. During this process and after the best-performing parameterization was determined, PhAST was employed in prospective virtual screenings for the identification of compounds that possess certain activity of biological interest.

11.1 - Bacterial Thymidinkinase of *Staphylococcus aureus*

This section discusses the publication listed as Appendix D. Methicillin-resistant *Staphylococcus aureus* is a widespread pathogenic bacterium.^{133,134} The combination of trimethoprim and sulfamethoxazole (SXT) has antimicrobial activity, as it inhibits the folic acid pathway, eventually blocking the bacterial synthesis of deoxythymidine monophosphate (dMTP) by thymidylate synthase. But *S. aureus* possesses a second pathway for the synthesis of dMTP by uptake of extracellular thymidine and subsequent phosphorylation via Thymidinekinase.^{135,136} In the presence of high extracellular levels of thymidine, a combination of halogenated 2'-deoxyuridine derivatives (see Figure 16) has been reported to exhibit synergistic antimicrobial activity against *S. aureus*.^{137,138} The downside of nucleoside analogues as inhibitors of bacterial thymidine kinase is their cytotoxicity. If they are phosphorylated to triphosphates and incorporated into DNA, they can lead to single-strand breaks.^{139,140} Therefore, screening for non-nucleoside analogues as inhibitors of bacterial Thymidinekinase was of special interest for this work.

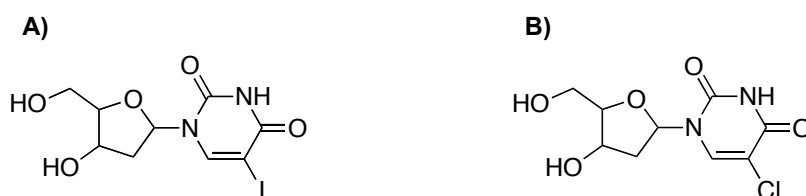
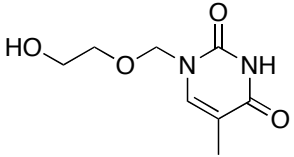
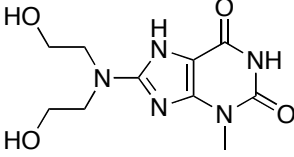
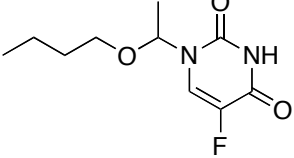
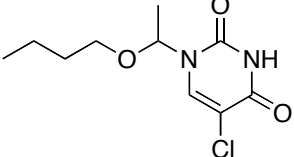


Figure 16. Structure diagrams of two halogenated 2'-deoxyuridine derivatives. A) 5-iodo-2'-deoxyuridine, B) 5-chloro-2'-deoxyuridine. Both are known to have antimicrobial activity against *S. aureus* in combination with SXT with minimal inhibitory concentrations of 0.0625 mgL⁻¹ for *S. aureus* strains ATCC 700699 and ATCC 29213.

Table 19. Results of virtual screening with PhAST for non-nucleoside inhibitors of bacterial thymidine kinase. MIC values represent the median of three experiments.

	Structure	Rank	MIC [mgL ⁻¹]	
			ATCC 700699	ATCC 29213
1		2	128	128
2		8	32	64
3		18	128	128
4		41	128	128

PhAST was used in a ligand-based virtual screening for non-nucleoside inhibitors of *S. aureus* thymidine kinase in the version performing best in retrospective experiments at that point in development: For graph canonization PhAST employed the Prabhakar algorithm,⁸⁰ sequence alignments were calculated with gap open penalty = 5 and gap extension penalty = 1. The screening library combined the vendor catalogues of Specs (v01/2009, Specs, Delft, The Netherlands) and Asinex Gold and Platinum collections (v11/2008, Asinex, Moscow, Russia). All compounds were protonated using the ‘wash’ function of MOE (v2008.10, Chemical Computing Group, Montreal, Canada). 5-chloro-2'-deoxyuridine was used as query. From the resulting ranked list, four compounds were selected and evaluated for their activity.¹⁴¹ Minimal inhibitory concentrations (MICs) of all selected candidate compounds were determined in combination with SXT against *S. aureus* strains ATCC 700699 (methicillin resistant) and 29213 (not methicillin resistant) in the presence of thymidine.

Obtained results are presented in Table 19. The only molecule with measured inhibitory activity was compound 2. With MICs of 32 mgL⁻¹ (ATCC 700699) and 64 mgL⁻¹ (ATCC 29213) activity was 50 and 100 fold lower compared to the query compound. Despite the decreased activity, PhAST succeeded in the identification of an active compound that is not a nucleoside-analog. Compound 2 is a purine-dione whereas the query is a pyrimidine-dione. This non-nucleoside-analog inhibitor could be further optimized in future studies.

11.2 - Application to γ -Secretase

γ -Secretase (GS) is an integral membrane protein. Among other substrates it processes the amyloid precursor protein (APP) as subsequent step to its proteolytic cleavage by β -Secretase. During this process it produces preferably amyloid- β (A β) peptides of length 40 and 42.^{142,143} A β 42 fragments are prone to oligomerize, eventually forming neurotoxic extracellular amyloid plaques, which are characteristically found in brains of patients suffering from Alzheimer's disease (AD).¹⁴⁴ The process of A β 42 oligomerization and the extracellular deposition of amyloid plaques are believed to be a major disease-causing step in the pathology of AD. This so called amyloid hypothesis served as the rationale for the development of GS inhibitors (GSIs) (in order to treat AD).¹⁴⁵ As GS processes ca. 80 peptidic substrates (*e.g.* the NOTCH receptor), the inhibition of GS has severe consequences besides inhibition of APP processing. Only recently (August 2010), Eli Lilly had to stop the development of the unselective GSI semagacestat (Figure 17) that reached phase III clinical trials.¹⁴⁶ The compound failed to slow progression of Alzheimer's disease. Furthermore, declines in cognitive function and a greater risk of skin cancer appeared as side effects. Consequently, NOTCH-sparing approaches of A β 42 reduction are of urgent need.

One of these alternative approaches besides GS inhibition is GS modulation. γ -Secretase modulators (GSMs) cause a product shift during APP processing at the expense of A β 42 to shorter and non-toxic fragments, such as A β 38.¹⁴⁷ Importantly, they do not influence the processing of other GS substrates.¹⁴⁸⁻¹⁵⁰ Four examples of GSMs characterized by their A β 42 inhibitory concentration 50% (A β 42 IC₅₀) are reported in Table 20. Compounds 5-7 are 'non-steroidal anti-inflammatory drugs' (NSAIDs) that constituted the first class of GSMs. They combine a carboxylic head group with lipophilic aromatic substituents. They exhibit

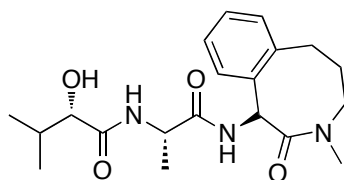
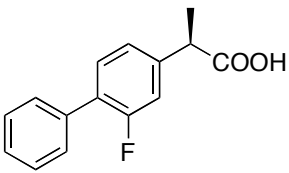
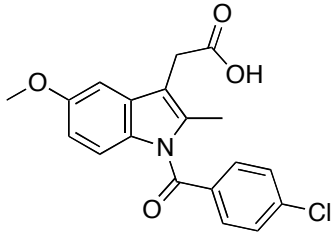
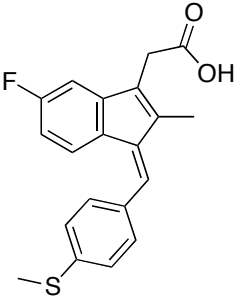
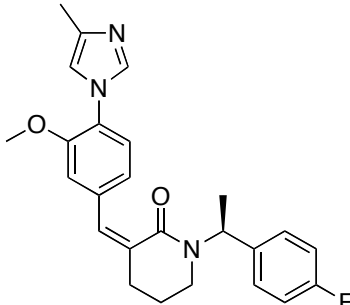


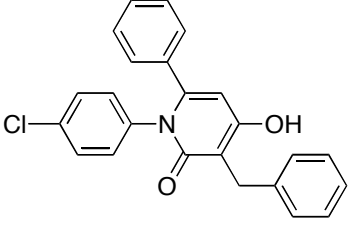
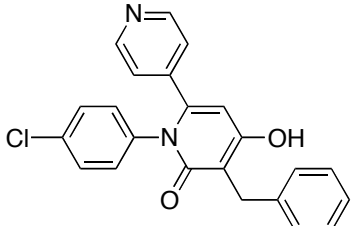
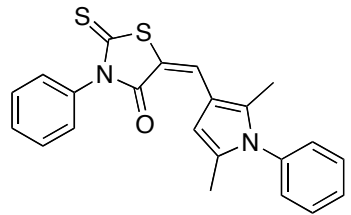
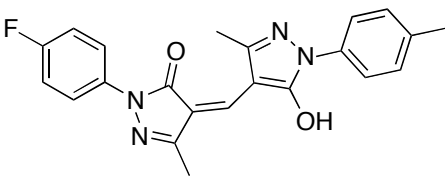
Figure 17. Structure diagram of semagacestat. This γ -Secretase inhibitor was developed by Eli Lilly. Development was stopped in 2010 after phase III clinical trials revealed its inability to stop the progression of Alzheimer's disease.

Table 20. Known γ -Secretase modulators and their in vitro activity. Activities for compounds 5-7 according to Peretto *et al.* 2008,¹⁵¹ activity for compound 8 according to Kimura *et al.* 2005.¹⁵²

	Structure	IC ₅₀ A β 42 [μ M]
5	 <p>(R)-Flurbiprofen</p>	305
6	 <p>Indomethacin</p>	25-50
7	 <p>Sulindac sulfide</p>	25-50
8		0.065

only weak inhibition of A β 42 production, but are reported to be highly active against cyclooxygenase (COX).¹⁴⁷ As a consequence, their long-term use is associated with COX-mediated side effects such as gastrointestinal ulceration and increased cardiovascular morbidity. Despite its weak activity (305 mM), compound 5 ((R)-Flurbiprofen) reached phase III clinical development, where the compound failed to show any beneficial effects. Its low potency and weak blood-brain-barrier permeability are discussed as major reasons for the failure.^{153,154} These problems with NSAIDs therapeutically applied as GSMs emphasize the need for not-NSAID-like GSMs.¹⁵⁵

Table 21. Results of virtual screening with PhAST for GSMs. ‘-‘ indicates no activity at 100 μ M.

	Structure	Rank	IC ₅₀ A β 42 [μ M]
9		12	-
10		13	10
11		20	-
12		77	-

PhAST was used in a ligand-based virtual screening for GSMs with the best-performing parameterization identified in this study: Canonization through Minimum Volume Embedding employing the Diffusion Kernel with diffusion parameter 0.4 and covalent connectivity, gap open penalty = 5, gap extension penalty = 1 and the alignment score normalized to alignment length for alignment evaluation. The screening library combined the vendor catalogues of Specs (v01/2010, Specs, Delft, The Netherlands) and Asinex Gold and Platinum collections (v11/2008, Asinex, Moscow, Russia). All compounds were protonated using the ‘wash‘ function of MOE. Compound 8 listed in Table 20 was used as query. From the resulting ranked list, four compounds were selected and evaluated. Activity was determined in an ELISA as described elsewhere.¹⁵⁶ The results are presented in Table 21.

Three of the four tested compounds (9, 11 and 12) were inactive at 100 μ M. Compound 10, that was ranked thirteenth, exhibited inhibition of A β 42 production with an IC₅₀ of 10 μ M without influencing A β 40 and A β 38. It is topologically identical to the inactive compound 9. The molecular graphs differ only in the exchange of an aromatic carbon atom to

a nitrogen atom from 9 to 10, indicating a steep structure-activity relationship. As a consequence, compounds 9 and 10 were assigned adjacent ranks with identical scores due to the fact that the graphs of potential pharmacophoric points created during the atom-typing step of PhAST are identical. Both are based on a bipyridine framework that is substituted with two additional aromatic ring systems. None of them has a carboxylic head group known from NSAIDs. Furthermore, they diverge from the linear assembly of four ring systems that can be observed in non-acidic GSMs.¹⁵⁵ These findings show that the identified active compound is pharmacologically disjunct from other classes of GSMs known so far.

Compound 10 has a 30-fold increased activity compared to compound 5 that reached phase III clinical trials. As compound 8 shows, molecules with even higher activity are known. But those are highly optimized structures. PhAST is meant as aid in the hit-identification phase of the drug development process. Promising compounds will be subject to further optimization in following stages. As hit, compound 10 is interesting because of its new pharmacology, and with its already acceptable activity it has high potential for optimization. PhAST succeeded in the identification of a promising GSM with a novel chemotype.

12 - Conclusions

This work prove that the concept of molecule comparison by global sequence alignment, until now applied to sequences of amino acids and nucleic acids, can be successfully used for the comparison small drug-like molecules. The obstacle of proteins and nucleic acids being linear and directed structures in contrast to small molecules that contain branches and cyclic systems was overcome by canonical atom labeling. Meaningful scoring systems for potential pharmacophoric points were calculated based on concepts developed for amino acids. The implementation of these methods lead to the development of the Pharmacophore Alignment Search Tool (PhAST). This concept was successfully applied to prospective screening scenarios.

This work investigated several concepts for graph linearization. Algorithms most suitable for PhAST were originally developed for dimensionality reduction. To the best of knowledge, this work represents the first application of these methods to the problem of graph linearization. The superiority of these methods to algorithms developed for calculating canonical sets of atom labels was significant.

It was shown that screening performance of PhAST is comparable to that of methods already applied in drug development campaigns. At the same time, PhAST was complementary to those methods, meaning that it ranked compounds in a unique order. This way, it can be applied to prospective scenarios along with other methods, generating a diverse set of hit candidates. The incorporation of knowledge about ligand-receptor interactions in the screening process by the application of position weights in the query sequence significantly increased screening performance of PhAST compared to the standard version. The concept of positional weights was used for pharmacophore elucidation: the determination of key interactions common to a diverse set of active compounds.

Double Dynamic Programming was developed to calculate alignments of amino acid sequences based on structural similarity. In this work, this technique was successfully transferred to the comparison of line notations of small molecules. This way it was shown that a linear molecular representation is sufficient for shape comparison.

The significance of similarity scores is of great importance, as the exclusion of insignificant hits obtained in a virtual screening from subsequent activity assessment can reduce time and cost of early stages in drug design campaigns. For the most methods used in

drug discovery, there are no proposed or established ways of significance estimation. PhAST employs sequence alignment for similarity assessment, where this problem has been investigated for nearly 40 years. As a consequence, methods developed for significance assessment of amino acid alignment scores were re-used for alignments of small molecule line notations. In this work, a technique for the sampling of rare events was applied successfully to the determination of alignment score distributions of PhAST-sequences. These distributions were used for the calculation of p-values, yielding E-values after a Bonferroni correction for library size. This way, PhAST can be used for the identification of significant hits.

PhAST was applied successfully in prospective screening campaigns. In both applications, to γ -Secretase and bacterial Thymidinkinase, active compounds with a structural distinction to the query structure were identified. These results prove that PhAST is a suitable and valuable method for the identification of diverse hits in the early stages of drug development campaigns.

Besides their impact for the development of PhAST, the results of this work disclosed general coherences important to virtual screening and drug design in general. The incorporation of ligand-receptor interactions in the screening process has high impact and helps to build more realistic models. There is no single-best screening method for all drug targets. The application of only one screening technique is inferior compared to data fusion approaches or screening cascades. As illustrated in the thymidine kinase project, the identification and refinement of hits with a diverse set of methods can lead to more potent compounds than one method alone. This is due to the multifaceted nature of biological activity.

This work evidenced the capabilities of text-based virtual screening and the effects of alterations to its components. The calculation of significance estimates of similarity scores, the flexible scoring scheme, the possibility to apply weights to key interaction, its pharmacophore elucidation capabilities and the unique rank order of compounds set PhAST apart from other screening techniques. Because of these reasons, PhAST has the potential to be a valuable asset to any drug development campaign.

13 - Outlook

During the development of PhAST it was shown that sequence alignment, developed for the comparison of amino acid sequences and nucleic acids, can be applied to the comparison of drug-like molecules as well. This analogy was not limited to the general concept of sequence alignment but applied in a variety of specific findings such as:

- the alignment score is an indicator for similarity superior to percent sequence identity,
- sequence alignment can be used for the comparison of three-dimensional structures,
- the efficient calculation of p-values is possible through Markov Chain Monte Carlo methods.

These pronounced similarities give plausible reason to the hope that improvements to sequence alignment and special variants of this technique developed for the application to amino acid sequences and nucleic acids are applicable to PhAST as well.

Sequence alignment algorithms can be adapted to specific hardware features. An implementation tailored to the Intel SSE2 extension calculated alignments 13 times faster.¹⁵⁷ Using field programmable gate arrays (FPGAs), 160-fold acceleration was observed.^{158,159} Sequence alignment is an embarrassingly parallel problem, as different sequence pairs can be aligned on different processing units without any necessary interprocess communication. Several implementations of sequence alignment algorithms that exploit this fact have been reported. Using the multi-core Cell processor embedded in the Sony Playstation 3, 36-fold speedup was achieved.¹⁶⁰ Graphics processing units (GPUs) can be used for parallelization as well, due to their large number of highly specialized processors. Using the OpenGL interface calculations could be accelerated by factor 5.¹⁶¹ An adaption to the ‘compute unified device architecture’ (CUDA) resulted in reported increases of 2-30-fold and 23-fold.^{162,163} Parallelization cannot only be achieved by distributing sequence alignment tasks to a large number of processing units. Another possibility is the modification of the alignment algorithm to calculate single alignments using several processors in parallel. So far, this strategy resulted in 2- to 16-fold increase depending on sequence length.¹⁶⁴ Besides parallelization alignment algorithms can be improved by the usage of ‘query-profiles’ that circumvent the time-consuming look-up of scores in the score matrix.^{157,165} Using heuristics instead of exact solutions for the optimal alignment problem results in 50-fold increase.¹⁶³ PhAST could benefit from increased throughput in two ways: First, the number of parameterizations that

can be evaluated in retrospective screenings would increase. Especially with regard to time-consuming calculations such as Double Dynamic Programming and Rare Event Sampling this would be a major improvement. Second, the number of screened compounds in prospective applications could be increased. Again, Double Dynamic Programming and significance estimation through Rare Event Sampling would benefit most, because these calculations take hours to days in the current implementation.

The capabilities of PhAST to compare three-dimensional molecular structures through Double Dynamic Programming have to be further investigated. The current implementation cannot differentiate between atoms that diverge in their directional component. For protein structures it has been reported that accounting for directional differences improves equivalency detection. The second level of dynamic programming could be eliminated if spatial equivalence was evaluated with a measure different from sequence alignment. This would significantly reduce computational costs. In addition to calculations of molecular similarity, matches in the generated alignments could be used as seeds in the calculation of molecular alignments.^{166,167}

Homologue searches applied to amino acid sequences can be refined in an iterative process that calculates a position specific score matrix for the query sequence, as shown with the 'Position-Specific Iterated Basic Local Alignment Search Tool' (PSI-BLAST).¹⁶⁸ Sequences considered for position-specific score calculations are selected based on a significance threshold. The availability of p-values and E-values permits such strategies in PhAST as well, if meaningful thresholds can be determined. Threshold determination goes hand in hand with further investigation of the significance of global alignments of PhAST-sequences. The impact of symbol frequency adaption to screening libraries as well as compound classes and effects of alternative sampling methods should be investigated for that matter.

So far multiple sequence alignment⁹⁶ seemed infeasible for pharmacophore elucidation because of the large number of possible sequence orientation combinations and because of multiple sequence alignment being a NP-complete problem.¹⁶⁹⁻¹⁷¹ A simple heuristic could solve this problem: One sequence is defined as origin, the remaining sequences are integrated in the multiple sequence alignment in the orientation that yields highest alignment score to the first one.

This work investigated pairwise optimal global sequence alignments for the comparison of drug-like molecules. For nucleotide or amino acid sequences the optimal alignment may not necessarily reflect the correct biological alignment. Since in most cases the

true alignment is unknown, methods that generate ‘suboptimal’ alignments close to the optimal one have been developed.¹⁷²⁻¹⁷⁵ An alignment score within a certain score difference to the optimal alignment characterizes these suboptimal alignments. Future studies could investigate the applicability of suboptimal alignments to the comparison of drug-like molecules.

The current implementation of PhAST compares linear representations that are created from molecules in one single step. The disadvantage of this approach is that features from different domains of a molecule might end up represented by adjacent symbols in the corresponding line notation. This behavior makes PhAST a non-additive similarity function. A fragment-based transformation of molecules to line notations would constrain this effect locally. Such a ‘Fragment Alignment Search Tool’ (FAST) could increase the sensitivity for local similarity. An important variable in this concept is the fragmentation strategy. A strict scaffold-and-side-chain-based set of rules as proposed by Bemis and Murcko^{176,177} requires the possibility of meaningful prioritization between sidechains. Alternatives are substructure prioritizations analogue to systematic compound name generation¹⁷⁸ or the step-wise decomposition of molecular structures as proposed by Schuffenhauer *et al.*¹⁷⁹ If preservation of locality is strong enough, the application of local sequence alignment as similarity measure might be possible.¹⁸⁰

Algorithms utilized for canonization so far either were developed for molecular graphs or belong to the field of dimensionality reduction. A field not investigated so far with similarities to dimensionality reduction is the projection of vertices to monster curves, such as the Hilbert space-filling curve.¹⁸¹ This approach might circumvent disadvantages observed in Principal Component Analysis, that places vertices with large distances in between orthogonal to the principal component adjacent in the one-dimensional projection.

The identified modulator of γ -Secretase and its inactive structural analogue point to a weakness in the pharmacophore model of PhAST as they have identical graphs of potential pharmacophoric points: Heteroatoms in aromatic rings get assigned no interaction possibility besides aromaticity. In addition, only recently,¹⁸² a statistical evaluation of hydrogen-bond donors and acceptors has been published than can be incorporated in a more meaningful pharmacophore model.

The multitude of aspects and concrete starting points for future investigation underline the flexibility of the PhAST concept and give confidence that this approach can be further enhanced in future studies.

14 - List of Publications

This section lists the contributions of Volker Dirk Hähnke to publications that are part of this cumulative dissertation.

- (1) Hähnke, V.; Rupp, M.; Krier, M.; Rippmann, F.; Schneider, G. (2010) Pharmacophore Alignment Search Tool: Influence of Canonical Atom Labeling on Similarity Searching, *Journal of Computational Chemistry* **31**, 2810-2826.
 - Development of the concept to use dimensionality reduction for canonical atom labeling
 - Implementation of canonization algorithms: Weininger, Prabhakar, Jochum-Gasteiger, Principal Component Analysis, Laplacian Eigenmaps, Isomap, Minimum Volume Embedding
 - Implementation of kernels: Diffusion Kernel, Euclidean distance kernel
 - Implementation of modified Needleman-Wunsch algorithm and FSM algorithm for sequence alignment
 - Development and implementation of the concept to use rank correlation for screening result comparison
 - Development and execution of the idea to compare alignment algorithms by correlation of retrospective performance obtained with PhAST
 - Implementation of percent sequence identity and alignment score variants
 - Development, execution and analysis of the test for canonization necessity
 - Development, execution and analysis of the test for canonization robustness through resistance to structure modifications, selection of modification fragments
 - Implementation of the BEDROC metric for enrichment assessment and the paired permutation test
 - Molecule preparation: determination of protonation states, calculation of 2D layouts
 - Execution of all retrospective screenings and assessment of enrichment through BEDROC scores
 - Design of all figures and tables in the manuscript
 - Draft of the complete manuscript

(2) Hähnke, V.; Klenner, A.; Rippmann, F.; Schneider, G. Pharmacophore Alignment Search Tool: Influence of the Third Dimension on Text-based Similarity Searching, *Journal of Computational Chemistry*, accepted.

- Implementation of canonization algorithms: Weininger, Prabhakar, Jochum-Gasteiger, Principal Component Analysis, Laplacian Eigenmaps, Isomap, Minimum Volume Embedding
- Implementation of kernel: Diffusion Kernel, Euclidean distance kernel, Gaussian radial basis function kernel, p-step random walk kernel
- Implementation of the modified Needleman-Wunsch and FSM algorithm for sequence alignment
- Implementation of the LINGO approach for virtual screening
- Development, implementation and execution of the concept to use Levenshtein- and Damerau-Levenshtein distance for similarity assessment of canonization algorithms as well as 2D and 3D representations of molecules
- Development, implementation and execution of the concept to use averaged rank correlation for screening method comparison
- Implementation and execution of the data fusion approach
- Implementation of Double Dynamic Programming and development as well as execution of parameterization concepts
- Implementation of the BEDROC metric for enrichment assessment and the paired permutation test
- Development, implementation and execution of the concept used for calculation of a symmetric distance matrix based on asymmetric significance estimations
- Molecule preparation: determination of protonation states, calculation of 2D layouts, calculation of single 3D conformation
- Execution of all retrospective screenings and assessment of enrichment through BEDROC scores with PhAST as well as every other screening method
- Analysis of screening runtimes
- Execution and analysis of all clusterings
- Design of all figures and tables in the manuscript
- Draft of the complete manuscript (except for the subsection comparing datasets by scaffold diversity)

(3) Hähnke, V.; Schneider, G. Pharmacophore Alignment Search Tool: Influence of Scoring Systems on Text-based Similarity Searching, *Journal of Computational Chemistry*, accepted.

- Implementation of Minimum Volume Embedding and Diffusion Kernel employed for canonization
- Implementation of the BEDROC metric for enrichment assessment and the paired permutation test
- Development, implementation and execution of the concept to use constrained pairwise sequence alignments for the calculation of alignment scores applicable to potential pharmacophoric points
- Development, implementation (except for the ISOA kernel itself) and execution of the concept to use constrained pairwise assignments for the calculation of alignment scores applicable to potential pharmacophoric points
- Parameterization of log-odds-score calculations
- Development, implementation, parameterization and execution of the stochastic optimization applicable to the problem of optimizing alignment scores of potential pharmacophoric points
- Development and implementation of the concept of weighted positions in the query sequence
- Development and implementation of the concept to use systematic weighted screenings for pharmacophore elucidation
- Molecule preparation: determination of protonation states
- Execution of all retrospective screenings and assessment of enrichment through BEDROC scores
- Similarity assessment of molecules by structural fingerprints
- Execution of all paired permutation tests for significance assessment
- Analysis of score matrices obtained with the proposed calculation strategies and their relations
- Analysis of weight sets obtained from systematic weighted retrospective screenings for pharmacophore elucidation
- Design of all figures and tables in the manuscript
- Draft of the complete manuscript

(4) Zander, J.; Hartenfeller, M.; Hähnke, V.; Proschak, E.; Besier, S.; Wichelhaus, T. A.; Schneider, G. (2010) Multistep Virtual Screening for Rapid and Efficient Identification of Non-Nucleoside Bacterial Thymidine Kinase Inhibitors, *Chemistry – a European Journal* **16**, 9630-9637.

- Preceding parameterization of PhAST and implementation of all employed algorithms: canonization (Prabhakap algorithm, Isomap), FSM algorithm for sequence alignment
- Molecule preparation: determination of protonation states
- Calculation of all 184 2D descriptors of MOE (Molecular Operating Environment, v2008.10, Chemical Computing Group, Montreal, QC, Canada)
- Execution of a Principal Component Analysis to the calculated 2D descriptors
- Selection of the first 40 principal components
- Execution of prospective screenings with PhAST
- Compound selection after the first screening level
- Substructure search in screening library based on results of the first screening level
- Compound selection after the second screening level

15 - References

- 1 DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. (2003) The price of innovation: new estimates of drug development costs, *J. Health Econ.* **22**, 151-185.
- 2 Rankovic, Z.; Morphy, R. *Lead Generation Approaches in Drug Discovery*, John Wiley & Sons Inc., Hoboken, New Jersey (2010).
- 3 Adams, C. P.; Brantner, V. V. (2006) Estimating The Cost Of New Drug Development: Is It Really \$802 Million?, *Health Affair.* **25**, 420-428.
- 4 Michne, W. F. (1996) Hit-to-lead chemistry: a key element in new lead generation, *Pharmaceutical News* **3**, 19-21.
- 5 Bleicher, K. H.; Böhm, H. J.; Müller, K.; Alanine, A. I. (2003) Hit And Lead Generation: Beyond High Throughput Screening, *Nat. Rev. Drug Discovery* **2**, 369-378.
- 6 PhRMA Pharmaceutical Industry Profile 2003, Chapter 1: Increased Length and Complexity of the Research and Development Process.
- 7 Alper, J. (1994) Drug discovery on the assembly line, *Science* **264**, 1399-1401.
- 8 Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases, *J. Chem. Inf. Model.* **46**, 1124-1133.
- 9 Heinrich, R.; Mamitsuka, H.; Kanehisa, M.; Miyano, S.; Tagaki, T. *Genome Informatics 2005*, Universal Academy Press, Tokyo (2010), pp. 281-285.
- 10 Bohacek, R. S.; McMartin, C.; Guida, W. C. (1996) The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective, *Med. Res. Rev.* **16**, 3-50.
- 11 Schneider, G.; Schneider, P. (2003) Collection of Bioactive Reference Compounds for Focused Library Design, *QSAR Comb. Sci.* **22**, 713-718.
- 12 Gasteiger, J. (2006) Chemoinformatics: A new field with a long tradition, *Anal. Bioanal. Chem.* **384**, 57-64.
- 13 Truchon, J. F.; Bayly, C. I. (2007) Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem, *J. Chem. Inf. Model.* **47**, 488-508.
- 14 Böhm, H. J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg (2002).

- 15 Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug. Del. Rev.* **46**, 3-26.
- 16 Ghose, A. K.; Viswanadhan, V. N.; Wenoloski, J. J. (1999) A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases, *J. Comb. Chem.* **1**, 55-68.
- 17 Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. (2001) Is There a Difference between Leads and Drugs? A Historical Perspective, *J. Chem. Inf. Comput. Sci.* **41**, 1308-1315.
- 18 Lipinski, C. A. (2004) Lead- and drug-like compounds: the rule-of-five revolution, *Drug. Discov. Today* **1**, 337-341.
- 19 Klebe, g. (2006) Virtual ligand screening: Strategies, perspectives and limitations, *Drug. Discov. Today* **11**, 580-594.
- 20 Johnson, A. M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York (1990).
- 21 Maggiora, G. M. (2006) On Outliers and Activity Cliffs - Why QSAR Often Disappoints, *J. Chem. Inf. Model.* **46**, 1535.
- 22 Eckert, H.; Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches, *Drug Discov. Today* **12**, 225-233.
- 23 Peltason, L.; Bajorath, J. (2007) SAR Index: Quantifying the Nature of Structure-Activity Relationships, *J. Med. Chem.* **50**, 5571-5578.
- 24 Guha, R.; Van Drie, J. H. (2008) Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs, *J. Chem. Inf. Model.* **48**, 646-658.
- 25 Douguet, D. (2008) Ligand-Based Approaches in Virtual Screening, *Curr. Comput.-Aided Drug Des.* **4**, 180-190.
- 26 Harris, C. J.; Stevens, A. P. (2006) Chemogenomics: structuring the drug discovery process to gene families, *Drug Discov. Today* **11**, 880-888.
- 27 Schneider, G.; Böhm, H. J. (2002) Virtual screening and fast automated docking methods, *Drug Discov. Today* **7**, 64-70.
- 28 King, R. B. *Chemical Applications of Topology and Graph Theory*, Elsevier Science Ltd., Amsterdam (1983).

- 29 Xu, J. (1996) GMA: A Generic Match Algorithm for Structural Homomorphism, Isomorphism, and Maximal Common Substructure Match and Its Applications, *J. Chem. Inf. Comput. Sci.* **36**, 25-34.
- 30 Raymond, J. W.; Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures, *J. Comput. Aided Mol. Des.* **16**, 521-533.
- 31 Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH Verlag GmbH, Weinheim (2000).
- 32 Gasteiger, J.; Engel, T. *Cheminformatics A Textbook*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim (2003).
- 33 Willett, P.; Barnard, J. M.; Downs, G. M. (1998) Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.* **38**, 983-996.
- 34 Kier, L. B. *Molecular Orbital Theory in Drug Research*, Academic Press, New York (1971), pp. 164-169.
- 35 Wermuth C., Ganellin C., Lindberg P., Mitscher L. (1998) Glossary of terms used in medicinal chemistry (IUPAC recommendations 1998), *Pure. Appl. Chem.* **70**, 1129–1143.
- 36 Schneider, G.; Baringhaus, K. H. *Molecular Design Concepts and Applications*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim (2008), pp. 58-62.
- 37 Schwarz, O., Jakupovic, S.; Ambrosi, H. D.; Haustedt, L. O.; Mang, C.; Müller-Kuhrt, L. (2007) Natural Products in Parallel Chemistry - Novel 5-Lipoxygenase Inhibitors from BIOS-Based Libraries Starting from a-Santonin, *J. Comb. Chem.* **9**, 1104-1113.
- 38 Franke, L.; Schwarz, O.; Müller-Kuhrt, L.; Hoernig, C.; Fischer, L.; George, S.; Tanrikulu, Y.; Schneider, P.; Werz, O.; Steinhilber, D.; Schneider, G. (2007) Identification of natural-product-derived inhibitors of 5-lipoxygenase activity by ligand-based virtual screening, *J. Med. Chem.* **50**, 2640-2646.
- 39 Wiswesser, W. J. *A Line-Formula Chemical Notation*, Thomas Crowell, New York (1954).
- 40 Wiswesser, W. J. (1982) How the WLN Began in 1949 and How It Might Be in 1999, *J. Chem. Inf. Comput. Sci.* **22**, 88-93.
- 41 Smith, E. G. *The Wiswesser Line-Formula Chemical Notation*, McGraw Hill Inc., New York (1983).
- 42 Granito, C. E.; Rosenberg, M. D. (1971) The Chemical Substructure Index (CSI), a New Research Tool, *J. Chem. Doc.* **11**, 251-256.

- 43 Ash, J. E.; Hyde, E. *Chemical Information Systems*, Ellis Horwood, Chichester (1975).
- 44 Barnard, J. M.; Jochum, C. J.; Welford, S. M. *ACS Symposium Series 400*, American Chemical Society, Washington DC (1989), pp. 76-81.
- 45 Weininger, D. (1988), SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* **28**, 31-36.
- 46 Weininger, D.; Weininger, A.; Weininger, J.L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation *J. Chem. Inf. Comput. Sci.* **29**, 97-101.
- 47 McNaught, A. (2006) The IUPAC international chemical identifier: InChI - A new standard for molecular informatics, *Chemistry International* **28**, 12-14.
- 48 Stein, S. E.; Heller, S. E.; Tchekhovskoi, D. V. *The IUPAC Chemical Identifier - Technical Manual*, National Institute of Standards and Technology, Gaithersburg, Maryland (2006).
- 49 Federal Information Processing Standards Publication 180-2, National Institute of Standards and Technology, 2002.
- 50 Feldman, R. J.; Koniver, D. A. (1971) Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation, *J. Chem. Doc.* **11**, 154-159.
- 51 Heller, S. R.; Koniver, D. A. (1972) Computer Generation of Wiswesser Line Notation. II. Polyfused, Perifused, and Chained Ring Systems, *J. Chem. Doc.* **12**, 55-59.
- 52 Miller, G. A. (1972) Encoding and Decoding WLN, *J. Chem. Doc.* **12**, 60-67.
- 53 Vidal, D.; Thormann, M.; Pons, M. (2005) LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities, *J. Chem. Inf. Model.* **45**, 386-393.
- 54 Vidal, D.; Thormann, M.; Pons, M. (2006) A Novel Search Engine for Virtual Screening of Very Large Databases, *J. Chem. Inf. Model.* **46**, 836-843.
- 55 Ming, J.; Vitanyi, P. *An Introduction to Kolmogorov Complexity and its Applications*, Springer, New York (1997).
- 56 Li M., Chen X., Li X., Ma B., Vitanyi P. (2004) The Similarity Metric, *IEEE Trans. Inf. Theory* **50**, 3250-3264.
- 57 Ziv J., Lempel A. (1977) A Universal Algorithm for Sequential Data Compression, *IEEE Trans. Inf. Theory* **23**, 337-343.

- 58 Huffman D. (1952) A Method for the Construction of Minimum Redundancy Codes, *Proceedings of the IRE* **40**, 1098-1101.
- 59 Melville, J. L.; Riley, J. F.; Hirst, J. D. (2007) Similarity by Compression, *J. Chem. Inf. Model.* **47**, 25-33.
- 60 Levenshtein, V. I. (1966) Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady* **10**, 707-710.
- 61 Damerau, F. J. (1964) A technique for computer detection and correction of spelling errors, *Communications of the ACM* **7**, 171-176.
- 62 Dice, L. R. (1945) Measures of the Amount of Ecologic Association Between Species, *Ecology* **26**, 297-302.
- 63 Proschak, E.; Wegner, J. K.; Schüller, A.; Schneider, G.; Fechner, U. (2007) Molecular Query Language (MQL) - A Context-Free Grammar for Substructure Matching, *J. Chem. Inf. Model.* **47**, 295-301.
- 64 Needleman, S. B.; Wunsch, C. D. (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *J. Mol. Biol.* **48**, 443-453.
- 65 Waterman, M. S.; Smith, T. F.; Beyer, W. A. (1976) Some Biological Sequence Metrics, *Adv. Math.* **20**, 367-387.
- 66 Gusfield, D. *Algorithms on Strings, Trees, and Sequences Computer Science and Computational Biology*, Cambridge University Press, New York (1997).
- 67 Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington DC (1978), pp. 345-358.
- 68 Henikoff, S.; Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. USA* **89**, 10915-10919.
- 69 Hähnke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. (2009) PhAST: Pharmacophore Alignment Search Tool, *J. Comput. Chem.* **30**, 761-771.
- 70 Fechner, U.; Schneider, G. (2004) Optimization of a Pharmacophore-based Correlation Vector Descriptor for Similarity Searching, *QSAR Comb. Sci.* **23**, 19-23.
- 71 Doolittle, R. F. (1981) Similar Amino Acid Sequences: Chance or Common Ancestry?, *Science* **214**, 149-159.
- 72 Rost, B. (1999) Twilight Zone of protein sequence alignments, *Protein Eng.* **12**, 85-89.

- 73 Raghava, G. P. S.; Barton, G. J. (2006) Quantification of the variation in percentage identity for protein sequence alignments, *BMC Bioinformatics* **7**, 415-418.
- 74 Hawkins, P.; Warren, G.; Skillman, G.; Nicholls, A. (2008) How to do an evaluation: pitfalls and traps, *J. Comput. Aided Mol. Des.* **22**, 179-190.
- 75 Hanley, J. A.; McNeil, B. J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve, *Diag. Radiol.* **143**, 29-36.
- 76 Swamidass, S. J.; Azencott, C. A.; Daily, K.; Baldi, P. (2010) A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval, *Bioinformatics* **26**, 1348-1356.
- 77 Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. (2009) A statistical framework to evaluate virtual screening, *BMC Bioinformatics* **10**, 225-237.
- 78 Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological sequence analysis*, Cambridge University Press, Cambridge UK (1998).
- 79 Jochum, C.; Gasteiger, J. (1977) Canonical Numbering and Constitutional Symmetry, *J. Chem. Inf. Comput. Sci.* **17**, 113-117.
- 80 Prabhakar, Y. S.; Balasubramanian, K. (2006), A Simple Algorithm for Unique Representation of Chemical Structures - Cyclic / Acyclic Functionalized Achiral Molecules, *J. Chem. Inf. Model.* **46**, 52-56.
- 81 Pearson, K. (1901) On lines and planes of closest fit to systems of points in space, *Phil. Mag.* **2**, 559-572.
- 82 Belkin, M.; Niyogi, P. (2003) Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Comput.* **15**, 1373-1396.
- 83 Tenenbaum, J. B.; de Silva, V.; Langford, J. C. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science* **290**, 2319-2323.
- 84 Shaw, R.; Jebara, T. "Minimum Volume Embedding" in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, Omnipress, Madison (2007).
- 85 Brenner, S. E.; Chothia, C.; Hubbard, T. J. P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc. Nat. Acad. Sci. USA* **95**, 6073-6078.
- 86 Bundschuh, R. (2002) Rapid Significance Estimation in Local Sequence Alignment with Gaps, *J. Comput. Biol.* **9**, 243-260.
- 87 Newberg, L. A. (2008) Significance of Gapped Sequence Alignments, *J. Comput. Biol.* **15**, 1187-1194.

- 88 Eddy, S. R. (2008) A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation, *PLoS Comput. Biol.* **4**(5).
- 89 Hartmann, A. (2001) Sampling rare events: Statistics of local sequence alignments, *Phys. Rev. E* **65**, 1-4.
- 90 Sheridan, R. P.; Kearsley, S. K. (2002) Why do we need so many chemical similarity search methods?, *Drug Discov. Today* **7**, 903-911.
- 91 Bajorath, J. (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening, *J. Chem. Inf. Comput. Sci.* **41**, 233-245.
- 92 Taylor, W. R.; Orengo, C. A. (1989) Protein Structure Alignment, *J. Mol. Biol.* **208**, 1-22.
- 93 Kendall, M. (1938) A New Measure of Rank Correlation, *Biometrika* **30**, 81-89.
- 94 Ginn, C. M. R.; Willett, P.; Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion, *J. Perspect. Drug. Discov.* **20**, 1-16.
- 95 Tanrikulu, Y.; Proschak, E.; Werner, T.; Geppert, T.; Todoroff, N.; Klenner, A.; Kottke, T.; Sander, K.; Schneider, E.; Seifert, R.; Stark, H.; Clark, T.; Schneider, G. (2008) Homology Model Adjustment and Ligand Screening with a Pseudoreceptor of the Human Histamine H4 Receptor, *ChemMedChem* **4**, 820-827.
- 96 Carrillo, H.; Lipman, D. D. (1988) The Multiple Sequence Alignment Problem, *SIAM Journal of Applied Mathematics* **48**, 1073-1082.
- 97 Rupp, M.; Proschak, E.; Schneider, G. (2007) Kernel approach to molecular similarity based on iterative graph similarity, *J. Chem. Inf. Model.* **47**, 2280-2286.
- 98 Rohrer, S. G.; Baumann, K. (2009) Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data, *J. Chem. Inf. Model.* **49**, 169-184.
- 99 Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. (2009) Critical Comparison of Virtual Screening Methods against MUV Data Set, *J. Chem. Inf. Model.* **49**, 2168-2178.
- 100 Kier, M.; Hutter, M. C. (2009) Bioisosteric Similarity of Molecules based on Structural Alignment and Observed Chemical Replacements in Drugs, *J. Chem. Inf. Model.* **49**, 1280-1297.
- 101 Köppen, H. (2009) Virtual screening - what does it give us? *Curr. Opin. Drug Discovery Dev.* **12**, 397-407.

- 102 Jaccard, P. (1901) Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines, *Bull. Soc. Vaudoise Sci. Nat.* **37**, 241-272.
- 103 Daylight Manual, Chapter 6 - Fingerprints, Daylight Chemical Information Systems Inc., Cabot Road, Laguna Niguel, California USA.
- 104 Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **45**, 4350-4358.
- 105 Baldie, P.; Nasr, R. (2010) When is chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and its Extreme Values, *J. Chem. Inf. Model.* **50**, 1205-1222.
- 106 Chia, N.; Bundschuh, R. (2006) A Practical Approach to Significance Assessment in Alignment with Gaps, *J. Comp. Biol.* **13**, 429-441.
- 107 Hähnke, V.; Rupp, M.; Krier, M.; Rippmann, F.; Schneider, G. (2010) Pharmacophore Alignment Search Tool: Influence of Canonical Atom Labeling on Similarity Searching, *J. Comput. Chem.* **31**, 2810-2826.
- 108 Karlin, S.; Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci. USA* **87**, 2264-2268.
- 109 Karlin, S.; Dembo, A. (1992) Limit distributions of the maximal segmental score among Markov-dependent partial sums, *Adv. Appl. Prob.* **24**, 113-140.
- 110 Karlin, S.; Altschul, S. F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences, *Proc. Natl. Acad. Sci.* **90**, 5873-5877.
- 111 Gumbel, E. J. *Statistics of Extremes*, Columbia University Press, New York (1958).
- 112 Yu, Y; Hwa, T. (2001) Statistical Significance of Probabilistic Sequence Alignment and Related Local Hidden Markov Models, *J. Comp. Biol.* **8**, 249-282.
- 113 Kschischo, M.; Lässig, M.; Yu, Y. (2004) Toward an accurate statistics of gapped alignments, *Bull. Math. Biol.* **67**, 169-191.
- 114 Sankoff, D.; Kruskal, J. B. *Time Warps, String Edits and Macromolecules: The Theory and Practice Sequence Comparison*, Addison-Wesley, Reading MA USA (1983), pp. 55-91.
- 115 Lilliefors, H. (1967) On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown, *J. Amer. Statistical Assoc.* **62**, 399-402.
- 116 Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. (1953) Equations of State Calculations by Fast Computing Machines, *J. Chem. Phys.* **21**, 1087-1092.

- 117 Hastings, W. K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika* **57**, 97-109.
- 118 Wolfsheimer, S.; Burghardt, B.; Hartmann, A. K. (2007) Local sequence alignments statistics: deviations from gumbel statistics in the rare-event tail, *Algorithm. Mol. Biol.* **2**(9).
- 119 Siegmund, D. (1976) Importance sampling in the Monte Carlo study of sequential tests, *Ann. Stat.* **4**, 673-684.
- 120 Levenberg, K. (1944) A Method for the Solution of Certain Non-Linear Problems in Least Squares, *Q. Appl. Math.* **2**, 164-168.
- 121 Natesh, R.; Schwager, S. L. U.; Sturrock, E. E.; Acharya, K. R. (2003) Crystal structure of the human angiotensin-converting enzyme-lisinopril complex, *Nature* **421**, 551-554.
- 122 Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilità, *Publicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3-62.
- 123 Neuwald, A. F.; Liu, J. S.; Lawrence, C. E. (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats, *Prot. Sci.* **4**, 1618-1632.
- 124 Hückel, E. (1931) Quantentheoretische Beiträge zum Benzolproblem I. Die Elektronenkonfiguration des Benzols und verwandter Verbindungen, *Z. Phys.* **70**, 204-286.
- 125 Hückel, E. (1931) Quantentheoretische Beiträge zum Benzolproblem II. Quantentheorie der induzierten Polaritäten, *Z. Phys.* **72**, 310-337.
- 126 Geyer, C. J. "Markoc chain Monte Carlo maximum likelihood" in *Proceedings of the 23rd Symposium on the Interface: Critical Applications of Scientific Computing: Biology, Engineering, Medicine, Speech*, Interface Foundation, Washington DC (1991), pp. 156-163.
- 127 Irwin, J. J.; Shoichet, B. K. (2005) ZINC - A Free Database of Commercially Available Compounds for Virtual Screening, *J. Chem. Inf. Model.* **45**, 177-182.
- 128 Kenney, J. F.; Keeping, E. S. *Mathematics of Statistics Pt. 2, 2nd Ed.*, Van Nostrand, Princeton, NJ (1951).
- 129 Wang, F.; Landau, D. P. (2001) Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States, *Phys. Rev. Lett.* **86**, 2050-2053.

- 130 Wang, F.; Landau, D. P. (2001) Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram, *Phys. Rev. E* **64**, 056101-1 - 056101-16.
- 131 Holm, S. (1979) A simple sequentially rejective multiple test procedure, *Scand. J. Stat.* **6**, 65-70.
- 132 Benjamini, Y.; Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.
- 133 Kitahara, T.; Aoyama, Y.; Hirakata, Y.; Kamihira, S.; Kohno, S.; Ichikawa, N.; Nakashima, M.; Sasaki, H.; Higuchi, S. (2006) In vitro activity of lauric acid or myristylamine in combination with six antimicrobial agents against methicillin-resistant *Staphylococcus aureus* (MRSA), *Int. J. Antimicrob. Agents* **26**, 51-57.
- 134 Zuo, G. Y.; Wang, G. C.; Zhao, Y. B.; Xu, G. L.; Hao, X. Y.; Han, J.; Zhao, Q. (2008) Screening of Chinese medicinal plants for inhibition against clinical isolates of methicillin-resistant *Staphylococcus aureus* (MRSA), *J. Ethnopharmacol.* **120**, 287-290.
- 135 Adra, M.; Lawrence, K. R. (2004) Trimethoprim/Sulfamethoxazole for Treatment of Severe *Staphylococcus aureus* Infections, *Ann. Pharmacoth.* **38**, 338-341.
- 136 Grim, S. A.; Rapp, R. P.; Martin, C. A.; Evans, M. E. (2005) Trimethoprim-Sulfamethoxazole as a Viable Treatment Option for Infections Caused by Methicillin-Resistant *Staphylococcus aureus*, *Pharmacotherapy* **25**, 253-264.
- 137 Kosinska, U.; Carnrot, C.; Eriksson, S.; Wang, L.; Eklund, H. (2005) Structure of the substrate complex of thymidine kinase from *Ureaplasma urealyticum* and investigations of possible drug targets for the enzyme, *FEBS J.* **272**, 6365 - 6372.
- 138 Voytek, P.; Chang, P. K.; Prusoff, W. H. (1972) Kinetic and Photochemical Studies of 3-N-Methyl-5-iodo-2'-deoxyuridine: SENSITIZATION OF ULTRAVIOLET INACTIVATION OF THYMIDINE KINASE BY 3-N-METHYL-5-IODO-2'-DEOXYURIDINE AND OTHER HALOGENATED ANALOGS OF THYMIDINE, *J. Biol. Chem.* **247**, 367-372.
- 139 Carnrot, C.; Vogel, S. R.; Byun, Y.; Wang, L.; Tjarks, W.; Eriksson, S.; Phipps, A. J. (2006) Evaluation of *Bacillus anthracis* thymidine kinase as a potential target for the development of antibacterial nucleoside analogs, *Biol. Chem.* **387**, 1575-1581.

- 140 Galanis, E.; Goldberg, R.; Reid, J.; Atherton, P.; Sloan, J.; Pitot, H.; Rubin, J.; Adjei, A. A.; Burch, P.; Safgren, S. L.; Witzig, T. E.; Ames, M. M.; Erlichmann, C. (2001) Phase I trial of sequential administration of raltitrexed (Tomudex) and 5-iodo-2'-deoxyuridine (IdUrd), *Ann. Oncol.* **12**, 701-707.
- 141 Kuroda, M.; Ohta, T.; Uchiyama, I.; Baba, T.; Yuzawa, H.; Kobayashi, I.; Cui, L.; Oguchi, A.; Aoki, K.; Nagai, Y.; Lian, J.; Ito, T.; Kanamori, M.; Matsumaru, H.; Maruyama, A.; Murakami, H.; Hosoyama, A.; Mizutani-Ui, Y.; Takahashi, N. K.; Sawano, T.; Inoue, R.; Kaito, C.; Sekimizu, K.; Hirakawa, H.; Kuhara, S.; Goto, S.; Yabuzaki, J.; Kanehisa, M.; Yamashita, A.; Oshima, K.; Furuya, K.; Yoshino, C.; Shiba, T.; Hattori, M.; Ogasawara, N.; Hayashi, H.; Hiramatsu, K. (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*, *Lancet* **357**, 1218-1219.
- 142 Bergmans, B. A.; De Strooper, B. (2010) Gamma-secretases: from cell biology to therapeutic strategies, *Lancet Neurol.* **9**, 215-226.
- 143 Steiner, H.; Fluhrer, R.; Haass, C. (2008) Intramembrane Proteolysis by gamma-Secretase, *J. Biol. Chem.* **283**, 29627-29631.
- 144 Golde, T. E. (2003) Alzheimer disease therapy: Can the amyloid cascade be halted?, *J. Clin. Invest.* **111**, 11-18.
- 145 Kreft, A. F.; Martone, R.; Porte, A. (2009) Recent advances in the identification of gamma-secretase inhibitors to clinically test the Abeta oligomer hypothesis of Alzheimer's disease, *J. Med. Chem.* **52**, 6169-6188.
- 146 McKee, S. (2010) Lilly hit by spectacular failure of Phase III Alzheimer's candidate, *PharmaTimes*, London.
- 147 Leuchtenberger, S.; Beher, D.; Weggen, S. (2006) Selective modulation of Abeta42 production in Alzheimer's disease: non-steroidal anti-inflammatory drugs and beyond, *Curr. Pharm. Des.* **12**, 4337-4355.
- 148 Takahashi, Y.; Hayashi, I.; Tominari, Y.; Rikimaru, K.; Morohashie, Y.; Kann, T.; Natsugari, H.; Fukuyama, T.; Tomita, T.; Iwatsubo, T. (2003) Sulindac sulfide is a noncompetitive gamma-secretase inhibitor that preferentially reduces Abeta 42 generation, *J. Biol. Chem.* **278**, 18664-18670.
- 149 Weggen, S.; Eriksen, J. L.; Das, P.; Sagi, S. A.; Wang, R.; Pietrzik, C. U.; Findley, K. A.; Smith, T. E.; Murphy, M. P.; Bulter, T.; Kang, D. E.; Marquez-Sterling, N.; Golde, T. E.; Koo, E. H. (2001) A subset of NSAIDs lower amyloidogenic Ab42 independently of cyclooxygenase activity, *Nature* **414**, 212-216.

- 150 Weggen, S.; Eriksen, J. L.; Sagi, S. A.; Pietrzik, C. U.; Golde, T. E.; Koo, E. H. (2003) Abeta42-lowering nonsteroidal anti-inflammatory drugs preserve intramembrane cleavage of the amyloid precursor protein (APP) and ErbB-4 receptor and signaling through the APP intracellular domain, *J. Biol. Chem.* **278**, 30748-30754.
- 151 Peretto, I.; La Porta, E. (2008) Gamma-secretase modulation and its promise for Alzheimer's disease: a medicinal chemistry perspective, *Curr. Top. Med. Chem.* **8**, 38-46.
- 152 Kimura, T.; Kawano, K.; Doi, E.; Kitazawa, N.; Takaishi, M.; Ito, K.; Kaneko, T.; Sasaki, T.; Sato, N.; Miyagawa, T.; Hagiwara, H.; Eisai Co., Ltd. Cinnamide Compound, WO 2005115990, 2005.
- 153 Green, R. C.; Schneider, L. S.; Amato, D. A.; Beelen, A. P.; Wilcock, G.; Swabb, E. A.; Zavitz, K. H. (2009) Effect of Tarenflurbil on Cognitive Decline and Activities of Daily Living in Patients With Mild Alzheimer Disease, *J. Am. Med. Assoc.* **302**, 2557-2564.
- 154 Galasko, D. R. (2007) Safety, Tolerability, Pharmacokinetics, and Abeta Levels After Short-term Administration of R-flurbiprofen in healthy elderly individuals, *Alzheimer Dis. Assoc. Disord.* **21**, 292-299.
- 155 Zettl, H.; Weggen, S.; Schneider, P.; Schneider, G. (2010) Exploring the chemical space of gamma-secretase modulators, *Trends Pharmacol. Sci.* **31**, 402-410.
- 156 Czirr, E.; Leuchtenberger, S.; Dorner-Ciossek, C.; Schneider, A.; Jucker, M.; Koo, E. H.; Pietrzik, C. U.; Baumann, K.; Weggen, S. (2007) Insensitivity to Abeta 42-lowering non-steroidal anti-inflammatory drugs (NSAIDs) and gamma-secretase inhibitors is common among aggressive presenilin-1 mutations. *J. Biol. Chem.* **282**, 24504-24513.
- 157 Farrar, M. (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations, *Bioinformatics* **23**, 156-161.
- 158 Oliver, T.F.; Schmidt, B.; Maskell, D.L.; Vinod, A.P. (2005) A Reconfigurable Architecture for Scanning Biosequence Databases, *IEEE International Symposium on Circuits and Systems 2005* **5**, 4799-4802.
- 159 Li I. S. T.; Shum, W.; Truong, K. (2007) 160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA), *BMC Bioinformatics* **8**, 185-191
- 160 Wirawan, A.; Kwo, C. K.; Hieu, N. T.; Schmidt, B. (2008) CBESW: Sequence Alignment on the Playstation 3, *BMC Bioinformatics* **9**, 377-386.

- 161 Liu, W.; Schmidt, B.; Voss, G.; Muller-Wittig, W. (2007) Streaming Algorithms for Biological Sequence Alignment on GPUs, *IEEE Transactions on Parallel and Distributed Systems* **18**, 1270-1281.
- 162 Manavski, S. A.; Valle, G. (2008) CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment, *BMC Bioinformatics* **9**, S10-S18.
- 163 Striemer, G.M.; Akoglu, A. Sequence Alignment with GPU: Performance and Design Challenges. *Proceedings of the 23rd IEEE International Parallel and Distributed Processing Symposium*, 1-10.
- 164 Batista, R. B.; Boukerche, A.; de Melo, A. C. M. A. (2008) A parallel strategy for biological sequence alignment in restricted memory space, *J. Parallel Distrib. Comput.* **68**, 548-561.
- 165 Rognes, T.; Seeberg, E. (2000) Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors, *Bioinformatics* **16**, 699-706.
- 166 Kabsch, W. (1976) A solution of the best rotation to relate two sets of vectors, *Act Crystallogr.* **32**, 922-923.
- 167 Kabsch, W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors, *Acta Crystallogr.* **A34**, 827-828.
- 168 Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25**, 3389-3402.
- 169 Wang, L.; Jiang, T. (1994) On the complexity of multiple sequence alignment, *J. Comput. Biol.* **1**, 337-348.
- 170 Just, W. (2991) Computational complexity of multiple sequence alignment with SP-score, *J. Comput. Biol.* **8**, 615-623.
- 171 Elias, I. (2006) Settling the intractability of multiple alignment, *J. Comput. Biol.* **13**, 1323-1339.
- 172 Waterman, M. S. (1983) Sequence Alignments in the Neighborhood of the Optimum with General Application to Dynamic Programming, *Proc. Natl. Acad. Sci.* **80**, 3123-3124.
- 173 Waterman, M. S.; Byers, T. H. (1985) A Dynamic Programming Algorithm to Find All Solutions in a Neighborhood of the Optimum, *Math. Biosci.* **77**, 179-188.

- 174 Zuker, M. (1991) Suboptimal Sequence Alignment in Molecular Biology, Alignment with Error Analysis, *J. Mol. Biol.* **221**, 403-420.
- 175 Naor, D.; Brutlag, D. (1993) On suboptimal alignments of biological sequences, *Lect. Notes Comput. Sc.* **684**, 179-196.
- 176 Bemis, G. W.; Murcko, M. A. (1996) The Properties of Known Drugs. 1. Molecular Frameworks, *J. Med. Chem.* **39**, 2887-2893.
- 177 Bemis, G. W.; Murcko, M. A. (1999) Properties of Known Drugs. 2. Side Chains, *J. Med. Chem.* **42**, 5095-5099.
- 178 Mockus, J.; Isenberg, A. C.; Vander Stouw, G. G. (1981) Algorithmic generation of Chemical Abstracts Index Names. 1. General design, *J. Chem. Inf. Comput. Sci.* **21**, 183-195.
- 179 Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. (2007) *J. Chem. Inf. Model.* **47**, 47-58.
- 180 Smith, T. F.; Waterman, M. S. (1981) Identification of Common Molecular Subsequences, *J. Mol. Biol.* **147**, 195-197.
- 181 Hilbert, D. (1891) Über die stetige Abbildung einer Linie auf ein Flächenstück, *Math. Ann.* **38**, 459-460.
- 182 Bissantz, C.; Kuhn, B.; Stahl, M. (2010) A Medicinal Chemist's Guide to Molecular Interactions, *J. Med. Chem.* **53**, 5061-5084.

16 - Acknowledgements

I would like to thank a lot of people for making this work possible.

- At first, of course, my supervisor Prof. Dr. Gisbert Schneider for the continuous support and belief in my abilities. Holding me on such a long leash gave me the freedom to explore even the most exotic aspects of sequence alignment.
- Prof. Dr. Ina Koch (Johann Wolfgang Goethe-University, Frankfurt) for her willingness to review my dissertation.
- Dr. Mireille Krier and Dr. Friedrich Rippmann (Merck KGaA, Darmstadt) for many helpful discussion and fruitful visits in Darmstadt and, of course, the fellowship granted during my PhD studies in Frankfurt.
- Dr. Dr. Thomas Wichelhaus and Dr. Johannes Zander (University Hospital, Frankfurt) for the cooperation on the Thymidinkinase project.
- Prof. Dr. Sascha Weggen (University Hospital, Düsseldorf) for the cooperation on the γ -Secretase project.
- Prof. Dr. Alexander Hartmann and Dr. Oliver Melchert (Carl von Ossietzky-University, Oldenburg) for the excellent supervision and great working atmosphere during my visit in Oldenburg. Their patience and detailed explanations made the adaption of Rare Event Sampling to PhAST possible.
- Adrian Fuchs and David Klemmer for continuous support and creative design of recreational time, even after my transfer to Zürich. Knowing to have good friends and sometimes getting reminded what really is important helped to keep everything in balance.
- My fellow lodgers and friends Alexander Klenner and Markus Hartenfeller for the unstressed and pleasant time in Zürich during the finalization of our PhD studies. Not having to worry and argue about who buys what or who used something up made things easy and relaxed. Sadly, after we accompanied each other for the last eight years through the complete time of our studies, we now most likely have to go separate ways. It is unbelievable how fast time flies by.

- My ‘cubemates’ Tim Geppert and Felix Reisen for many interesting and fruitful discussions about work-related and other topics and many joyful events and hours in Frankfurt and in Zürich.
- Dr. Heiko Zettl for many discussions about γ -Secretase, theories about modulation mechanisms, pharmacology in general and for his unique and admirable attitude towards life.
- Christian Koch for his expertise on the English language.
- Dr. Matthias Rupp for the continuous support on everything related to statistics and mathematics.
- The rest of the ‘modlab’ team in Frankfurt and in Zürich, including all members that had to leave while I was still working on my thesis.

My special thanks go to Helga Winkler, who was my biology teacher from 1999 to 2001. She introduced me to the idea to study bioinformatics and encouraged me to go this way. Without her I might have chosen another subject, and consequently this work would not exist.

And, of course, I thank my parents and my family, without whose continuous support this work would not have been possible.

17 - Appendix

This sections presents full reprints of the publications (including the complete respective supplementary material) that are part of this cumulative dissertation. The order is as follows:

- A Hähnke, V.; Rupp, M.; Krier, M.; Rippmann, F.; Schneider, G. (2010) Pharmacophore Alignment Search Tool: Influence of Canonical Atom Labeling on Similarity Searching, *Journal of Computational Chemistry* 31, 2810-2826.

- B Hähnke, V.; Klenner, A.; Rippmann, F.; Schneider, G. Pharmacophore Alignment Search Tool: Influence of the Third Dimension on Text-based Similarity Searching, *Journal of Computational Chemistry*, accepted.

- C Hähnke, V.; Schneider, G. Pharmacophore Alignment Search Tool: Influence of Scoring Systems on Text-based Similarity Searching, *Journal of Computational Chemistry*, accepted.

- D Zander, J.; Hartenfeller, M.; Hähnke, V.; Proschak, E.; Besier, S.; Wichelhaus, T. A.; Schneider, G. (2010) Multistep Virtual Screening for Rapid and Efficient Identification of Non-Nucleoside Bacterial Thymidine Kinase Inhibitors, *Chemistry – a European Journal* 16, 9630-9637.

Appendix A

Authors: Hähnke, V.
Rupp, M.
Krier, M.
Rippmann, F.
Schneider, G.

Title: Pharmacophore Alignment Search Tool: Influence of Canonical
Atom Labeling on Similarity Searching

Publication Year: 2010

Journal: Journal of Computational Chemistry
Volume 31
Pages 2810-2826

Pharmacophore Alignment Search Tool: Influence of Canonical Atom Labeling on Similarity Searching

VOLKER HÄHNKE,^{1,2} MATTHIAS RUPP,³ MIREILLE KRIER,⁴ FRIEDRICH RIPPMANN,⁴ GISBERT SCHNEIDER²

¹Chair for Chem- and Bioinformatics, Johann Wolfgang-Goethe University,
Frankfurt am Main 60323, Germany

²Eidgenössische Technische Hochschule (ETH), Institute of Pharmaceutical Sciences,
Zürich 8093, Switzerland

³Helmholtz Zentrum München, German Research Center for Environmental Health,
Neuherberg 85764, Germany

⁴Merck KGaA, Merck Serono Research, Bio- and Chemoinformatics,
Darmstadt 64293, Germany

Received 10 February 2010; Revised 7 April 2010; Accepted 7 April 2010
DOI 10.1002/jcc.21574

Published online 1 June 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Previously, (Hähnke et al., *J Comput Chem* 2009, 30, 761) we presented the Pharmacophore Alignment Search Tool (PhAST), a ligand-based virtual screening technique representing molecules as strings coding pharmacophoric features and comparing them by global pairwise sequence alignment. To guarantee unambiguity during the reduction of two-dimensional molecular graphs to one-dimensional strings, PhAST employs a graph canonization step. Here, we present the results of the comparison of 11 different algorithms for graph canonization with respect to their impact on virtual screening. Retrospective screenings of a drug-like data set were evaluated using the BEDROC metric, which yielded averaged values between 0.4 and 0.14 for the best-performing and worst-performing canonization technique. We compared five scoring schemes for the alignments and found preferred combinations of canonization algorithms and scoring functions. Finally, we introduce a performance index that helps prioritize canonization approaches without the need for extensive retrospective evaluation.

© 2010 Wiley Periodicals, Inc. *J Comput Chem* 31: 2810–2826, 2010

Key words: global alignment; line notation; molecular graph; similarity; virtual screening

Introduction

The Pharmacophore Alignment Search Tool (PhAST) is a string-based approach to virtual screening.¹ It reduces each molecule to an unambiguous linear representation describing its pharmacophore—called “PhAST-sequence”—in three steps: (i) each nonhydrogen atom in the structure graph is replaced by a potential pharmacophoric point symbol, hydrogen atoms are removed; (ii) vertices of this pharmacophore graph are canonically labeled by the algorithm of Weininger et al.²; and (iii) vertex symbols are concatenated into a string in increasing order of their canonical labels. For virtual screening, both the screening compound collection (“library”) and the query molecules are converted, and the resulting PhAST-sequences are compared using pairwise global sequence alignment.³ As a result, molecular similarity values are computed from the alignment, which can be used for the retrieval of pharmacophorically similar molecules from a compound database.

Here, we present some modifications to the original method. To speed up the alignment process, we exchanged the Needleman Wunsch algorithm³ with an algorithm proposed by Durbin

et al.⁴ that has the same asymptotic runtime complexity but a lower constant, i.e., it runs faster in practice. We compared the retrospective virtual screening performance of both algorithms using BEDROC scores⁵ with Pearson’s ρ ⁶ and complete ranked result lists with Kendall’s τ as rank-correlation coefficient.⁷ We further investigated alternatives for the evaluation of sequence alignments, namely the alignment score and the normalized alignment score. This was motivated by a previous comparison of these methods for determination of homology of protein sequences.⁸ There, sequence identity was inferior to the (normalized) alignment score, and both performed worse than significance-based evaluation methods like the *E*-value measure.⁹

The focus of the present work lies in comparison with different canonical labeling algorithms in step (ii) (*vide supra*) of

Additional Supporting Information may be found in the online version of this article.

Correspondence to: G. Schneider; e-mail: gisbert.schneider@pharma.ethz.ch

PhAST. We demonstrate the necessity of this canonization step, which improves performance over random symbol ordering. In addition to the algorithms proposed by Weininger et al.² and Jochum and Gasteiger,¹⁰ we implemented a third canonization algorithm for molecular graphs suggested by Prabhakar and Balasubramanian.¹¹ All three algorithms were tested in their original version and in a modified version that excludes some of the original vertex prioritization rules. In addition, we used several dimensionality reduction methods, namely linear principal component analysis¹² (PCA) and the nonlinear methods Laplacian Eigenmaps,¹³ Isomap,¹⁴ and minimum volume embedding¹⁵ (MVE) to reduce two-dimensional graphs to one-dimensional representations.

We compared the different canonization methods by retrospective virtual screening of a collection of drugs and lead compounds (collection of bioactive reference analogues (COBRA)).¹⁶ For statistical evaluation, we used BEDROC⁵ scores (with $\alpha = 20$ as suggested as default value for evaluation⁵), the permutation test proposed by Zhao et al.,¹⁷ and the Kolmogorov–Smirnov test.¹⁸ Finally, we investigated properties of the canonization algorithms related to their impact on virtual screening performance. To this end, we quantified the extent to which neighborhoods of graph vertices are preserved by the algorithms. Because the small structural modifications of molecules should result in similar PhAST-sequences, we investigated the effect of adding small fragments to the original molecules.

Methods

Definitions of pharmacophoric points used in step (i) of PhAST are shown in Tables 1 and 2. We use the original PhAST version¹ as baseline with the only modification being a gap open penalty of three instead of five. The reason for this was a change in preprocessing of compounds [“washing” with MOE (Molecular Operating Environment, version 2010.06, Chemical Computing Group, Montreal, Canada) instead of using fully protonated structures], resulting in best retrospective performance for the new gap penalty. All retrospective screens were performed using the COBRA library¹⁶ (version 6.1, 8311 bioactive compounds; see Table 3 for a list of the selected targets).

Sequence alignment is used in bioinformatics to decide how related two sequences (deoxyribonucleic acid (DNA), ribonu-

Table 2. Pharmacophoric Point Definitions in Terms of Molecular Query Language (MQL)¹⁹ Queries.

MQL query	PPP symbols
C	R
N	R
*[charge < 0]	N
*[charge > 0]	P
C(=O)–O–H	O;N;E
P(=O)–O–H	O;N;E
S(=O)–O–H	O;N;E
N[allHydrogens=0&totalConnections=3]	Q
N[allHydrogens=1&totalConnections=3](–C')–C'	U
N[allHydrogens=2&totalConnections=3]–C'	U
N[allHydrogens=1&totalConnections=2]=C'	E
N[allHydrogens=0&totalConnections=2](=C')–C'	A
O–H	E
C=O	O;A
C[!bound(~ N)&!bound(~O)].*[ClFClBrIIS]	L
Cl	L
Br	L
I	L
S[!bound(~N)bound(~O)].*[ClH]	L

Symbols are assigned to atoms in the query from left to right; queries are used from top to bottom.

cleic acid (RNA), and amino acid sequences) are. To create the alignment of two sequences $X = x_1x_2 \dots x_n$ and $Y = y_1y_2 \dots y_m$, their symbols are matched. The symbol order is retained, and gaps may be inserted to improve the matching (insertion of paired gaps is forbidden). Three cases exist: (i) x_i is aligned to y_j and $x_i = y_j$ (match), (ii) x_i is aligned to y_j and $x_i \neq y_j$ (mismatch), (iii) x_i is aligned to a gap in Y , or y_j is aligned to a gap in X . In protein sequence alignment, matches represent conserved residues, mismatches may arise from mutations, and gaps from insertions or deletions in an assumed evolutionary process of the sequences. Consequently, matches are rewarded with a positive score, mismatches are either rewarded with a positive score or penalized with a negative score (depending on the particular scoring scheme), and gaps are penalized with a negative score. The optimal alignment is the one with the highest overall score (summed over the whole alignment). It can be computed using dynamic programming.³

Table 1. Potential Pharmacophoric Points Used in PhAST.

Possible interactions	PPP symbol
Hydrogen bond acceptor	A
Charge positive	P
Charge negative	N
Lipophilic	L
Aromatic	R
Hydrogen bond acceptor, hydrogen bond donor	E
Hydrogen bond acceptor, charge positive	Q
Hydrogen bond acceptor, hydrogen bond donor, charge positive	U
No possible interactions	O

Table 3. Targets, Taken From the COBRA Library (Version 6.1, $n = 8,311$).

Target	Abbreviation	No. of actives
Angiotensin-converting enzyme	ACE	34
Cyclooxygenase 2	COX2	136
Dihydrofolate-reductase	DHFR	64
Factor Xa	FXA	228
Peroxisome-proliferator-activated receptor type γ	PPAR γ	44
Thrombin	THR	183
Total		689

Previously, we used an implementation of the Needleman Wunsch algorithm³ for sequence alignment by PhAST.¹ The algorithm was adapted to the affine gap penalty model with a fixed gap open and gap extension penalty.⁴ This version runs in $O(nm)$ instead of the original $O(nm(n+m))$ for any gap penalty model. To further speed up virtual screenings, we implemented the global pairwise sequence alignment algorithm conceived by Durbin et al.,⁴ hereafter referred to as finite-state-machine ("FSM") algorithm. It has the same $O(nm)$ runtime but runs noticeably faster in practice. In some cases, the two algorithms alignments are not identical, because the FSM algorithm prohibits the insertion of a gap in Y directly following a gap in X . This simplification reduces computational cost and causes the speedup, but it does not change the asymptotic runtime.

To assess whether this influences results, we conducted the same retrospective virtual screenings twice, once for each algorithm. For each target (Table 3), each active was used as query, resulting in 689 ranked lists for each of the two alignment algorithms. For each list, the BEDROC score⁵ was calculated (with $\alpha = 20$). The correlation between the two sets of BEDROC scores was determined using Pearson's ρ .⁶ Statistical significance of the observed correlation was estimated from the p -value of a t test under the null hypothesis that the correlation equals zero.

The BEDROC score is based on ranks and thus invariant under permutations of the actives' ranks. To investigate differences in the complete ranked lists produced by both algorithms, we compared the two ranked lists of each query with Kendall's τ^7 as rank-correlation coefficient. Because ties can occur, we used τ_b , which corrects for this scenario. The significance of the observed rank correlation was calculated as the p -value of a z test under the null hypothesis that the correlation equals zero.

The focus of this work is on the influence of the canonization step on PhAST performance. We compared canonization methods as follows: With each method and for each target, each active was used as query in a retrospective virtual screening, resulting in 689 ranked lists. For each ranked list, the BEDROC score was calculated ($\alpha = 20$). The mean BEDROC score was used as overall performance index. For each canonization method and target, gap open and extension penalties were optimized using a grid search (starting from gap open penalty = 2 and gap extension penalty = 1, each combination with gap extension penalty lower than gap open penalty was tested, resulting in 190 penalty combinations), as it is hard to choose them by intuition.²⁰ Gap penalties greater than 20 seem unreasonable as they exceed the highest mismatch penalties.

To prove that the canonization step is essential, we compared baseline PhAST (Weininger canonization) against PhAST with random labeling in step (ii) of the algorithm. To avoid bias (default gap penalties are optimized for Weininger algorithm), we used the same simple scoring scheme for both labeling methods: Matches are rewarded with +1, mismatches are penalized with -1, and both gap penalties are 1. For random labeling, we generated 100 pairs of PhAST-sequences for each pair of molecules and used the average score as final similarity value.

To assess whether two different versions of PhAST have significantly different performance, we used the permutation test

proposed by Zhao et al.¹⁷ It has the null hypothesis that virtual screening method P performs significantly better than method Q . Assuming p and q are rank lists of actives resulting from the virtual screening methods, the null hypothesis requires that $\text{BEDROC}(p) > \text{BEDROC}(q)$. As each active has two ranks, one in p and one in q , new rank lists p^* and q^* can be created by swapping its rank in p with its rank in q for each active with probability 1/2. This is repeated 10,000 times and the frequency of the event that $\text{BEDROC}(p) - \text{BEDROC}(q)$ is less than $\text{BEDROC}(p^*) - \text{BEDROC}(q^*)$ is recorded. The frequency of this event is the type I error rate for the null hypothesis. In addition, we used a Kolmogorov-Smirnov test¹⁸ for the same purpose. Both methods were used to assess the significance of the difference between using the Weininger algorithm for graph canonization and using random concatenation of symbols and to assess the improvement from baseline PhAST to the best combination of algorithms identified in this work. In both cases, calculations were based on the 689 BEDROC scores resulting from each version of PhAST.

Canonization Algorithms

The atom-typing step in PhAST yields a graph of potential pharmacophoric points that has the same topology as the molecular graph with suppressed hydrogen atoms. Each vertex is colored with a symbol (Table 1) that represents a potential pharmacophoric point. Edges represent covalent bonds. Canonization is the labeling of the vertices with the natural numbers 1,2,3... In a previous study,¹ we compared the canonization algorithms of Weininger et al.² and Jochum and Gasteiger¹⁰; we reevaluate them, here, because of changes in molecule preprocessing. In contrast to these two algorithms, the one by Prabhakar and Balasubramanian¹¹ is based on paths, a property thought to be beneficial for PhAST. We modified all algorithms by using pharmacophoric points as prioritization criterion instead of the element number.

Jochum-Gasteiger Method

The canonical labels created by the Jochum and Gasteiger algorithm¹⁰ are in most cases identical to those obtained by the Morgan²¹ algorithm.¹⁰ The first step is the separation of all vertices into two sets—terminal vertices (vertices with exactly one single bond) and core vertices (all others). All core vertices with the same buriedness are members of the same equivalence class. The algorithm divides the vertices of each class further using a set of prioritization criteria, until only one atom remains that gets the next label, starting from the vertices in the most buried class. Prioritization criteria are (i) priority of the potential pharmacophoric point (atom number in the original application) and (ii) number of free electrons. In both cases, the vertex with the highest value has priority. The next criteria involve the environment of the vertices organized in spheres around each vertex. The vertex with the highest of these values in his neighborhood gets priority: (iii) number of vertices, (iv) priority of potential pharmacophoric points, (v) number of free electrons, (vi) number of bonds in the next sphere, (vii) bond order of these bonds, (viii) neighborhood to an already labeled vertex, and (ix) bond

order to the vertex in (viii). If more than one vertex remains after (ix), all of them are marked as indistinguishable and the remaining vertices have priority over them. After all distinguishable vertices are labeled, (viii) is used to label the undistinguishable vertices. After all core vertices are prioritized, terminal vertices are prioritized by criteria (i) and (viii).

Weininger Method

The canonization algorithm by Weininger et al. was proposed as part of canonical simplified molecular input line entry system (SMILES) generation.² Its idea is to assign vertices to topological symmetry classes. It first assigns a property vector to each vertex that consists of different atomic invariants mainly based on the original molecular graph: (i) number of connected vertices, (ii) number of connected nonhydrogen atoms, (iii) priority of the pharmacophoric point (atom number in the original version), (iv) sign of charge, (v) absolute charge, and (vi) number of connected hydrogen atoms. Vertices with identical vectors form an equivalence class, and all vertices are sorted ascending by this vector. For each vertex, its extended connectivity is calculated as follows: Beginning with the equivalence class with the lowest index, the vertices in each class are assigned the same prime number, starting with 2. For each vertex in the graph, the product of the primes of its neighbors is calculated. These product values define new equivalence classes on the vertices. Each equivalence class, in order of product values, is assigned an index, starting from 1. This process is repeated until the number of equivalence classes does not change in a step. If, after extended connectivity calculation, an equivalence class contains two or more vertices, these ties are broken by an additional step: the index of each equivalence class is doubled, and one vertex from the equivalence class with the lowest index is randomly chosen to form an own equivalence class with the index of its original equivalence class lowered by 1. After that, all equivalence classes are renumbered starting from 1. These two steps (computing the extended connectivity and breaking ties) are alternated until the number of equivalence classes equals the number of vertices in the graph.

Prabhakar–Balasubramanian Method

The canonization algorithm by Prabhakar and Balasubramanian¹¹ uses more graph-based prioritization rules than the other two algorithms and progresses along paths through the graph. First, the number of incident bonds with respect to bond order is determined for each vertex (c_n). As with the Jochum and Gasteiger algorithm, vertices are divided into two sets, terminal vertices ($c_n = 1$) and core vertices ($c_n > 1$). Labeling starts with the core atoms. Using the following prioritization rules, they are divided into smaller subsets, until only one atom (which will get the next canonical label) remains: (i) number of incident bonds, (ii) number of incident bonds with respect to bond order, and (iii) pharmacophoric point priority (atom number in the original version of the algorithm). In these cases, the vertex with the highest value has priority. If more than one atom with highest priority remains, copies of the original graph are created, called "fragments." If there are n vertices left for prioritization, $n-1$

copies of the original graph are created for each vertex v . In each copy, the first bond in the shortest path between v and one of the other competing vertices is deleted. Only the part of the copy that includes v is retained. The remaining prioritization rules are applied to these fragments; a vertex has the highest priority, if one of his fragments has a higher priority than the fragments created for all other vertices: (iv) the length of the path starting in the competing vertex and following the highest pre-computed c_n values, until it reaches a vertex already visited or labeled, (v) number of loops, (vi) length of the longest path in the fragment, (vii) number of pharmacophoric points not lipophilic, aromatic or no interaction, (viii) summed symbol priorities of vertices in the fragment, (ix) averaged distances between all vertices not lipophilic, aromatic, or no interaction in the fragment. In all cases except the last one, the fragment with the highest value has priority. If there remains more than one vertex and there is no already labeled vertex, one of them is chosen arbitrarily and has priority over all other vertices. If there is already at least one labeled vertex, (x) the label of the connected vertex is used. These rules are used in a depth-first search. All neighbors of the last labeled vertex are the potential candidates for the next label. If this search reaches an end point, all vertices adjacent to an already labeled vertex are candidates for the next label. After all core vertices are labeled, terminal vertices are labeled according to criteria (ii), (iii), and the label of the neighboring core atom.

Irrespective of the canonization method used, the PhAST-sequences created from two identical graphs of pharmacophoric points are identical. If a pharmacophoric point is changed, but the topology remains the same, the relative order of symbols in the PhAST-sequence should remain unchanged as well. Yet all three algorithms use pharmacophoric point priority as a prioritization criterion. Consequently, the changes in a PhAST-sequence because of exchange of a single vertex symbol can be more severe than intended. To attenuate this, each canonization algorithm was tested in a modified version:

- In the Jochum and Gasteiger algorithm, clipping of terminal atoms was omitted, and the criteria symbol priority and number of free electrons were eliminated.
- For the Weininger algorithm, the creation of the initial prioritization vector was changed: the priority of the pharmacophoric point, the total number of neighbors, and the number of neighboring hydrogen vertices were omitted and both charge criteria.
- For the Prabhakar algorithm, the initial clipping of terminal vertices was omitted. The priority of the pharmacophoric point, the number of pharmacophoric points not lipophilic, aromatic or no interaction in a fragment and averaged distances between all vertices not lipophilic, aromatic or no interaction in a fragment were removed.

Canonization by Dimensionality Reduction

An alternative approach to canonization that to our knowledge has not been used before for canonization and does not suffer from the mentioned drawbacks is the use of dimensionality reduction algorithms. We implemented four such methods.

Principal Component Analysis

Principal component analysis¹² is a linear dimensionality reduction method often used to visualize high-dimensional data. We used PCA to calculate one-dimensional coordinates from two-dimensional graph layouts generated by the 2D depiction algorithm of MOE (version 2010.06, Chemical Computing Group, Montreal, Canada). Therefore, the coordinates of the vertices in each graph were mean-centered, and the covariance matrix between the position vectors of all vertices calculated. The computation of the eigenvectors and eigenvalues of the covariance matrix gives the loading vectors that are used for the computation of the new coordinates of the vertices. To get the one-dimensional coordinate for each vertex, the dot product between its original position vector and the loading vector with the highest absolute eigenvalue was calculated. Beginning with the vertex with lowest one-dimensional coordinate, we assigned labels in ascending order. For identical one-dimensional coordinates, we used their coordinate in the second dimension of the principal component space as prioritization criterion. In all cases, the vertex with the lowest coordinate had the highest priority.

PCA finds a low-dimensional embedding of data points that best preserves their variance. However, PCA fails when a data set contains nonlinear structures. Nonlinear approaches that overcome this problem start with the assignment of vertex neighborhoods by using a connectivity algorithm like k -nearest neighbors,²² b -matching²³ (each vertex gets assigned exactly b neighbors), or ϵ -balls (a vertex is connected to all vertices within distance ϵ), resulting in a neighborhood graph with edges between neighboring vertices. We directly used the topology of the molecular graph instead of connectivity algorithms. In the embedding, these methods aim at preserving the pairwise distances between neighbors.

Laplacian Eigenmaps

Laplacian Eigenmaps¹³ start by calculating three matrices from the neighborhood graph: the weight matrix W with [eq. (1)]

$$W_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{else} \end{cases}, \quad (1)$$

the degree weight matrix D with the column sums of W as entries [eq. (2)],

$$D_{ii} = \sum_j W_{ij}, \quad (2)$$

and the positive semidefinite Laplacian matrix L [eq. (3)] with

$$L = D - W. \quad (3)$$

Then, the eigenvalues and eigenvectors of the generalized eigenvector problem [eq. (4)] are calculated.

$$Lf = \lambda Df \quad (4)$$

Eigenvectors (f) are sorted according to their eigenvalues (λ) in ascending order. The first eigenvector with $\lambda = 0$ is omitted. The next d eigenvectors are used for embedding. In our case,

the second eigenvector contains the coordinates for the one-dimensional embedding.

Isomap

The Isomap algorithm by Tenenbaum et al.¹⁴ uses the neighborhood graph to estimate geodesic distances between the vertices. A matrix D of shortest distances between all vertices is computed, e.g., using the Floyd–Warshall algorithm.^{24,25} Using D , the matrix $\tau(D)$ is calculated as [eq. (5)]:

$$\tau(D) = -\frac{1}{2} * HSH \quad (5)$$

where S is the matrix of squared distances [eq. (6)]

$$S_{ij} = D_{ij}^2 \quad (6)$$

and H the centering matrix [eq. (7)]:

$$H_{ij} = \delta_{ij} - \frac{1}{n} \quad (7)$$

with δ_{ij} the Kronecker delta and n the number of vertices. The eigenvectors and eigenvalues of $\tau(D)$ are computed. To embed in d dimensions, the first d eigenvectors sorted according to their eigenvalues in decreasing order are used. If λ_p is the p th eigenvalue of $\tau(D)$ and v_p^i is the i th component of the p th eigenvector, then the p th component of the d -dimensional coordinate vector of a vertex is equal to $\sqrt{\lambda_p} v_p^i$.

Minimum Volume Embedding

The two previous methods lose all information contained in the eigenvectors that are not used for the embedding. None of them aims at minimizing the amount of information lost this way. MVE¹⁵ preserves as much information as possible in the d dimensions used for embedding. This is achieved by an iterative process based on semi-definite programming (SDP). First, an affinity matrix A is calculated for the vertices using a kernel function k . A is positive semidefinite and must be centered. This matrix is used in the neighborhood definition process (instead of given vertex coordinates) to obtain a binary connectivity matrix C . A third matrix K is set equal to A . The following procedure is repeated until convergence: (i) Calculate the eigenvectors f_i and eigenvalues λ_i of K , sort the f_i descending to their corresponding λ_i . (ii) Calculate the matrix B using eq. (8),

$$B = -\sum_{i=1}^d f_i f_i^T + \sum_{i=d+1}^N f_i f_i^T \quad (8)$$

(iii) use SDP to solve eq. (9)

$$K = \arg \min_{K \in \mathcal{K}} \text{tr}(KB) \quad (9)$$

under constraints \mathcal{K} defined by Shaw and Jebara,¹⁵ tr denotes the matrix trace (sum of the diagonal elements).

After convergence, kernel PCA²⁶ is performed with K to get the d eigenvectors used for embedding. MVE works with any positive semidefinite kernel k . We used MVE with two different kernel functions. The first one is a diffusion kernel.²⁷ For each pair of vertices v_i and v_j , it returns the probability that a random walk starting in v_i will be in v_j after an infinite number of steps, with only a low probability of leaving the current vertex in each step. The kernel matrix can be calculated according to eq. (10).²⁸

$$K = e^{(-\beta L)} \quad (10)$$

where L is the Laplacian matrix introduced in (3) and β is the diffusion parameter. If β equals 0, no diffusion is allowed and K equals the unit matrix. K is computed by matrix exponentiation, which is different from componentwise exponentiation. We used 12 values for β to determine its influence on PhAST: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, and 10. The second kernel function calculates inner products from the Euclidean coordinates of the vertices. It is defined as given in eq. (11).

$$k(x, y) = \frac{1}{2} * (\|x\|^2 + \|y\|^2 - \|x - y\|^2). \quad (11)$$

To obtain Euclidean coordinates for each vertex, as for PCA, we used the 2D depiction algorithm of MOE. This particular kernel is parameter free.

When using dimensionality reduction methods, slight modifications of a molecule can switch the direction of the axis of the one-dimensional coordinate system. We addressed this issue by repeating every sequence comparison during the virtual screening with one of the sequences inverted, if dimensionality reduction was involved. We used the higher of the two resulting values as similarity measure.

Evaluation of the Alignment

In addition to canonization methods, we analyzed the effect of different alignment evaluation methods. The alignment of nucleic or amino acid sequences is a traditional field in bioinformatics and well analyzed. It was shown that there are methods to determine peptide sequence homology from pairwise alignments that yield better results than the sequence identity used in PhAST.⁸ These methods are alignment scores and significance estimations of alignments. We evaluated two variants of sequence identity and three variants of alignment scores. There is no significant overhead compared with sequence identity calculation.

Sequence Identity

The original PhAST used the percent identity (PID₁) between two sequences X and Y [eq. (12)],

$$\text{PID}_1 = \frac{M(A(X, Y))}{L(A(X, Y))} \quad (12)$$

with $A(X, Y)$ the alignment of X and Y , $M(A(X, Y))$ the number of matches in the alignment, and $L(A(X, Y))$ the length of the align-

ment (including all gaps). Comparing two sequences of molecules active on the same target but of different size might result in a low PID₁, because the global alignment has to extend the shorter sequence to the length of the longer one with gaps. To counteract this effect, we correct for the size of the sequences: We first calculate the maximum reachable PID₁ of two sequences by inserting the maximum number of matches (length of the shorter sequence) and the minimum length of the alignment (length of the longer sequence) into eq. (12). We then normalize PID₁ according to eq. (13):

$$\text{PID}_2 = \frac{\text{actual PID}_1}{\text{max PID}_1} \quad (13)$$

Alignment Score

Besides sequence identity, we investigated alignment scores as evaluation measures. The raw alignment score S_1 is the sum of matches, mismatches, and gap penalties. Alignments of long sequences tend toward higher scores, so S_1 depends on sequence similarity and length. We normalize S_1 by dividing through the length of the alignment, yielding S_2 . The resulting score measures the average contribution of each event in the alignment (match, mismatch, or gap). S_1 can also be normalized by dividing through the length of the shortest original sequence, yielding S_3 , a measure of the maximum averaged contribution of each symbol in a sequence. All alignment score methods are measures of the similarity between aligned sequences, but are no longer bounded by 0 and 1.

Alignment evaluation methods involving alignment length or the number of occurrences of an event suffer from the drawback that there can be more than one optimal alignment of two sequences, which one is found depends on implementation details. The alignment of the sequence pair (X, Y) can, therefore, differ from that of (Y, X) in length, number of matches, number of mismatches, number of gaps, and gap length. To ensure the symmetry of our method, i.e., identical scores for $A(X, Y)$ and $A(Y, X)$, we modified the affected evaluation methods. In case of PID₁, we compute $A(X, Y)$ and $A(Y, X)$, and use the average PID₁ as final evaluation measure. This correction is used in the calculation of actual PID₁ for PID₂ as well. In case of S_2 , we align both sequence pairs and use the averaged alignment length for normalization.

In total, we compared 11 graph canonization methods combined with five alignment evaluation methods and 190 gap penalty combinations by conducting 689 virtual screenings for each combination and averaging the resulting BEDROC scores ($\alpha = 20$). To assess whether our modifications lead to significant improvements of PhAST, we used the permutation test proposed by Zhao et al.¹⁷ and a Kolmogorov–Smirnov test.¹⁸

Canonization Analysis

To further quantify the differences between canonization methods, we determined how well the neighborhood relations in the original pharmacophoric point graph are retained and represented in the resulting PhAST sequences. We did this by counting how often the vertices, which are neighbors in the graph, are

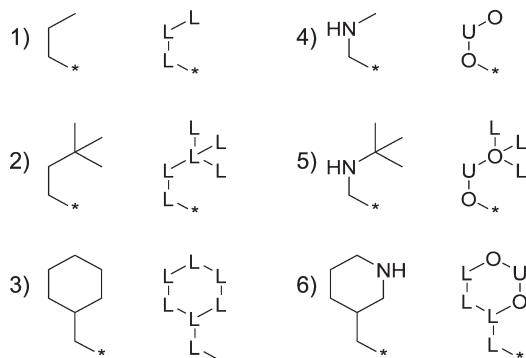


Figure 1. Fragments for the comparison of modified molecules. * Denotes the attachment point. Shown are molecular fragments (left) and corresponding pharmacophoric point graphs (right).

neighbors in the PhAST-sequence. This was done for every molecule in the COBRA library of reference compounds.¹⁶ For all graph neighbors that are not neighbored in the one-dimensional representation, we counted how many vertices were inserted between them, resulting in a histogram of these distances. As each pair of neighboring vertices was viewed twice, once from each vertex as origin, the resulting counts were divided by 2.

For use with PhAST, canonization algorithms should be robust against small changes in molecular structure, i.e., similar molecules should have similar PhAST-sequence. This in turn means that neighbors in the PhAST-sequences of a molecule should remain neighbors even if the molecule is slightly modified. To test this, the compounds in the COBRA library were modified by attaching small fragments. After conversion to PhAST-sequence, we counted (i) neighboring vertices in the original PhAST-sequence that are neighbors in the modified PhAST-sequence, (ia) with their relative orientation as in the original PhAST-sequence, (ib) with their relative orientation changed; (ii) neighboring vertices in the original PhAST-sequence that are not neighbors in the modified PhAST-sequence, (iia) with their relative orientation as in the original PhAST-sequence, (iib) with their relative orientation changed. If two vertices of the same type that are neighbors in the original PhAST-sequence are still neighbors in the modified PhAST-sequence, but changed their relative orientation, this event is counted as (ia) because the change of positions has no effect on the PhAST-sequence due to identical types. Cases (ib) and (iib) present a problem for global sequence alignment as a string comparison method. As the relative position of symbols cannot be changed, the only two operations that can be reconstructed in the alignment process are mutations (a symbol changed to another one) and insertion/deletions (compensated by gaps). A transversion, i.e., the swapping of two different symbols, cannot be properly modeled by only one event.

Figure 1 presents the six used fragments. Three fragments consist only of carbon atoms that are typed as L in the atom typing step. The other three are topologically identical to the first three fragments, with one carbon changed to nitrogen typed as

U. This way, the algorithms are confronted with topological modifications and changes in vertex priorities. Fragment attachment points should not change the atom typing. We used carbon atoms that were typed as L, R, or O in the original molecule and were connected to at least two hydrogen atoms. One of the hydrogen atoms was replaced by the first atom of the fragment. Molecules (1612/8311) from COBRA were omitted, because they had less than 10 possible attachment points. Each single fragment attachment was repeated five times at random positions. In addition, each possible combination between fragments (resulting in 21 unique pairs) was used five times as well, again at random positions. In total, each molecule was compared to 135 variants of itself.

All modifications were undertaken in a single preprocessing of the COBRA compound collection to ensure that all algorithms are compared with the same modified molecules. The resulting molecular graphs were depicted using MOE, because one variant of MVE depends on Euclidean distances. In the case of dimensionality reduction methods, again both possible canonization results were used in the analysis, and we used the one with a higher preservation of neighborhood relationships. This is justified, because for these methods, both orientations of the PhAST-sequence are used during a virtual screen.

All programing was done using the Java Programming Language (version 6). Eigen decompositions were done with the java linear algebra package (JLAPACK) library.²⁹ SDP problems were solved with the c semi-definite programming library (CSDP) solver.³⁰ Productive runs and calculations were performed on a Linux cluster with 40 advanced micro devices (AMD) Opteron 8214 processors and 320 gigabyte (GB) random access memory (RAM).

Results and Discussion

Choice of Alignment Algorithm

To determine whether the exchange of the alignment algorithm of Needleman and Wunsch by the faster FSM algorithm affected the performance of PhAST, we determined the correlation of BEDROC scores obtained from virtual screening with each active as reference ($n = 689$). The Pearson-correlation coefficient was 0.9996 with a p -value of below 10^{-1051} . To quantify the differences within the complete ranked lists, Kendall's τ was computed (see Fig. 2). The minimum correlation observed was 0.945, the maximum correlation observed was 0.998, and the average correlation observed was 0.984. The p -value for each τ was below $0.5 * \operatorname{erfc}\left(\frac{82,779,141,588 + \sqrt{2}}{109,242,709}\right)$. The FSM algorithm used only 40% of the runtime of the Needleman Wunsch algorithm. On the basis of the gain in computation speed and the high correlations, we decided to employ the FSM algorithm for all experiments in this study.

Necessity for Canonization

To verify the importance of the canonization step, we compared BEDROC values ($\alpha = 20$) of baseline PhAST and PhAST with random labeling (average of 100 random labeling procedures

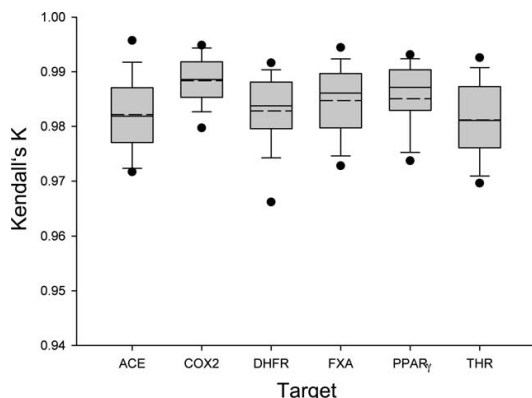


Figure 2. Box-whisker plots of rank correlation coefficients between the ranked lists obtained using the Needleman–Wunsch and the FSM algorithm for sequence alignment per target ($n = 34, 136, 64, 228, 44,$ and 183). Shown are 5th/95th (points), 10th/90th (whiskers), 25th/75th (box borders) percentiles, median (solid line), and mean (dashed line).

per comparison; matches = +1, mismatches = -1, gap penalties = 1).

Figure 3 presents the distribution of BEDROC scores per target. There is almost no overlap, with the exception of cyclooxygenase 2 (COX2). The latter can be explained by the distribution of pharmacophoric points: As given in Table 4, COX2 ligands have 56% pharmacophoric points of type R. Random concatenations of pharmacophoric points of the same type do not change the resulting PhAST sequence (more symbols of the same type

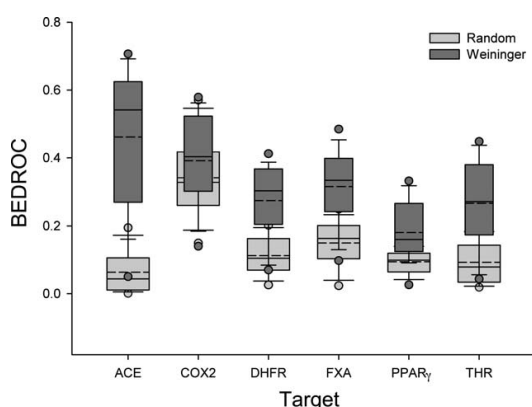


Figure 3. BEDROC ($\alpha = 20$) scores of Weininger versus random canonization. A simple alignment scoring system with +1 (-1) for matches (mismatches) and gap open and extension penalty of 1. For random canonization, the mean similarity of 100 random sequences was used. Shown are 5th/95th (points), 10th/90th (whiskers), 25th/75th (box borders) percentiles, median (solid line), and mean (dashed line).

Table 4. Symbol Frequencies in the COBRA Library.

	ACE	COX2	FXA	PPAR γ	DHFR	THR
\emptyset Symbols	27.12	24.61	34.34	28.57	27.34	35.95
σ	7.43	3.17	5.28	7.03	5.71	7.90
A	8.89	2.27	4.87	5.73	4.91	6.99
E	0	0.24	0.28	0.08	0.29	1.08
L	23.64	17.99	9.45	17.10	13.83	21.01
N	6.40	0.69	0.46	3.10	2.57	00.43
O	23.64	22.80	25.57	19.97	17.49	29.47
P	3.25	0.09	2.85	0.56	0.97	3.02
Q	1.74	0.18	2.92	0.88	0.74	2.60
R	31.45	55.48	48.43	51.31	49.94	29.21
U	0.98	0.27	5.17	1.27	9.26	6.19

“ \emptyset Symbols” and “ σ ” indicate the average number of pharmacophoric points per molecule and the standard deviation.

only result in fewer possible sequences). The global sequence alignment matches regions of identical symbols between sequences, resulting in a high score. Subsequences of different lengths are compensated by gaps, lowering the score. PhAST-sequences of COX2 ligands have shortest average length and standard deviation, i.e., they are least affected by this effect because they are of comparable length. In agreement with this reasoning, the target with second largest overlap, peroxisome-proliferator-activated receptor type γ (PPAR γ), is also second in type R symbols (51%) and of comparatively small size.

Table 5 presents the results of the Zhao permutation test. PhAST with Weininger canonization performed significantly better than random labeling in 91% of all screenings at a significance level of 0.05. In 5% of the screens, random labeling performed better. The latter cases are dominated by screenings on COX2. A Kolmogorov–Smirnov test showed that the difference between these two methods is significant with a p -value of 1.0835×10^{-78} . We conclude that the canonization step is necessary for PhAST.

Comparison of Canonization Methods

We compared 11 canonization algorithms, five alignment evaluation methods, and 190 gap penalty combinations for their effect on PhAST in a set of virtual screening experiments employing

Table 5. Permutation Test Results for Weininger Canonization Versus Random Canonization.

	No. of queries	Weininger	Random
ACE	34	100 (100)	0 (0)
COX2	136	70 (68)	22 (21)
DHFR	64	95 (95)	2 (2)
FXA	228	98 (98)	1 (1)
PPAR γ	44	84 (84)	7 (5)
THR	183	97 (97)	0 (0)
Total	689	91 (91)	5 (5)

Shown are the percentages of cases where one contestant performs significantly better than the other at a significance level of 0.05 (0.01).

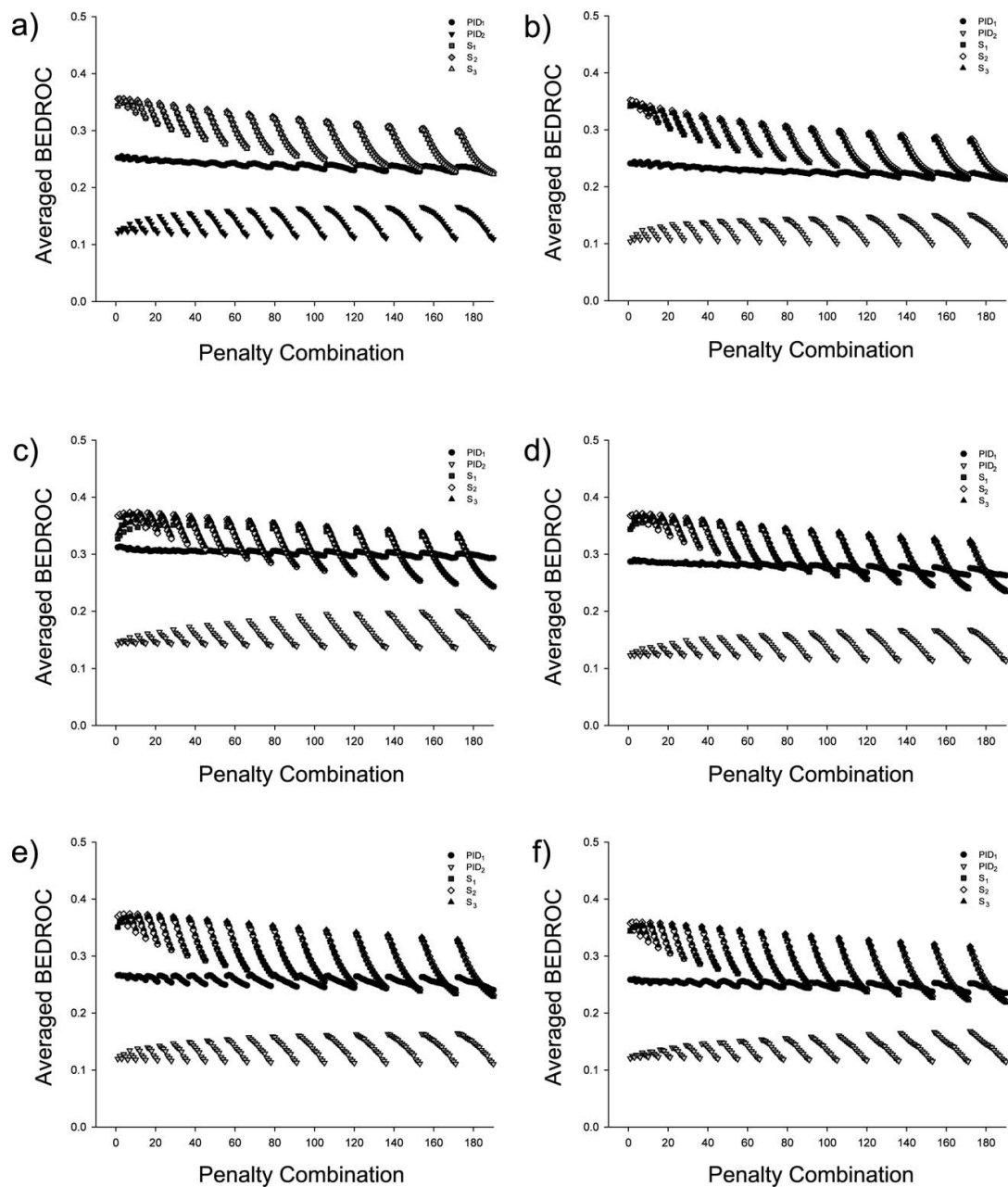


Figure 4. BEDROC ($\alpha = 20$) scores for combinations of alignment evaluation methods and gap penalties. (a) Jochum and Gasteiger algorithm, (b) modified Jochum and Gasteiger algorithm, (c) Weininger algorithm, (d) modified Weininger algorithm, (e) Prabhakar algorithm, (f) modified Prabhakar algorithm, (g) MVE with diffusion kernel and diffusion parameter 0.4, (h) MVE with Euclidean distance kernel, (i) Laplacian Eigenmaps, (j) Isomap, and (k) PCA.

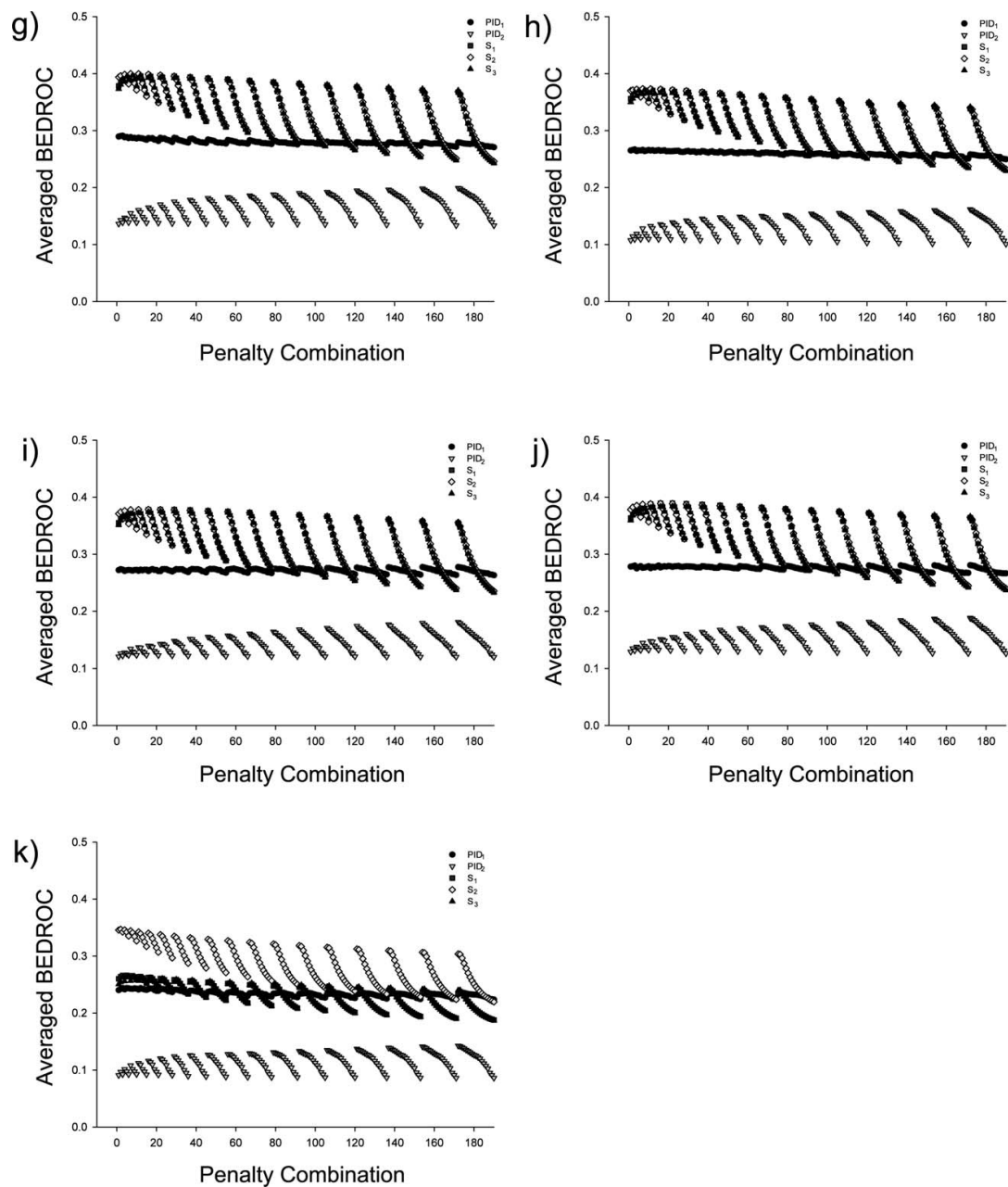


Figure 4. (Continued)

six targets (Table 3) from the COBRA library of bioactive compounds. For each combination of canonization, alignment, and gap penalties, each active molecule of each target was used as query in a virtual screening. Screening success was assessed by

average ($n = 689$) BEDROC ($\alpha = 20$) scores. For MVE with diffusion kernel, we compared 12 values of the diffusion parameter β . Only the best performing version with $\beta = 0.4$ that reaches the highest averaged BEDROC score is included in the

Table 6. BEDROC Scores ($\alpha = 20$).

	Scoring scheme				
	PID ₁	PID ₂	S ₁	S ₂	S ₃
Jochum Gasteiger	0.26 (3, 2)	0.17 (20, 1)	0.35 (5, 1)	0.36 (4, 1)	0.35 (4, 2)
Jochum Gasteiger modified	0.24 (3, 2)	0.15 (20, 1)	0.35 (4, 1)	0.35 (2, 1)	0.35 (3, 2)
Isomap	0.28 (20, 1)	0.19 (20, 1)	0.39 (8, 1)	0.39 (7, 1)	0.38 (8, 1)
Laplacian Eigenmaps	0.28 (20, 1)	0.18 (20, 1)	0.38 (8, 1)	0.38 (7, 1)	0.38 (9, 1)
MVE diffusion kernel 0.4	0.29 (3, 2)	0.20 (20, 1)	0.39 (7, 1)	0.40 (5, 1)	0.39 (6, 2)
MVE Euclidean kernel	0.27 (3, 2)	0.16 (20, 1)	0.37 (7, 1)	0.37 (5, 1)	0.37 (5, 2)
PCA	0.24 (3, 2)	0.14 (20, 1)	0.27 (4, 2)	0.35 (3, 1)	0.26 (4, 3)
Prabhakar	0.27 (13, 2)	0.16 (20, 1)	0.37 (6, 1)	0.37 (5, 1)	0.37 (7, 1)
Prabhakar modified	0.26 (3, 2)	0.17 (20, 1)	0.26 (6, 1)	0.36 (4, 1)	0.36 (6, 1)
Weininger	0.31 (3, 1)	0.20 (20, 1)	0.36 (7, 2)	0.37 (6, 1)	0.37 (7, 2)
Weininger modified	0.29 (3, 2)	0.17 (29, 1)	0.36 (5, 2)	0.37 (4, 1)	0.37 (5, 2)

The best gap open/gap extension penalties are shown in parentheses.
Higher scores indicate better virtual screening performance.

following analysis. The results for the remaining settings of β are presented in Figure S1 (cf. Supporting Information). Figure 4 presents the outcome of the comparison.

For each canonization algorithm, alignment evaluation methods and gap penalty combinations have similar effects. For the algorithms of Weininger, Prabhakar, and Jochum–Gasteiger, the respective modified versions display a retrospective performance that is comparable to the original versions of the algorithms. Our modifications, therefore, neither worsened nor improved their performance. For MVE, the effect of different kernel functions is minor. The best parameterization of the diffusion kernel was slightly superior to the Euclidean distance kernel. Isomap and Laplacian Eigenmaps performed comparably but with slightly lower BEDROC scores. Among the dimensionality reduction methods, PCA performed worst. Considering the comparison of baseline PhAST with random labeling, we conclude that having a canonization method at all is more important than the particular method used.

PID₁ is the original alignment evaluation method, PID₂ penalizes differences in sequence lengths to a lesser extent than PID₁, and performs worse than PID₁ manifesting in lower averaged BEDROC scores (see Fig. 4). We introduced PID₂ to compensate for the difference in sequence lengths of actives of the same target. Although PID₂ yielded greater similarity values than PID₁, it also did so for sequences from different targets. This effect generates false positives, thereby, diminishing screening performance.

The evaluation methods S₁, S₂, and S₃ are all based on the alignment score with different normalization techniques. They perform similar and all of them appear to be superior to alignment evaluation methods based on sequence identity. We explain this observation by the improved weighting of matches and mismatches. Sequence identity is influenced only by the number of (mis)matches. The alignment score, however, is influenced by the exact type of match or mismatch depending on the symbols involved. For similar numbers of matches and mismatches, this enables a more differentiated evaluation of the alignment.

All combinations of canonization and alignment evaluation strongly depend on the gap penalties used. Retrospective results for one particular combination of canonization algorithm and alignment evaluation method show strong variation with different penalty combinations. For each gap open penalty, retrospective performance decreases with increasing gap extension penalty. This can be explained by the alignment process itself. The optimal alignment is the combination of positive scores for matches and negative scores for mismatches and gaps that is highest in sum. If gap penalties exceed mismatch scores, gaps will decrease the alignment score more than mismatches, thus increasing the number of mismatches. This results in alignments dominated by mismatches due to this effect and not because of the exchange of functional groups in the molecular graph. The resulting alignments do not reflect molecular similarity anymore and decrease virtual screening accuracy.

The averaged BEDROC score of the best performing gap penalty combination is given in Table 6 for each canonization algorithm and alignment evaluation method. The performance of baseline PhAST (Weininger canonization, gap open penalty = 3, gap extension penalty = 1, alignment evaluation PID₁) is 0.29 (bottom left in the table). The best performance was 0.40 [MVE canonization using the diffusion kernel ($\beta = 0.4$), gap open penalty = 5, gap extension penalty = 1, alignment evaluation S₂]. To see whether this improvement is significant, we performed the permutation test proposed by Zhao. Table 7 presents the results per target. The lowest fraction of significantly improved screenings is for angiotensin-converting enzyme (ACE) with ~24%. On average, the performance was significantly increased in 71% of all screening experiments. Baseline PhAST performs better only in 21%. In combination with the increased average BEDROC scores, we conclude that the improvement of our method is significant. This is supported by a Kolmogorov–Smirnov test producing a p -value of 1.37×10^{-21} .

Average BEDROC performance for the globally optimal gap penalty combination and gap penalties optimized for each target

Table 7. Permutation Test Results for MVE ($\beta = 0.4$) Canonization with S_2 Versus Weininger Canonization With PID₁.

	No. of queries	MVE 0.4 S_2 (5, 1)	Weininger PID ₁ (3, 1)
ACE	34	24 (24)	59 (56)
COX2	136	56 (56)	40 (40)
DHFR	64	95 (95)	3 (3)
FXA	228	67 (64)	22 (21)
PPAR γ	44	64 (61)	16 (14)
THR	183	90 (90)	6 (5)
Total	689	71 (70)	21 (20)

Shown are the percentages of cases where one contestant performs significantly better than the other at a significance level of 0.05 (0.01).

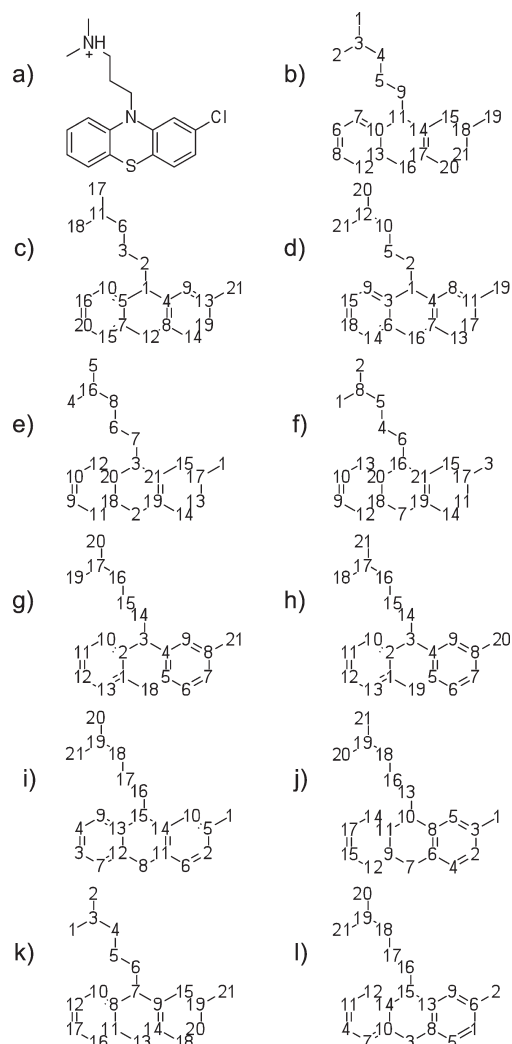
are very close (Table 8), with a maximum difference of 0.024 for COX2. We conclude that the global optimum (gap open penalty = 5, gap extension penalty = 1) provides reasonable default values for practical applications.

We varied two major components of PhAST, the canonization algorithm and the alignment evaluation method. Changing both improved baseline PhAST significantly. But which of these two variables is more important? The lowest score achieved with PID₁ (baseline PhAST) is 0.24 (modified Jochum–Gasteiger canonization), the highest score is 0.31 (Weininger canonization). With S_2 (best performance), even the lowest score of 0.35 (PCA canonization) lies above the highest score of PID₁. The highest score with S_2 for alignment evaluation is 0.40 (MVE canonization with diffusion kernel, $\beta = 0.4$). In 32 of 33 cases, the alignment evaluation methods based on the alignment score yield higher values with the same canonization algorithm than PID₁. For PID₂, this is true for all 33 cases. In Table 6, the mean coefficient of variation (σ/μ)⁶ of the columns is 0.09, whereas that of the rows is 0.27. Varying the alignment evaluation method, therefore, influences the performance three times stronger than varying the canonization method. We conclude that the choice of the alignment evaluation method is more important than the choice of the canonization algorithm.

Table 8. BEDROC Scores ($\alpha = 20$) Per Target for Global and Target-Optimal Gap Penalties.

	Average BEDROC with global optimum gap penalties (5, 1)	Average BEDROC with optimum gap penalties per target	
ACE	0.4034	0.4081	(2, 1)
COX2	0.4011	0.4251	(11, 2)
DHFR	0.5654	0.5704	(9, 1)
FXA	0.3563	0.3676	(2, 1)
PPAR γ	0.2612	0.2612	(5, 1)
THR	0.4130	0.4165	(2, 1)

Global optimum gap penalties are gap open penalty 5 and gap extension penalty 1. The best performing gap penalties per target are shown in parentheses. Higher scores indicate better virtual screenings performance.

**Figure 5.** Canonical labels for (a) chlorpromazine with (b) PCA, (c) Jochum and Gasteiger algorithm, (d) modified Jochum and Gasteiger algorithm, (e) Weininger algorithm, (f) modified Weininger algorithm, (g) Prabhakar algorithm, (h) modified Prabhakar algorithm, (i) MVE with diffusion kernel and diffusion parameter 0.4, (j) MVE with Euclidean distance kernel, (k) Laplacian Eigenmaps, and (l) Isomap.

Neighborhood Preservation

As an example, Figure 5 presents the canonical labels generated for chlorpromazine with each of the 11 compared canonization algorithms. Both versions of the Prabhakar algorithm have a tendency to label consecutive paths in the graph. For both versions of the Gasteiger algorithm, the ground concept of number-

Table 9. Preserved Neighborhood Relations.

Algorithm	No. of preserved neighborhoods	% Preservation
Jochum Gasteiger	20,587	8
Jochum Gasteiger modified	22,801	9
Laplacian Eigenmaps	129,082	49
Isomap	105,152	40
MVE diffusion kernel 0.4	79,289	30
MVE Euclidean kernel	82,363	31
PCA	89,465	34
Prabhakar	163,691	62
Prabhakar modified	174,861	66
Weininger	24,734	9
Weininger modified	25,866	10

ing the vertices in spheres around the most buried atom is recognizable. The other methods tend to spread the labels in a non-intuitive manner. To further assess the differences between canonization algorithms, we analyzed to which extent each method is capable of preserving neighborhood relations during the reduction of the two-dimensional graph of potential pharmacophoric points to its one-dimensional form, the PhAST-sequence. For each method, 264,220 neighborhood relations were checked. The results are summarized in Table 9.

The large number of preserved neighborhoods with both variants of the Prabhakar algorithm (original: 62%, modified: 66%) is no surprise as both perform a depth-first search with a complex set of rules to decide which vertex is visited next. The creation of paths of consecutive canonical labels is inherent to this approach. The modified version preserves even more neighborhoods as the original algorithm, because it lacks the removal of terminal atoms as initial steps. As a consequence, paths through the molecule can include more atoms, and the fragments with consecutive canonical labels may be elongated.

Both variants of the Jochum–Gasteiger algorithm preserve least neighborhoods (original: 8%, modified: 9%). This is no surprise as well, because originating from the most buried vertices in the graph, they work in spheres around this centre. In this approach, it cannot be anticipated that the resulting canonical labels reflect neighborhoods to a high extent. If a vertex v_i in a sphere with n_i vertices receives canonical label x , all adjacent vertices from the next sphere of size n_j will get assigned canonical labels that are bound between $x + 1$ and $x + (n_i - 1) + n_j$. But the algorithm does not guarantee that these canonical labels will be directly subsequent to x or to each other. As with the Prabhakar algorithm, the modified version preserves more neighborhoods because of the fact that the terminal atoms were not treated separately.

As the number of connected vertices is the main criterion used for prioritization in the Weininger algorithm, it tends to work its way from the outside to the inside of a molecule. The initial equivalence classes are created based on further properties regarding all atoms at once, not limited to a certain subset as the vertices connected to the last labeled vertex as in the Prabhakar algorithm or atoms with the same buriedness as the Jochum–Gasteiger method. So, it does not group atoms by their affiliation to

a certain region of the molecule but by the similarity of their overall properties. Further on, the algorithm divides vertices from the same equivalence class into unique subsets comprising only one vertex, so this initial partition cannot be reversed. This behavior is reflected in the low preservation of neighborhoods with 9% for the original and 10% for the modified version.

All methods for dimensionality reduction only moderately preserve the neighborhoods. For MVE (diffusion kernel: 30%, Euclidean distance kernel: 31%), Isomap (40%), and Laplacian Eigenmaps (49%), this was expected, as they were developed to preserve local distances between neighboring points in datasets as good as possible. They do not work in a greedy approach like the depth-first search used by the Prabhakar algorithm. By preserving the distances between neighboring pairs, distances between nonadjacent vertices may be changed. So, it was anticipated that the degree of preservation is lower than for the Prabhakar algorithm. For PCA (34%), however, this result is surprising. During PCA projection neighborhoods of vertices are not regarded explicitly, and different parts of a molecule may collapse in the same region of the PhAST-sequence, merging vertices from different parts of a molecule in the process.

MVE with the diffusion kernel that performed best in the retrospective comparison does not perform best in neighborhood preservation. The Prabhakar algorithm, which in the modified version preserves the most neighborhoods, does not perform best in the retrospective comparison. Pearson's correlation coefficient between the retrospective results using the best performing alignment evaluation method S_2 , and the percentage of preserved neighborhoods during the reduction of the two-dimensional graph to a PhAST-sequence is 0.46. So, despite the fact that different approaches for graph canonization yield notably different results in this analysis; the percentage of preserved neighborhoods seems unsuited to explain why MVE performed best in the retrospective comparison.

For each vertex pair in the two-dimensional graph of pharmacophoric points, we checked whether their corresponding symbols are adjacent to each other in the PhAST-sequence. If not,

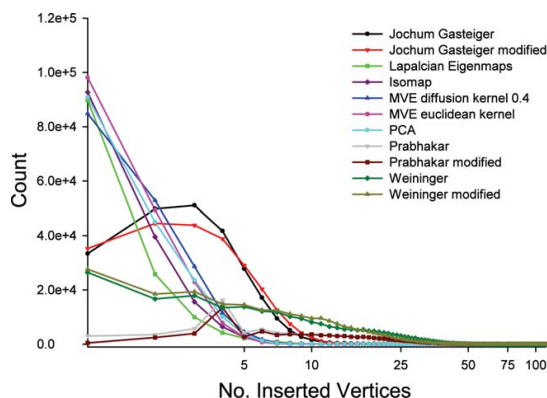


Figure 6. Number of inserted vertices between vertex pairs that are neighbors in the molecule graph, but not the PhAST-sequence. X-axis logarithmic to emphasize the interval in which the behavior of the canonization algorithms diverges most.

Table 10. Percentages of Neighborhood Relations Preserved and Changed Between PhAST-Sequences When Modifying a Molecule by Attaching Fragments.

	% Preserved	% Preserved original orientation	% Preserved transversed orientation	% Changed	% Changed original orientation	% Changed transversed orientation
Jochum Gasteiger	43.41	42.76	0.65	56.59	44.64	11.95
Jochum Gasteiger modified	40.65	39.47	1.17	59.35	46.43	12.93
Laplacian Eigenmaps	69.86	53.61	16.24	30.14	23.99	23.99
Isomap	64.63	43.25	21.38	35.37	24.55	10.82
MVE diffusion kernel 0.4	65.52	42.70	22.82	34.48	22.84	11.64
MVE Euclidean kernel	53.62	36.54	17.07	46.38	29.47	16.91
PCA	38.42	29.46	8.96	61.58	44.61	16.97
Prabhakar	75.60	72.76	2.84	24.40	19.25	5.15
Prabhakar modified	81.54	77.81	3.74	18.46	14.24	4.22
Weininger	62.03	62.01	0.01	37.97	36.75	1.22
Weininger modified	56.18	55.21	0.97	43.82	40.31	3.51

For both cases, the percentage of neighborhood relations in original orientation and in transversed orientation is shown in addition.

we counted by how many symbols they are separated. The result is presented in Figure 6, exact values are available from Table S1 in the Supporting Information. Overall, all methods for dimensionality reduction behave similarly. They separate neighboring vertices in the graph more often by only one vertex in the PhAST-sequence compared to the other canonization algorithms. For separations consisting of more than five vertices, they have the lowest count of occurrence among all methods. As both variants of the Prabhakar algorithm have the highest neighborhood preservation, their overall count of insertions between neighboring vertices is the lowest. They perform a depth-first search, and only reaching a dead end in this search can cause an event counted in this experiment. In case of these two algorithms, the number of vertices inserted between neighboring vertices gives the path length from a dead end in the depth-first search and the next unlabeled vertex.

The Jochum–Gasteiger algorithm works in spheres around the most buried vertex. Vertices from the same sphere receive consecutive labels. This explains why both variants of the algorithm have a high number of insertions with less than 10 inserted vertices. Most molecules analyzed here do not possess enough vertices to create spheres of more than 10 members, so the count of insertions of this size is very low.

Both variants of the Weininger algorithm lead to a medium number of insertions of sizes 1–7, but from thereon, they have the highest count for insertions. MVE performs best in retrospective studies for our method. But as well as the percent of preserved neighborhoods, the number of vertices inserted between originally neighboring vertices in the graph during the generation of the PhAST-sequence seems not to be suitable to explain this good performance.

We have demonstrated that different approaches for graph canonization show different behavior regarding the number of vertices they insert between vertices in the PhAST-sequences that were connected in the two-dimensional graph of pharmacophoric points. MVE with diffusion kernel turned out to be the best performing method in the retrospective comparison.

Robustness Against Structural Modification

We tested the robustness of the compared canonization methods by comparing the PhAST-sequence generated from a molecule with that generated from a molecule similar to the original but with a slight structural modification. For each pair of neighboring vertices in the original sequence, we checked whether they remain adjacent in the modified PhAST-sequence and whether they changed their relative orientation. We used six fragments for chemical structure modification (see Fig. 1) that were attached individually and in pairs. Each original PhAST-sequence was compared with 135 variants, 6699 molecules from the COBRA molecule library were investigated that way. The results are presented in Table 10.

Both variants of the Jochum–Gasteiger algorithm preserve around 40% of the neighborhoods from the original PhAST-sequence in the modified variants. Nearly all of these preserved relationships are kept in the original orientation, only a small amount of transversions is generated (original: 0.65%, modified: 1.17%). Nonpreserved neighborhoods are kept mostly in the original orientation as well (original: ~45%, modified: ~46%), enabling the global sequence alignment to compensate these changes by inserting gaps. Only around 12% of all neighborhood relations are not kept and transversed at the same time.

All methods for nonlinear dimensionality reduction keep more neighborhood relations but introduce transversions at the same time to a higher extent, foremost MVE in combination with the diffusion kernel with over 22% transversions. In case of disrupted neighborhoods, the fraction of created transversions is as high as for the Jochum–Gasteiger algorithm or even higher.

The aim of these algorithms for nonlinear dimensionality reduction is to keep pairwise distances between neighboring points while embedding a set of data points in a lower-dimensional space. So, they only consider relationships between pairs of points. Distances between two points are kept even if these points switch coordinates. So, these methods introduce a high amount of transversions, because this is a valid operation in their functioning. PCA preserves the least neighborhoods from the

Table 11. Time (Seconds) to Canonize the COBRA Library (Version 6.1, $n = 8,311$) on a Single CPU.

	Total	Mean	Max	Min	σ
Jochum Gasteiger	95.61	0.01150	2.93918	0.00007	0.07857
Jochum Gasteiger modified	116.65	0.01404	2.85762	0.00009	0.08386
Laplacian Eigenmaps	21.48	0.00258	0.08001	0.00019	0.00327
Isomap	20.32	0.00245	0.07376	0.00017	0.00320
MVE diffusion kernel 0.4	9639.19	1.15981	96.29566	0.08336	2.72437
MVE Euclidean kernel	10045.93	1.20875	101.30420	0.09399	2.86952
PCA	2.02	0.00024	0.03797	0.00013	0.00073
Prabhakar	35904.97	4.32017	3159.95603	0.00005	66.66446
Prabhakar modified	135843.44	16.34502	23524.26899	0.00005	402.00449
Weininger	4.46	0.00054	0.01046	0.00006	0.00048
Weininger modified	4.51	0.00054	0.07926	0.00006	0.00098

original to the modified PhAST-sequence, but most of them in original orientation, not as transversions. It introduces transversions in the disrupted neighborhoods to an extent comparable to the nonlinear methods for dimensionality reduction.

Both variants of the Weininger algorithm introduce the lowest fraction of transversions (original version: 1%, modified version: 4%). Neighborhood preservation is well above the Jochum–Gasteiger method.

Preservation of neighborhoods between related PhAST-sequences is highest for both variants of the Prabhakar algorithm (original version 76%, modified version 82%). This indicates that paths in the original and altered molecules are very similar. This is in perfect agreement with the low-transversion rate in preserved (3% and 5%) and nonpreserved neighborhoods (4% and 4%).

The correlation between retrospective results using the best performing alignment evaluation method S_2 and (i) the percentage of preserved neighborhoods is 0.57, (ii) the percentage of kept but transversed neighborhoods is 0.65, and (iii) the percentage of neighborhoods that are disrupted and transversed is 0.09. This indicates that transversions do not affect performance as drastically as we expected. The sequence alignment uses only mutations and insertions/deletions; a transversion can, thus, be treated by two mismatches or a combination of gap and mismatch. This explanation is backed up by (iii). None of the correlations is sufficiently strong to qualify the corresponding property as necessary for “good” retrospective results. However, when both variants of the Prabhakar algorithm are omitted as outliers, the correlation (i) increases to 0.93. We interpret this observation as an indication that the Prabhakar algorithm differs from the other canonization approaches (omitting other algorithms does not increase correlation as much: without the Jochum–Gasteiger algorithms: 0.39, without the Weininger algorithms: 0.57, and without the methods for nonlinear dimensionality reduction: 0.65). Indeed, there is such a difference: The subset of vertices from which the next vertex is chosen is the smallest of all approaches because of the depth-first like canonization process of the Prabhakar algorithm. The number of candidates is four or even less in most cases because of the distribution of vertex degrees in molecular graphs.³¹ The Jochum–Gasteiger algorithm limits the number of candidates by the size of the current sphere, which is potentially larger than four levels.

For the Weininger algorithm, the limit is given by the number of vertices with the same properties, which typically exceeds four as well. For the dimensionality reduction methods, the next vertex can potentially be chosen from all remaining vertices.

Until this point, we have combined the results for single and pairwise modifications of molecules. Treating both cases separately does not dramatically change the picture (Tables S3 and S4 of the Supporting Information). Both the Jochum–Gasteiger and the Weininger algorithms introduce the fewest transversions. Both versions of the Prabhakar algorithm preserve neighborhoods best, followed closely by the methods for nonlinear dimensionality reduction that preserve most neighborhood relations but not in the original orientation. PCA preserves neighborhoods least but does not introduce as many transversions as the nonlinear methods. Calculating Pearson’s correlation coefficient between the retrospective evaluation scores and the percentage of neighborhoods kept in original orientation, kept in transversed orientation, and neighborhoods changed and transversed at the same time reveals in both cases a slightly different relationship as the combined evaluation. Using only the results of single (double) modifications, the correlation coefficient between retrospective performance and percentage of neighborhoods preserved is 0.57 (0.57). In case of transversed neighborhoods, the correlation coefficient is 0.64 (0.66). The major difference to the combined case is the relationship between retrospective performance and percentage neighborhoods changed and transversed, which correlates with -0.24 (-0.27).

In summary, transversions in the PhAST sequences of similar molecules do not affect the performance to a great extent. None of the investigated properties correlate strongly with retrospective virtual screening performance.

Canonization Time

Computational efficiency is important for rapid virtual screening. We compared our implementation of the canonization algorithms with respect to the time needed to process the COBRA library on a single central processing unit (CPU) of our cluster, atom typing excluded (Table 11). With only ~ 2 s for all 8311 molecules, PCA was fastest. Both variants of the Weininger algorithm are fast with a time requirement of about 4 s. The Jochum–Gasteiger algorithm and all methods for nonlinear

dimensionality reduction take more time but are still feasible. For a prospective application, the Prabhakar algorithm might be too slow. For a medium-sized screening library with 0.5×10^6 compounds, canonization with this algorithm would take 25 days for the original and 95 days for the modified version (not including atom-typing). MVE with the diffusion kernel (best retrospective performance) would need ~ 7 days.

Conclusions and Outline

The canonization algorithm influences the performance of PhAST in virtual screening, although to a minor extent. None of the investigated properties of canonization algorithms seems useful as an *a priori* indicator of retrospective virtual screening performance by PhAST. The best retrospective performance was achieved using MVE with the diffusion kernel ($\beta = 0.4$), gap open penalty = 5, and gap extension penalty = 1. Different kernel functions vary MVE performance only slightly. Future work could investigate alternative canonization approaches based on the molecular graph³² or MVE using other kernels like p -step random walks³³ to further improve performance. Kernels operating on the graph topology (as opposed to spatial vertex coordinates) have the advantage of being independent from the used layout/conformation algorithm. We used covalent bonds to define atom neighborhoods. In the original applications of these algorithms, neighborhoods were defined by connectivity algorithms like k -nearest neighbors,²² b -matching,²³ or ϵ -balls. Using these instead of covalent bonds would render PhAST-sequences independent from the original connectivity, possibly increasing the chance for scaffold-hopping. The recently developed structure preserving embedding method by Shaw and Jebara³⁴ preserves global connectivity and might be a promising candidate for further investigation. This technique already showed good results in embedding 3D structures into two dimensions.³⁴

MVE inserts many transversions into PhAST-sequences of similar molecules. Global sequence alignment can treat these only by mutations and insertions/deletions, and thus might not be the best metric in this situation. Other string metrics such as the Damerau–Levenshtein-distance³⁵ that are capable of using transversions as well as mutations and insertions/deletions might be promising alternatives, not only for PhAST but also for other string representations like SMILES.

Our study demonstrated that the alignment evaluation method influences performance more than the canonization algorithm. For all canonization methods, the alignment evaluation by alignment score yielded better results than the sequence identity calculated from the alignment. This has also been observed by studies on the alignment of protein sequences⁸; there, significance-based methods perform even better than the actual alignment score. Some techniques originally developed for local alignments seem promising for global alignments also.^{36–38} A first simple step in this direction could be the use of Z-scores.^{1,39} Until recently, one had to create a population of alignment scores by shuffling and realigning the originally compared sequences to estimate mean and standard deviation of alignment scores from alignments of random sequences. Booth et al. showed that it is possible (for the ungapped case) to calculate

mean and standard deviation efficiently, avoiding the time-consuming realignment step.⁴⁰

An important parameter of PhAST that was not changed in our study is the score matrix used to score matches and mismatches in the alignments. It directly influences the alignments, and thus the similarity score as well. The systematic development of a new score matrix that no longer depends on chemical intuition alone will be the subject of our future studies. Krier and Hutter⁴¹ recently proposed a process for building a scoring scheme based on aligning SMILES of molecular fragments. Their score matrix reflects the frequencies of chemical replacements in pharmaceutical substances. For PhAST, a similar approach might be possible based on pharmacophoric points, resulting in a score matrix close to the original concept of Dayhoff et al.⁴² Modification of the pharmacophoric points is another option, one should address at the same time.

With the alignment score as a measure for the evaluation of global alignments (instead of percent identity) the weighting of the influence of certain pharmacophoric points seems reasonable. These points could represent interactions that are necessary for binding. By upweighting the match and mismatch scores of important pharmacophoric points, one could force isofunctional points to be matched. If no such points exist, key interactions are missing resulting in a low score. Incorporating domain knowledge in this way could further improve the performance of PhAST.

Acknowledgments

V.H. is grateful for a Ph.D. scholarship granted by Merck KGaA.

References

1. Hähnke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. *J Comput Chem* 2009, 30, 761.
2. Weininger, D.; Weininger, A.; Weininger, J. L. *J Chem Inf Comput Sci* 1989, 29, 97.
3. Needleman, S. B.; Wunsch, C. D. *J Mol Biol* 1970, 48, 443.
4. Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis*; Cambridge University Press: Cambridge, 1998.
5. Truchon, J. F.; Bayly, C. I. *J Chem Inf Model* 2007, 47, 488.
6. Pearson, K. *Phil Trans R Soc* 1896, 187, 253.
7. Kendall, M. *Biometrika* 1938, 30, 81.
8. Brenner, S. E.; Chothia, C.; Hubbard, T. J. P. *Proc Natl Acad Sci USA* 1998, 95, 6073.
9. Karlin, S.; Altschul, S. F. *Proc Natl Acad Sci USA* 1990, 87, 2264.
10. Jochum, C.; Gasteiger, J. *J Chem Inf Comput Sci* 1977, 17, 113.
11. Prabhakar, Y. S.; Balasubramanian, K. *J Chem Inf Model* 2006, 46, 52.
12. Pearson, K. *Philos Mag* 1901, 2, 559.
13. Belkin, M.; Niyogi, P. *Neural Comput* 2003, 15, 1373.
14. Tenenbaum, J. B.; de Silva, V.; Langford, J. C. *Science* 2000, 290, 2319.
15. Shaw, B.; Jebara, T. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*; Omnipress: Madison, 2007.
16. Schneider, P.; Schneider, G. *QSAR Comb Sci* 2003, 22, 713.
17. Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. *BMC Bioinformatics* 2009, 10, 225.

18. Massey, F. J. *J Am Stat Ass* 1951, 46, 68.
19. Proschak, E.; Wegner, J. K.; Schüller, A.; Schneider, G.; Fechner, U. *J Chem Inf Model* 2007, 47, 295.
20. Vingron, M.; Waterman, M. S. *J Mol Biol* 1994, 235, 1.
21. Morgan, H. L. *J Chem Doc* 1965, 5, 107.
22. Cover, T.; Hart, P. *IEEE Trans Info Theory* 1967, 13, 21.
23. Müller-Hannemann, M.; Schwartz, A. *J Exp Algor* 2000, 5.
24. Floyd, R. W. *Comm ACM* 1962, 5, 345.
25. Warshall, S. *J ACM* 1962, 9, 11.
26. Schölkopf, B.; Smola, A.; Müller, K. *Neural Comput* 1998, 10, 1299.
27. Kondor, R. I.; Lafferty, J. D. *Proceedings of the Nineteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, 2002.
28. Tsuda, K.; Noble, W. S. *Bioinformatics* 2004, 20, 326.
29. Doolin, D.; Dongarra, J.; Seymour, K. *Sci Program* 1999, 7, 111.
30. Borchers, B. *Opt Meth Soft* 1999, 11, 613.
31. Rupp, M. *Kernel Methods for Virtual Screening*; Johann Wolfgang Goethe-University: Frankfurt am Main, 2009.
32. Faulon, J. L.; Collins, M. J.; Carr, R. D. *J Chem Inf Comput Sci* 2004, 44, 427.
33. Smola, A. J.; Kondor, R. I. *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*; Springer: Berlin and Heidelberg, 2003.
34. Shaw, B.; Jebara, T. *Proceedings of the 26th International Conference on Machine Learning*; Omnipress: Madison, 2009.
35. Damerau, F. J. *Commun ACM* 1964, 7, 171.
36. Newberg, L. A. *J Comput Biol* 2008, 15, 1187.
37. Chia, N.; Bundschuh, R. *J Comput Biol* 2006, 13, 429.
38. Hartmann, A. K. *Phys Rev E* 2002, 65, 056102.
39. Lipman, D. J.; Pearson, W. R. *Science* 1985, 227, 1435.
40. Booth, H. S.; Mairdonald, J. H.; Wilson, S. R.; Gready, J. E. *J Comput Biol* 2004, 11, 616.
41. Krier, M.; Hutter, M. C. *J Chem Inf Model* 2009, 49, 1280.
42. Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1978.

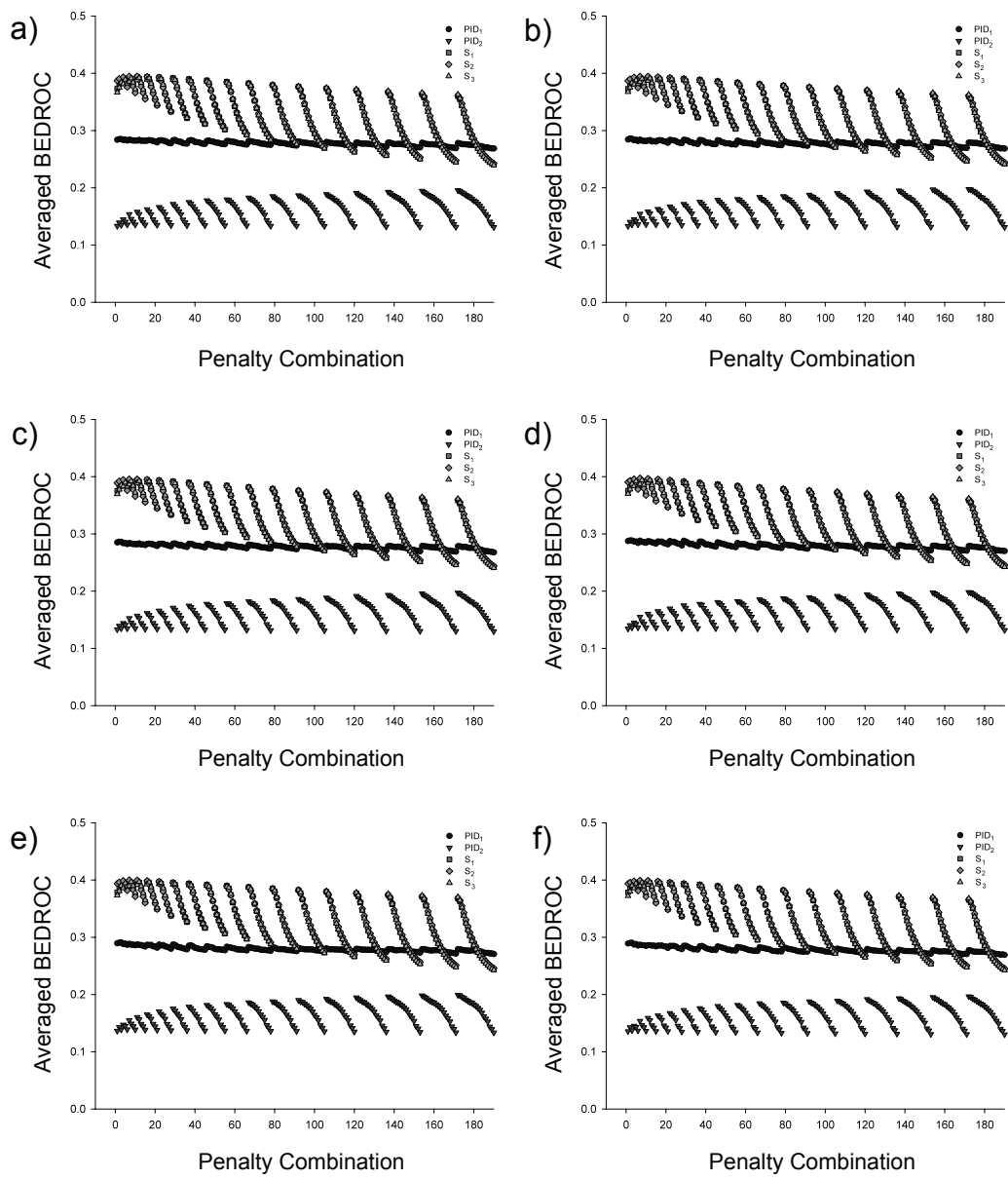


Figure S1 a-f BEDROC ($\alpha=20$) scores for combinations of alignment evaluation methods and gap penalties for Minimum Volume Embedding with different settings for the diffusion parameter β . **a)** $\beta = 0.01$, **b)** $\beta = 0.1$, **c)** $\beta = 0.2$, **d)** $\beta = 0.3$, **e)** $\beta = 0.4$, **f)** $\beta = 0.5$.

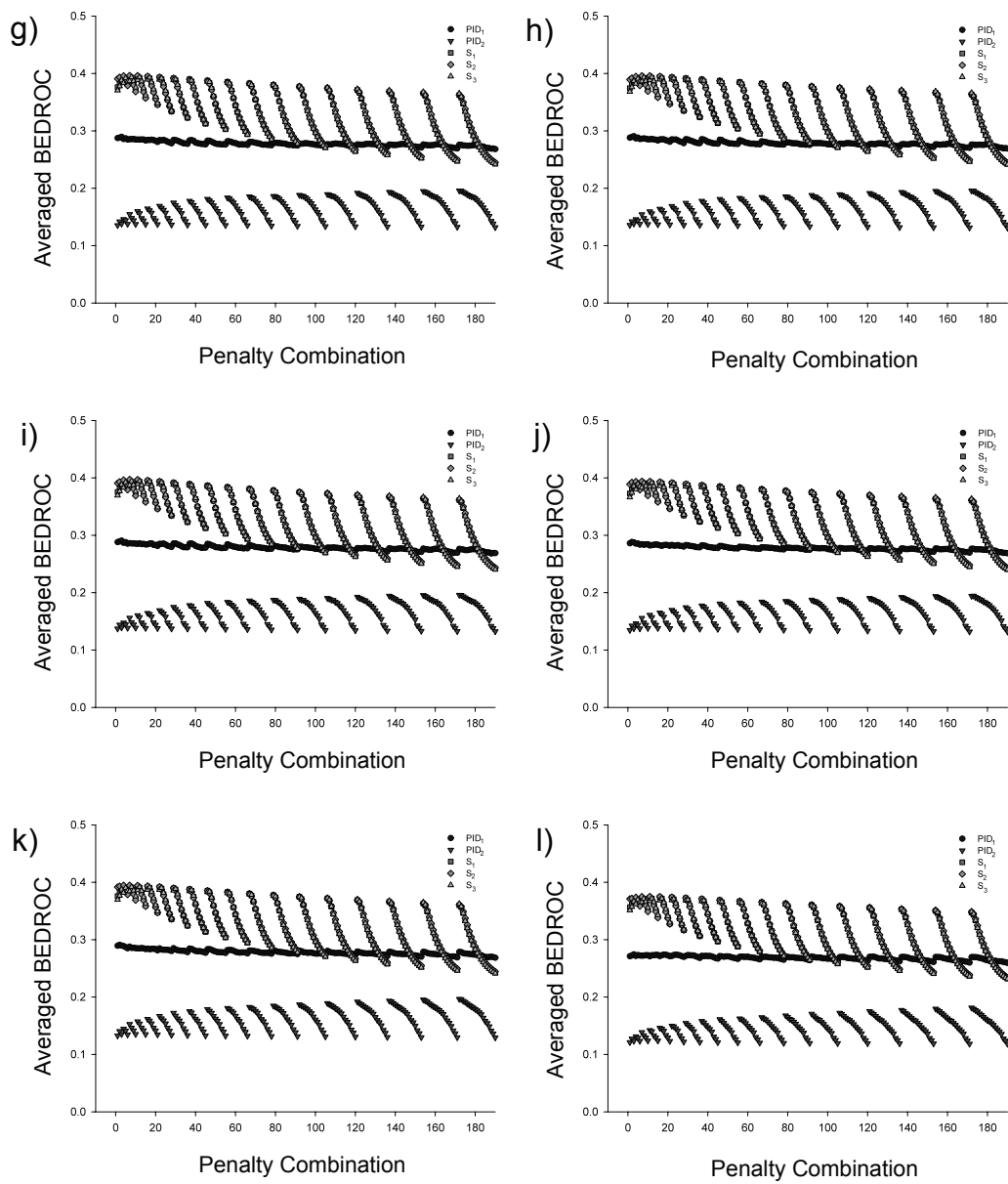


Figure S1 g-l BEDROC ($\alpha=20$) scores for combinations of alignment evaluation methods and gap penalties for Minimum Volume Embedding with different settings for the diffusion parameter β . **g)** $\beta = 0.6$, **h)** $\beta = 0.7$, **i)** $\beta = 0.8$, **j)** $\beta = 0.9$, **k)** $\beta = 1.0$, **l)** $\beta = 10$.

Table S2 Number of inserted vertices between vertex pairs neighbored in the molecular graph but not in the PhAST-sequence (COBRA library, n=8,311). The highest number of inserted vertices observed was 110. JG (m) = Jochum and Gasteiger algorithm (modified), LE = Laplacian Eigenmaps, Iso = Isomap, MVE (D 0.4) [E] = Minimum Volume Embedding (diffusion kernel with diffusion parameter 0.4) [euclidean distance kernel], PCA = principal component analysis, P (m) = Prabhakar algorithm (modified), W (m) = Weininger algorithm (modified).

No. Inserted Vertices	JG	JGm	LE	Iso	MVE D 0.4	MVE E	PCA	P	Pm	W	Wm
1	33368	35263	89758	92706	84852	98122	90823	3000	453	26429	27637
2	49765	44516	25799	39531	52831	49390	44994	3542	2463	16662	18500
3	51027	43826	9996	15597	28625	22973	23575	5697	3934	17894	19349
4	41708	38830	4236	6390	11451	7835	10105	16258	13478	13439	14756
5	27821	29045	2206	2606	4557	2549	3500	4237	2729	13821	14510
6	17127	20340	1316	1211	1705	746	1265	5500	4706	12159	12575
7	9484	12578	872	521	561	193	365	4157	3373	11835	12486
8	5126	7505	439	228	203	37	90	3890	3656	10228	11123
9	2658	4159	224	142	91	8	33	3626	3506	9436	10403
10	1597	2484	141	71	36	4	4	3645	3635	8127	9548
11	958	1176	68	38	9	0	1	3316	3278	7488	9571
12	580	727	34	10	6	0	0	3284	3220	6506	8548
13	344	386	25	6	2	0	0	2943	3016	6041	7521
14	312	241	13	2	0	0	0	2981	3010	5540	6473
15	241	116	5	1	2	0	0	2755	2734	5526	6072
16	181	88	0	3	0	0	0	2587	2615	5182	5237
17	155	43	3	3	0	0	0	2524	2713	5372	5066
18	110	31	1	1	0	0	0	2366	2551	4902	4565
19	130	18	0	1	0	0	0	2186	2298	4796	4075
20	92	13	0	0	0	0	0	2020	2115	4477	3728
21	113	11	1	0	0	0	0	1951	2045	4254	3344
22	94	7	1	0	0	0	0	1822	1800	3958	3071
23	87	3	0	0	0	0	0	1625	1650	3530	2576
24	62	4	0	0	0	0	0	1565	1572	3342	2182
25	80	5	0	0	0	0	0	1430	1416	3059	2010
26	43	4	0	0	0	0	0	1349	1229	2740	1723
27	64	0	0	0	0	0	0	1181	1199	2562	1537
28	37	0	0	0	0	0	0	1079	1087	2178	1386
29	46	0	0	0	0	0	0	1012	965	2257	1205
30	29	0	0	0	0	0	0	852	873	1763	973
31	28	0	0	0	0	0	0	724	723	1630	863
32	23	0	0	0	0	0	0	668	688	1511	808
33	21	0	0	0	0	0	0	588	558	1389	718
34	14	0	0	0	0	0	0	570	530	1122	587
35	14	0	0	0	0	0	0	476	474	1108	388
36	7	0	0	0	0	0	0	399	409	884	348
37	14	0	0	0	0	0	0	403	286	801	352
38	4	0	0	0	0	0	0	280	272	694	289
39	8	0	0	0	0	0	0	255	284	564	234
40	4	0	0	0	0	0	0	220	214	466	151
41	9	0	0	0	0	0	0	164	169	417	223
42	9	0	0	0	0	0	0	142	172	331	170
43	8	0	0	0	0	0	0	155	138	372	161
44	13	0	0	0	0	0	0	87	143	232	153

No. Inserted Vertices	JG	JGm	LE	Iso	MVE D 0.4	MVE E	PCA	P	Pm	W	Wm
45	0	0	0	0	0	0	0	111	106	240	129
46	2	0	0	0	0	0	0	75	82	197	125
47	2	0	0	0	0	0	0	86	80	228	57
48	3	0	0	0	0	0	0	55	62	165	51
49	1	0	0	0	0	0	0	75	58	143	68
50	3	0	0	0	0	0	0	49	49	141	43
51	3	0	0	0	0	0	0	51	40	93	34
52	0	0	0	0	0	0	0	41	45	72	41
53	0	0	0	0	0	0	0	52	55	74	60
54	1	0	0	0	0	0	0	44	30	62	22
55	1	0	0	0	0	0	0	38	39	83	16
56	0	0	0	0	0	0	0	26	32	82	27
57	0	0	0	0	0	0	0	28	34	40	28
58	0	0	0	0	0	0	0	16	27	75	30
59	0	0	0	0	0	0	0	22	21	68	14
60	0	0	0	0	0	0	0	23	30	75	29
61	0	0	0	0	0	0	0	18	19	46	39
62	0	0	0	0	0	0	0	13	20	36	24
63	0	0	0	0	0	0	0	18	9	39	25
64	0	0	0	0	0	0	0	19	12	14	16
65	1	0	0	0	0	0	0	14	16	45	26
66	0	0	0	0	0	0	0	22	14	33	16
67	0	0	0	0	0	0	0	11	14	20	25
68	0	0	0	0	0	0	0	11	10	76	35
69	0	0	0	0	0	0	0	10	10	41	19
70	0	0	0	0	0	0	0	12	8	14	12
71	0	0	0	0	0	0	0	7	3	11	39
72	0	0	0	0	0	0	0	12	5	10	10
73	0	0	0	0	0	0	0	7	2	7	11
74	0	0	0	0	0	0	0	6	3	17	2
75	0	0	0	0	0	0	0	2	2	32	11
76	0	0	0	0	0	0	0	1	7	28	4
77	0	0	0	0	0	0	0	4	7	19	35
78	0	0	0	0	0	0	0	3	3	12	0
79	0	0	0	0	0	0	0	3	1	30	0
80	0	0	0	0	0	0	0	3	5	1	0
81	0	0	0	0	0	0	0	4	2	0	1
82	0	0	0	0	0	0	0	4	1	9	0
83	0	0	0	0	0	0	0	1	2	0	0
84	0	0	0	0	0	0	0	1	3	1	0
85	0	0	0	0	0	0	0	2	3	4	0
86	0	0	0	0	0	0	0	1	0	5	0
87	0	0	0	0	0	0	0	1	0	23	0
88	0	0	0	0	0	0	0	1	1	8	0
89	0	0	0	0	0	0	0	1	0	1	0
90	1	0	0	0	0	0	0	1	0	2	0
91	0	0	0	0	0	0	0	0	2	0	0
92	0	0	0	0	0	0	0	1	0	0	0
93	0	0	0	0	0	0	0	0	1	1	9
94	0	0	0	0	0	0	0	2	0	4	4
95	0	0	0	0	0	0	0	0	0	0	0
96	0	0	0	0	0	0	0	1	1	0	13

No. Inserted Vertices	JG	JGm	LE	Iso	MVE D 0.4	MVE E	PCA	P	Pm	W	Wm
97	0	0	0	0	0	0	0	0	0	0	4
98	0	0	0	0	0	0	0	2	0	0	0
99	0	0	0	0	0	0	0	1	0	5	1
100	0	0	0	0	0	0	0	1	0	0	3
101	0	0	0	0	0	0	0	0	0	0	0
102	0	0	0	0	0	0	0	1	2	2	0
103	0	0	0	0	0	0	0	0	0	3	1
104	0	0	0	0	0	0	0	1	0	12	0
105	0	0	0	0	0	0	0	0	0	11	0
106	0	0	0	0	0	0	0	1	2	0	0
107	0	0	0	0	0	0	0	0	0	1	0
108	0	0	0	0	0	0	0	1	0	0	0
109	0	0	0	0	0	0	0	0	0	1	0
110	0	0	0	0	0	0	0	1	0	0	0

Table S3 Percentages of neighborhood relations preserved and changed between PhAST-sequences when modifying a molecule by attaching one fragment. For both cases, the percentage of neighborhood relations in original orientation and in transversed orientation is shown in addition.

	% preserved	% preserved original Orientation	% preserved transversed orientation	% changed	% changed original orientation	% changed transversed orientation
Jochum Gasteiger	53.18	52.67	0.51	46.82	37.32	9.50
Jochum Gasteiger modified	51.04	50.06	0.98	48.96	38.67	10.29
Laplacian Eigenmaps	74.78	57.57	17.21	25.22	20.11	5.11
Isomap	70.34	46.14	24.20	29.66	20.57	9.08
MVE diffusion kernel 0.4	70.67	44.66	26.01	29.33	19.25	10.08
MVE euclidean kernel	57.73	38.73	19.00	42.27	26.71	15.55
PCA	44.28	33.89	10.39	55.72	40.16	15.56
Prabhakar	78.84	76.73	2.11	21.16	16.82	4.35
Prabhakar modified	85.14	82.06	3.08	14.86	11.48	3.37
Weininger	71.31	71.30	0.01	28.69	27.98	0.72
Weininger modified	67.50	66.76	0.74	32.50	0.74	2.20

Table S4 Percentages of neighborhood relations preserved and changed between PhAST-sequences when modifying a molecule by attaching two fragments. For both cases, the percentage of neighborhood relations in original orientation and in transversed orientation is shown in addition.

	% preserved	% preserved original Orientation	% preserved transversed orientation	% changed	% changed original orientation	% changed transversed orientation
Jochum Gasteiger	40.62	39.93	0.69	59.38	46.73	12.65
Jochum Gasteiger modified	37.68	36.45	1.23	62.32	48.64	13.68
Laplacian Eigenmaps	68.45	52.48	15.97	31.55	25.10	6.45
Isomap	62.99	42.43	20.57	37.01	25.69	11.32
MVE diffusion kernel 0.4	64.04	42.14	21.91	35.96	23.86	12.09
MVE euclidean kernel	52.44	35.92	16.52	47.56	30.26	17.30
PCA	36.75	28.19	8.56	63.25	45.88	17.37
Prabhakar	74.67	71.62	3.05	25.33	19.95	5.38
Prabhakar modified	80.51	76.59	3.92	19.49	15.03	4.46
Weininger	59.38	59.36	0.02	40.62	39.25	1.37
Weininger modified	52.95	51.92	1.03	47.05	43.17	3.89

Appendix B

Authors: Hähnke, V.
Klenner, A.
Rippmann, F.
Schneider, G.

Title: Pharmacophore Alignment Search Tool: Influence of the Third Dimension on Text-based Similarity Searching

Journal: Journal of Computational Chemistry
Accepted for publication (see letter from the editor)

Letter from the Editor:

Date:03-Dec-2010

Ref.: JCC-10-0498.R1

Dear Prof. Schneider:

I am pleased to inform you that your manuscript, Pharmacophore Alignment Search Tool: Influence of the Third Dimension on Text-based Similarity Searching, is acceptable for publication in the Journal of Computational Chemistry. Thank you for publishing your work in our journal.

We must receive a completed Copyright Transfer Agreement, which can be downloaded at www.wiley.com/go/ctapsus.

Please fax this form, with a cover page bearing the JCC manuscript number, directly to the attention of the Production Staff at (USA) 717-738-9478 or (USA) 717-738-9479.

Thank you for your support of the Journal of Computational Chemistry. I look forward to seeing more of your work in the future.

Sincerely,

Prof. Gernot Frenking
Editor, Journal of Computational Chemistry



Pharmacophore Alignment Search Tool: Influence of the Third Dimension on Text-based Similarity Searching

Journal:	<i>Journal of Computational Chemistry</i>
Manuscript ID:	JCC-10-0498.R1
Wiley - Manuscript type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Haehnke, Volker; ETH Klenner, Alexander; ETH Rippmann, Friedrich; Merck Schneider, Gisbert; ETH, DCHAB
Key Words:	Double dynamic programming, Global alignment, Line notation, Molecular graph, Similarity, Virtual screening

SCHOLARONE™
Manuscripts

Hähnke et al.

1

Pharmacophore Alignment Search Tool: Influence of the Third Dimension on Text-based Similarity Searching

Volker Hähnke,¹ Alexander Klenner,¹ Friedrich Rippmann,² Gisbert Schneider^{1,*}

¹Swiss Federal Institute of Technology (ETH), Institute of Pharmaceutical Sciences, Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland

²Merck KGaA, Merck Serono, Frankfurter Str. 250, D 64293 Darmstadt, Germany

* author to whom correspondence should be sent:

Prof. Dr. Gisbert Schneider, Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, HCI H411, Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland

Email: gisbert.schneider@pharma.ethz.ch

Phone: +41 44 633 7327

Hähnke et al.

2

ABSTRACT

Previously (Hähnke *et al.*, J Comput Chem 2010, 31, 2810) we introduced the concept of non-linear dimensionality reduction for canonization of two-dimensional layouts of molecular graphs as foundation for text-based similarity searching using our Pharmacophore Alignment Search Tool (PhAST), a ligand-based virtual screening method. Here we apply these methods to three-dimensional molecular conformations and investigate the impact of these additional degrees of freedom on virtual screening performance and assess differences in ranking behavior. Best-performing variants of PhAST are compared to 16 state-of-the-art screening methods with respect to significance estimates for differences in screening performance. We show that PhAST sorts new chemotypes on early ranks without sacrificing overall screening performance. We succeeded in combining PhAST with other virtual screening techniques by rank-based data fusion, significantly improving screening capabilities. We also present a parameterization of double dynamic programming for the problem of small molecule comparison, which allows for the calculation of structural similarity between compounds based on one-dimensional representations, opening the door to a holistic approach to molecule comparison based on textual representations.

KEYWORDS

Double dynamic programming; Global alignment; Line notation; Molecular graph; Similarity; Virtual screening

INTRODUCTION

The Pharmacophore Alignment Search Tool (PhAST) is a string-based approach to virtual screening utilizing topological molecule information.^{1,2} It reduces each molecule to an unambiguous linear representation describing its pharmacophore in three steps: i) each non-hydrogen atom in the structure graph is replaced by a potential pharmacophoric point symbol and hydrogen atoms are removed, ii) vertices of this pharmacophoric feature graph receive canonic labels, and iii) vertex symbols are concatenated as a string according to their canonic labels. For virtual screening, both the screening compound collection ('library') and the query molecules are converted, and the resulting PhAST-sequences are compared using pairwise global sequence alignment.³ Molecular similarity is calculated as the ratio of the alignment score and the alignment length for the retrieval of pharmacophorically similar molecules from the compound library.

Previously,² we introduced the concept of canonizing molecular graphs with dimensionality reduction algorithms. In retrospective experiments we identified minimum volume embedding⁴ employing a combination of diffusion kernel⁵ with diffusion parameter 0.4 and covalent connectivity between potential pharmacophoric points as the best-performing canonization algorithm for the application to two-dimensional (2D) molecular graphs. Here, we expand this concept by applying canonization algorithms to three-dimensional (3D) conformations. In addition, we investigate new algorithms for dimensionality reduction in combination with connectivity algorithms defining the edges of the graph created in step (i) of PhAST. The canonization algorithm that performs best for single conformations is evaluated with regard to the impact of multiple conformations on screening performance. In contrast to our previous studies,² we did not perform an optimization of gap penalties for each algorithm but used fixed preferred penalty combinations.

All canonization algorithms and screening methods are evaluated using the COBRA collection of drugs and lead compounds.⁶ For statistical evaluation we use Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) scores ($\alpha = 20$),⁷ a paired permutation test for significance assessment,⁸ and Kendall's τ as rank correlation coefficient.⁹ Differences in PhAST-sequences generated by the same canonization algorithm from two-dimensional layouts and three-dimensional conformations of the same molecule are quantified by calculating their Levenshtein¹⁰ and Damerau-Levenshtein distances.¹¹

Information about the spatial arrangement of pharmacophoric points deduced from three-dimensional conformations can be used in the sequence comparison step of PhAST as well. For this purpose we parameterized double dynamic programming,¹² which – to the best of our knowledge – until now has only been used for the calculation of global pairwise sequence alignments based on structural residue-equivalence of protein sequences.

The main objectives of this study were to i) assess the screening performance of PhAST with different canonization algorithms applied to 2D molecular layouts and 3D conformations, ii) quantify the impact of conformer structure on screening performance, iii) investigate the effect of multiple conformations per molecule, iv) assess the novelty of PhAST compared to established 2D and 3D virtual screening techniques, and v) investigate the effect of using structural information for sequence comparison in PhAST through double dynamic programming.

Hähnke et al.

4

METHODS

Canonization

The atom-typing step of PhAST yields a graph of potential pharmacophoric points. It has the same topology as the original molecular graph without hydrogen atoms yielding a total of n vertices. Each vertex is colored with a symbol corresponding to one out of nine potential pharmacophoric features. Edges correspond to covalent bonds. Canonization is the labeling of the vertices with the natural numbers 1,2,3,..., n . The algorithms compared in this work are described in detail in the supplemental material. Short descriptions are presented in the following paragraph.

Centroid linearization prioritizes vertices by their distance to the geometric centre of a molecule. Performing principal component analysis¹³ (PCA) on vertex coordinates yields the first principal component as a one-dimensional coordinate system. Deterministic non-linear methods for dimensionality reduction applied to 2D layouts and 3D conformations of molecules are Laplacian eigenmaps,¹⁴ Isomap¹⁵ and minimum volume embedding⁴ (MVE). The latter employs a kernel function. Kernels evaluated in this work are a diffusion kernel^{5,16} (DK), a p -step random walk kernel (PRW),¹⁷ a method for calculating inner products from Euclidean coordinates² (referred to as 'Euclidean distance kernel' (EDK)) and a Gaussian radial basis function kernel (RBF).¹⁸ All non-linear methods rely on neighborhood relationships between vertices. As connectivity algorithms used for neighborhood assignment we evaluated covalent bonds (cov) and k nearest neighbors (kNN).¹⁹ Proximity embedding was compared to the deterministic embedding algorithms in a stochastic variant²⁰ (SPE) and based on canonical indices pre-calculated with MVE (MVEPE). These algorithms were compared amongst each other and to results obtained with algorithms that are independent from layouts and conformations of molecular structures. These are the Jochum-Gasteiger,²¹ Weininger,²² and Prabhakar algorithm²³ implemented as described previously.²

Sequence Alignment

Sequence alignment is used in bioinformatics to decide how related two sequences (DNA, RNA, amino acid sequences) are. To create the alignment of two sequences $X = x_1x_2\dots x_n$ and $Y = y_1y_2\dots y_m$, their symbols are matched. Thereby the symbol order is retained and gaps may be inserted to improve the matching (insertion of paired gaps is forbidden). Three cases exist: (i) x_i is aligned to y_j and $x_i = y_j$ (match), (ii) x_i is aligned to y_j and $x_i \neq y_j$ (mismatch), (iii) x_i is aligned to a gap in Y , or y_j is aligned to a gap in X . In protein sequence alignment, matches represent conserved residues; mismatches may arise from mutations, and gaps from insertions or deletions in an assumed evolutionary process of the compared sequences. Consequently, matches are rewarded with a positive score, mismatches are -- depending on the specific case -- either rewarded with a positive score or penalized with a negative score, and gaps are always penalized with a negative score. The optimal alignment is the one with the highest score (summed over the whole alignment). It can be computed using dynamic programming.²⁴ Instead of the original Needleman-Wunsch²⁴ algorithm we employed a faster method described in Durbin *et al.*³ It can be derived from a simple finite state machine and therefore will be referred to as 'FSM algorithm'. We could show that it runs 60% faster than the Needleman-Wunsch algorithm and the calculated alignments are nearly identical.²

Gap Penalties. Previously,^{1,2} we optimized gap penalties in a grid search with 190 penalty combinations. Here all retrospective screenings were carried out with only one combination: Gap open penalty = -5 and gap extension penalty = -1. This decision was made based on earlier findings.² Of all gap open penalties of best-performing combinations

1
2
3 involving the alignment score for alignment evaluation, -5 is the median value of the three
4 values with the highest frequency and so it is least extreme. Gap extension penalty -1 (-2) [-3]
5 occurs with a frequency of 70% (27%) [3%]. As differences in retrospective performance
6 between top-performing combinations are marginal when other parameters remain
7 unchanged, the combination we chose is an educated guess with minor sacrifice in
8 performance if wrong.

9
10 *Alignment Evaluation.* In previous studies² we identified the alignment score
11 normalized to the alignment length to be the best performing alignment evaluation method so
12 far. For this reason we only considered this evaluation method for comparison of PhAST
13 screening performance with different canonization algorithms.
14
15

16 17 **Library Preparation**

18
19 We used the COBRA library of reference compounds⁶ as screening library (version 6.1, 8311
20 compounds). Each compound was protonated using the 'wash' function of MOE (Molecular
21 Operating Environment, v2010.06, Chemical Computing Group Inc., Montreal, Canada). 2D
22 layouts were created for each compound using the 'depict' algorithm of MOE. Single 3D
23 conformations were created for each compound using CORINA (v3.46, Molecular Networks
24 GmbH, Erlangen, Germany) invoking the 'canon' option. Ten 3D conformations were
25 generated for each compound using our stochastic conformer generator SOCGER.²⁵
26
27

28 *The only publicly available dataset compiled specifically for the evaluation of ligand-*
29 *based virtual screening methods is the 'maximum unbiased validation' (MUV) dataset.²⁶ It*
30 *was shown that many methods fail in achieving any significant enrichment for most of the*
31 *targets present in MUV.²⁷ MUV therefore disqualified as a reference dataset for retrospective*
32 *comparisons of PhAST with other methods. The 'directory of useful decoys' (DUD) contains*
33 *actives and decoys for 40 targets.²⁸ DUD was especially designed for the evaluation of*
34 *docking methods. In an attempt to remove analogue bias in the sets of active molecules, only*
35 *actives were filtered according to certain criteria, causing some artificial enrichment.²⁹ A*
36 *second approach processed both, active and decoy compounds,³⁰ but still, a high ratio of*
37 *actives to decoys renders it unfavorable for virtual screening.⁷ In addition, DUD targets are*
38 *limited to structurally resolved proteins, and for example GPCRs are excluded. So we relied*
39 *on our own collection of bioactive reference compounds (COBRA)⁶. COBRA exhibits the*
40 *same degree of scaffold diversity as trade drugs (1.7 compounds per graph scaffold) and may*
41 *thus be considered as a druglike compound set also from a structural perspective. Both MUV*
42 *(3.8) and DUD (6.3) contain more compounds per scaffold on average. Scaffolds were*
43 *determined as graph frameworks defined by Bemis and Murcko.^{31,32}*
44
45
46
47
48

49 **Screening Protocol 1**

50
51 PhAST in combination with each canonization algorithm was used in a series of retrospective
52 screenings. For each target (Table 1) each active was used once as query, resulting in 689
53 screenings. Each screening run was evaluated with the Boltzmann-enhanced discrimination of
54 receiver operating characteristic (BEDROC) metric.⁷ BEDROC scores were calculated with α
55 = 20, the suggested default value for evaluation.⁷ We first evaluated screening performance
56 for each target by averaging the corresponding BEDROC scores. Final retrospective
57 performance is expressed as the mean of these averages. We used the mean of averages to
58 give equal weight to each target although the COBRA library contains unequal numbers of
59 actives for different targets. Each canonization algorithm described in the Methods section
60

Hähnke et al.

6

was tested on 2D graph layouts. Except for MVE DK and MVE PRW all versions were evaluated on single 3D conformations as well.

To assess whether differences in screening performance are significant we employed a paired permutation test⁸ that was recently found to be the most powerful available significance test for this purpose.³³ It has the null hypothesis that *virtual screening method P performs significantly better than method Q*. Assuming p and q are rank lists of actives resulting from the virtual screening methods, the null hypothesis requires that $\text{BEDROC}(p) > \text{BEDROC}(q)$. As each active has two ranks, one in p and one in q , new rank lists p^* and q^* can be created by swapping ranks in p with corresponding ranks in q for each active with a probability of 50%. This was repeated 10,000 times and the frequency of the event that $\text{BEDROC}(p) - \text{BEDROC}(q)$ is less than $\text{BEDROC}(p^*) - \text{BEDROC}(q^*)$ is the type I error rate for the null hypothesis used as p -value for significance estimation. As significance levels we used 0.05 and 0.01.

Assessment of Novelty

Each canonization algorithm based on Euclidean coordinates of vertices was used on 2D layouts of molecular graphs and on 3D single conformations of the complete COBRA library. Besides the retrospective screening performance we assessed differences between canonization algorithms in two and three dimensions based on differences in compound rankings and PhAST-sequences.

Ranking Differences: Screening protocol 1 yielded 689 ranked lists for each canonization algorithm. Rankings of actives were compared using Kendall's rank correlation coefficient.⁹ Before calculating Kendall's τ , ranked lists resulting from a virtual screening were reduced by eliminating all inactive compounds, yielding a ranking only of actives. These reduced lists were used to calculate τ_b that corrects for ties. We first calculated the average rank correlation per target and used the mean of these averages as final measure to express how similar PhAST ranks actives with two different canonization algorithms to give equal weight to each target although the COBRA library contains unequal numbers of actives for different targets.

Sequence Differences. Applying each canonization method on a 2D layout and a 3D conformation of the same molecule yields two PhAST-sequences. If the additional degrees of freedom in three instead of two dimensions have big impact on the canonization process, these two PhAST-sequences should be dissimilar. We measured sequence similarity employing the Levenshtein distance.¹⁰ It is defined as the minimum number of edit operations necessary to transform one sequence into the other with insertion, deletion and substitution of a single symbol being the allowed edit operations. To compare sequences of the whole COBRA library obtained with the same canonization algorithm applied to 2D layouts and 3D conformations we calculated the Levenshtein distance for all 8,311 pairs of PhAST-sequences generated from the same molecule and used the average of these values as final measure of dissimilarity. As observed earlier,² MVE in particular tends to introduce a fourth kind of events in PhAST-sequences of similar molecules: transpositions, defined as the exchange of position between neighboring symbols. Accounting for this fact we used an extension to the original Levenshtein distance, the Damerau-Levenshtein distance¹¹ that allows transpositions as edit operations. As the Damerau-Levenshtein distance uses an additional edit operation the calculated distances should be smaller compared to Levenshtein distances. We used both distances and compared obtained results because it is unknown which one is more suitable for our purpose.

Clustering of Canonization Methods

The averaged rank correlation, the averaged Levenshtein and the averaged Damerau-Levenshtein distance were used to quantify the difference between versions of PhAST with different canonization algorithms. Using these distances we performed a Ward clustering³⁴ of the different PhAST versions to assess the similarity between canonization algorithms and to get an idea whether screening behavior and sequence generation are more influenced by the dimensionality of molecular representation (2D layouts vs. 3D conformations) or the canonization algorithm. Therefore we first selected a representative of each canonization algorithm from the possible parameterizations by using the retrospective performance averaged over both dimensionalities as selection criterion. In case of the Kendall rank correlation we used $(1 - \tau)$ as distance measure.

Data Fusion

Retrospective results obtained with the top-performing canonization algorithms applied to 2D layouts and 3D conformations were combined using data fusion. This way we assessed whether the combination of topological and spatial information can further improve screening performance of PhAST. To do so, each screening described in screening protocol 1 was performed with both versions of PhAST. To avoid complications with different ranges of similarity values and re-scaling steps we chose a data fusion approach that combines the ranks of each compound in both ranked lists.³⁵ To combine two methods, a new ranked list was created according to Eq (1).

$$r_i^* = \min(r_i^{m_1}, r_i^{m_2}) \quad (1)$$

with m_1 the first screening method, m_2 the second screening method and r_i the rank of the i -th compound. This version of PhAST is referred to as PhAST DF.

Multiple Conformations

The retrospective performance of PhAST applied to 3D conformations of molecules with different canonization algorithms was assessed using single 3D conformations generated for each molecule in the COBRA library. But there is more than one possible low-energy conformation for most molecules.³⁶ This is why the canonization algorithm with best retrospective performance on single 3D conformations was re-evaluated with ten conformations per molecule. The retrospective analysis was similar to screening protocol 1 with one modification: All ten conformations of each query molecule were compared to all ten conformations of each screening compound, and the maximum of these 100 similarity scores per molecule comparison was used as final similarity measure. To quantify the influence of the additional degrees of freedom in three dimensions on the generation process of PhAST-sequences we again used alternative sequence distance measures. We calculated the Levenshtein and Damerau-Levenshtein distance for each single conformations of a molecule in the COBRA library to the corresponding ten multi-conformations.

Other Virtual Screening Methods

The best-performing versions of PhAST applied to 2D layouts and 3D conformations of molecules were compared to other popular virtual screening methods in the same series of retrospective screenings as described in screening protocol 1.

(i) MDL MACCS substructural search keys³⁷ were originally developed to encode common substructure features found in organic molecules. Each molecule is represented as a vector of 166 bits corresponding to a predefined set of 166 features, where each 'on' bit indicates the presence of the corresponding feature. We used the implementation of MACCS keys available in MOE, and binary vectors were compared using the Tanimoto coefficient.³⁸

(ii) LINGO^{39,40} is based on the fragmentation of SMILES into overlapping words of length q (LINGOs). Counts of LINGOs generated by the fragmentation of SMILES representations of two molecules were used to quantify molecular similarity between 0 and 1. SMILES were generated using MOE. These were modified by unitizing ring numbers as well as replacing 'Cl' with 'L', and 'Br' with 'R' as suggested for preprocessing.³⁹ We used $q = 4$ as this value showed highest retrospective performance in a comparison of lengths between 1 and 20 (data not shown).

(iii) ESshape3D and ESshape3D HYD are eigenvalue shape fingerprints implemented in MOE. ESshape3D compares 3D shapes made from heavy atoms of a molecule, ESshape3D HYD those from hydrophobic heavy atoms. Both are based on the calculation of eigenvalues for the Euclidean distance matrix between atoms, encoding this eigenspectrum into a fingerprint and using the inverse distance between fingerprints as similarity score.

(iv) The property vector referred to as SIMPLE was used during the creation of the MUV dataset.²⁶ It contains the number of all atoms, heavy atoms, boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus, and sulfur atoms, the number of acceptors, donors, logP, the number of chiral center and the number of ring systems in a molecule. For comparison the Euclidean distance between these vectors was calculated.

(v) The TGD/TGT/TAD/TAT fingerprint family of MOE is based on a common definition of pharmacophoric points. Each atom is typed either as donor, acceptor, polar, anion, cation or hydrophobe. TGD (TGT) codes all pairs (triplets) of atoms by their types and topological distance as features. TAD and TAT use Euclidean distance between atoms. Fingerprints were compared using the Tanimoto coefficient.

(vi) The *piDAPH# group of fingerprints in MOE is based on a more elaborate pharmacophore model. Each atom is assigned a type from the eight possible combinations between 'in pi system', 'is donor' and 'is acceptor'. GpiDAPH3 codes triplets of atoms by their types and topological distances. piDAPH3 (piDAPH4) is the spatial analogue using inter-atomic distances between triplets (quadruplets). Fingerprints were compared using the Tanimoto coefficient.

(vii) CATS (Chemically Advanced Template Search),⁴¹ (viii) LIQUID (Ligand-based Quantification of Interaction Distributions),⁴² and (ix) PRPS (Pseudoreceptor Point Similarity)⁴³ are in-house implementations of the correlation vector concept.⁴⁴ For CATS, each atom is assigned one type of donor, acceptor, anion, cation and lipophilic. For all 15 pairs between these types, their occurrence in topological distances from zero to nine bonds was counted, yielding the 'raw' version of CATS. The sensitive ('sens') variant scales these values by the sum of involved atom type counts. LIQUID uses only three atom types (donor, acceptor, lipophilic) and creates a 3D pharmacophore model. Pharmacophoric points were clustered with cluster radius 2 Å and used to create feature densities modeled by trivariate Gaussians. The correlation vector was calculated between all six possible atom pairs in binned distances between one and 20 Å (or more) in steps of 1 Å. Finally, values were scaled so that the sum of the 20 bins for each pair equals 1. PRPS models a pseudoreceptor around each ligand based on known interaction directions with interaction types donor, acceptor and lipophilic. These interaction possibilities were translated into a correlation vector analogous to

LIQUID with bins from 1 to 15 Å in 1 Å steps. For all correlation vectors we used the Euclidean distance for similarity assessment.

Clustering of Virtual Screening Methods

In order to identify similarities between virtual screening methods, to assess the novelty of PhAST applied to molecular representations of different dimensionality compared to already existing virtual screening techniques, and to quantify the influence of the dimensionality of the molecular representation we performed a Ward clustering. As distance measures we used the averaged Kendall rank correlation as well as significance estimates for the superiority of one method over another. For this purpose we calculated the average percentage of virtual screens one method performs significantly better than the other in the paired permutation test at significance levels 0.05 and 0.01 for each target. We used the average of these per-target-values to assess the significance of differences in retrospective performance between methods, yielding an asymmetric distance measure. Using Eq. (2)

$$d_s(m_i, m_j) = |d_a(m_i, m_j) - d_a(m_j, m_i)| \quad (2)$$

with m a screening method, d_a the asymmetric distance between methods and d_s the symmetric distance between methods we calculated a symmetric distance matrix.

Double Dynamic Programming

Besides the canonization algorithm during generation of PhAST-sequences spatial information can be used for scoring in the alignment process. This technique called 'double dynamic programming' (DDP) has proven to be successful in the comparison of protein structures as alternative to structure superposition.^{12,45,46} We will first describe the DDP alignment algorithm for proteins. Then we will explain our modifications to apply DDP to textual representations of small molecules.

Algorithms calculating the optimal pairwise global sequence alignment use dynamic programming.^{3,24} During the alignment process matches and mismatches of residues have to be scored. Typically these scores come from score matrices like PAM⁴⁷ or BLOSUM⁴⁸ and relate to the functional similarity of residues. Using DDP these scores are calculated by a second level of dynamic programming based on structural instead of functional similarity. We will refer to these two levels as 'residue level' for the dynamic programming level equal to the normal dynamic programming and 'distance level' for the dynamic programming level calculating the scores for the residue level. The simplest approach for proteins is to consider only C_α atoms in these calculations.

When sequences $X = x_1x_2\dots x_n$ and $Y = y_1y_2\dots y_m$ are aligned and the score for aligning residues x_i and y_j on residue level have to be calculated, a position-specific distance score matrix D^{ij} with entries [Eq. (3)]

$$D_{kl}^{ij} = \frac{a}{|X_{d_{i,k}} - Y_{d_{j,l}}| + b} \quad (3)$$

is created where $X_{d_{i,k}}$ is the Euclidean distance between x_i and x_k , $Y_{d_{j,l}}$ is the Euclidean distance between y_j and y_l , b prevents division by 0 and the ratio of a to b defines the

1
2
3 maximum score where $k = i$ and $l = j$. This matrix is used as score matrix for the distance level
4 alignment. This alignment is calculated under the assumption that that x_i and y_j are structurally
5 equivalent, hence the alignment of x_i and y_j has to be part of the distance level alignment.^{12,49}
6 The alignment score of the distance level alignment is used as score for the alignment of x_i
7 and y_j on residue level. As dynamic programming has complexity $O(n^2)$, and in each step
8 dynamic programming has to be performed to calculate the score of the alignment of two
9 residues with again $O(n^2)$, DDP has complexity $O(n^4)$.

10
11
12 The parameterization of DDP poses some new problems compared to standard
13 dynamic programming. Three parameterization solutions are described in the supplemental
14 material. As they mainly differ in the determination of gap penalties used on distance level,
15 they are referred to as 'static', 'flexible' and 'dynamic', based on the particular penalty
16 choices.

17 We evaluated our implementation of DDP with the best-performing canonization
18 algorithms based on 2D and 3D information that were identified in this study. This way we
19 wanted to assess whether the combination of 2D or 3D canonization with 3D sequence
20 comparison is advantageous.
21
22

23 24 Screening Protocol 2

25
26 As DDP has a complexity of $O(n^4)$ and we implemented it using the exact but slower
27 Needleman-Wunsch algorithm, average screenings as described in screening protocol 1 would
28 take huge amounts of time. For evaluation we used a protocol proposed earlier¹ with two
29 queries per target taken from PDB⁵⁰ structures (see Table 2 for detailed query list). Each
30 query is used in a retrospective screening and the resulting ranked list is evaluated by
31 calculating their BEDROC score with $\alpha = 20$, the suggested default value for evaluation.⁷
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

RESULTS AND DISCUSSION

Canonization algorithms

We compared the retrospective performance of PhAST employing different canonization approaches applied to 2D layouts and 3D conformations of molecules in a series of virtual screens. Canonization algorithms with highest retrospective performance in PhAST are minimum volume embedding (MVE) combined with the diffusion kernel (diffusion parameter 0.4) and covalent connectivity for 2D molecular representations (averaged BEDROC = 0.40) and MVE combined with the Gaussian radial basis function kernel ($\sigma = 2^2$) and k nearest neighbors connectivity with $k = 3$ (averaged BEDROC = 0.39). These versions of PhAST will be referred to as 'PhAST 2D' and 'PhAST 3D'. Full results of all canonization algorithms are provided in the supplemental material.

The difference in screening performance between PhAST 2D and PhAST 3D is significant in more than 50% of all screenings at the 0.05 and 0.01 significance level. Compared to all methods evaluated in this study, PhAST 2D performs significantly better in 50% of all screenings in 98% (84%) of all cases at a significance level of 5% (1%). This demonstrates that MVE in combination with a diffusion kernel (diffusion parameter = 0.4) and covalent connectivity is an appropriate canonization algorithm for the generation of molecule linearizations for sequence alignment.

Impact of dimensionality on PhAST-sequences

Table 3 presents the results of comparing the application of the same canonization algorithm to 2D layouts and 3D conformations of molecules by differences in PhAST-sequences and active ranks. The average difference between the averaged Levenshtein and Damerau-Levenshtein distance is 0.66, indicating that transpositions allowed as additional edit operation in the Damerau-Levenshtein distance are not used very often to explain differences in PhAST-sequences generated from different representations of the same molecules. For all canonization algorithms utilizing a neighborhood definition of vertices, the biggest difference in PhAST-sequences is observed using k nearest neighbor neighborhoods with $k = 2$. This shows that the changes in vertex neighborhoods introduced by the additional dimension using 3D conformations are only slight displacements of the same groups of vertices. Looking at only the two nearest neighbors, these displacements may result in selecting two different nodes, but as the number of considered neighbors increases, the shared fraction of nearest neighbors increases resulting in less distant PhAST-sequences. At the same time lowest rank correlation is observed for the same group of canonization algorithms with $k = 2$ as well. Lowest distance between PhAST-sequences and highest rank correlation are observed when identical connectivity is used in form of neighborhoods defined by covalent bonds.

The rank correlation of actives varies between 0.71 and 0.36 with a mean of 0.59. Pearson's correlation coefficient⁶⁵ between Levenshtein (Damerau-Levenshtein) distance and Kendall's τ is -0.95 (-0.96). This shows that whenever a canonization algorithm generates very dissimilar PhAST-sequences from molecular representations with different dimensionalities, this results in very dissimilar rankings of actives.

Comparison 2D vs. 3D

For all 42 canonization algorithms and parameterizations applied to molecular representations in both dimensionalities, the application to 2D layouts (3D conformations) has higher

1
2
3 averaged performance in 15 (27) cases. Focusing on the number of cases where more than
4 50% of all screenings have significantly better averaged retrospective performance this is true
5 in 2 (14) cases at 0.05 and only 0 (4) cases at 0.01. This implied superiority of using three-
6 dimensional molecular representations in general does not hold necessarily for a particular
7 target. Table 4 lists the number of cases where the application to one dimensionality results in
8 significantly higher retrospective performance compared to the other in more than 50% of
9 performed screenings per target at both significance levels. By that criterion, more than 50%
10 of the compared canonization algorithms have significantly higher retrospective performance
11 when applied to 3D conformations on THR at 0.01. On this target, the application to 2D
12 layouts is superior in only 7%. On COX2 both representations excel the other in 45%. On
13 ACE in 43% of the compared cases the application on 3D conformers yields significantly
14 better results with only 5% the other way. Nearly identical percentages are obtained for both
15 dimensionalities on DHFR, FXA and PPAR γ . These results demonstrate that the usage of 3D
16 conformations seems to be advantageous only in some cases, but not in general. This is in
17 agreement with other studies evaluating 2D and 3D methods.^{63,64}

18
19 The similarity of results obtained from 2D and 3D representations can be explained by
20 the similarity of the molecular representations. A comparison of MOE 2D layouts and
21 CORINA single 3D low-energy conformations for the COBRA compounds by calculation of
22 their pairwise root mean square deviation (RMSD) revealed that they are similar with an
23 averaged RMSD of 1.9 Å (standard deviation 0.9 Å). This high similarity between
24 representations is most likely due to a large number of atoms being part of arene systems. As
25 PhAST employs the Hueckel definition of aromaticity, all atoms typed as aromatic are part of
26 such planar systems. An analysis of PPP frequencies in the COBRA library reveals that 42%
27 of all atoms are typed as aromatic. As a consequence, on average 42% of all non-hydrogen
28 atoms present in a molecule have identical conformation in 2D and 3D, and slight differences
29 occur only due do different positioning of these fragments. As a result of this limited
30 difference between representations, Euclidean and topological distances between vertices are
31 highly correlated. The Pearson correlation coefficient⁶⁵ of the Euclidean and topological
32 distances between all vertex pairs in all molecules of the COBRA library is $r = 0.94$.
33 Linearization of structures based on distances from 2D and 3D representations with the same
34 canonization algorithm results in smaller differences in the generated PhAST-sequences
35 compared to the same molecular representation being processed by different canonization
36 algorithms. This observation motivated the next part of our study, namely a comparison of
37 canonization algorithms by the similarity of their corresponding PhAST-sequences.

44 45 46 Canonization Clustering

47
48 In order to assess general differences between algorithmic concepts, to categorize algorithms
49 and to quantify the impact of the dimensionality of molecular representation in comparison to
50 algorithmic differences we clustered the canonization algorithms compared in this study by
51 different distance measures. Therefore, we first selected a representative from every group of
52 canonization algorithm that was evaluated in more than one parameterization. As selection
53 criterion we used the averaged retrospective performance obtained in the application to both
54 dimensionalities of molecular representation. To get a more complete picture, we included
55 results obtained with algorithms compared previously. As distance measures we utilized
56 Kendall's rank correlation of actives averaged over all targets as well as the averaged
57 Levenshtein and Damerau-Levenshtein distance between PhAST-sequences generated from
58 the same molecular representation with different algorithms. As in contrast to distance
59 measures high rank correlation indicates similar behavior, we used $(1 - \tau)$ as distance. We
60

1
2
3 used Ward's algorithm to create a hierarchic clustering. Distance matrices are available in the
4 supplemental material.

5
6 The dendrogram obtained using Kendall's rank correlation coefficient as distance is
7 shown in Figure 1A. The algorithms explicitly designed for the canonization of molecular
8 graphs (Jochum Gasteiger, Prabhakar, Weininger) are grouped in a sub-tree. Despite the fact
9 that the Jochum Gasteiger algorithm labels vertices in spheres around the most buried one, the
10 Prabhakar algorithm labels long paths through the graph and the Weininger algorithm groups
11 vertices with similar properties in equivalence classes, they are more dissimilar to the
12 approaches using dimensionality reduction than among themselves. That may be due to the
13 fact that all three algorithms in the end use functional and topological vertex properties for
14 prioritization instead of topological or spatial distances. Within this substructure lie both
15 variants of centroid linearization. This is an expected result as they are the spatial analogue to
16 the topological procedure performed by the Jochum Gasteiger algorithm that lies next to
17 them. Besides centroid linearization, there is no case where both variants of the same
18 algorithm applied to 2D and 3D molecular representations are grouped together. This implies
19 that the dimensionality of molecular representation has bigger impact on rank orders than
20 algorithmic differences. With MVE DK and MVE PRW both variants of MVE based on a
21 kernel performing a random walk on the graph are grouped together. They form a sub-tree
22 with covalent variants of Isomap and laplacian eigenmaps. Identical algorithm versions with
23 different connectivity are never direct neighbors in the tree, showing that results obtained with
24 chemical reasonable connectivity differ from those originating from close proximity between
25 vertices. PCA and SPE, both independent from neighborhood definitions, behave similar,
26 whereat differences between algorithms are of lesser importance compared to dimensionality
27 of molecular representation for our problem to embed structures in only one dimension. For
28 each connectivity variant and dimensionality of molecular representation, MVE variants EDK
29 and RBF are grouped together, meaning they result in similar rankings of actives. This is
30 reasonable, as both kernels depend on the same distances and coordinates for vertices in
31 Euclidean space.

32
33 These observations are substantiated by the dendrograms created with the averaged
34 Levenshtein and Damerau-Levenshtein distances (dendrograms are identical) presented in
35 Figure 1B: i) Dimensionality of molecular representation results in smaller differences as
36 using different algorithms (except for centroid linearization) as does using covalent
37 connectivity instead of k nearest neighbors, ii) MVE EDK and MVE RBF behave similar, iii)
38 dimensionality reduction algorithms used for graph canonization result in active rankings
39 quite dissimilar from those obtained with algorithms explicitly designed for this purpose.

40
41 Cluster analysis of canonization algorithms revealed equivalent behavior of MVE RBF
42 and MVE EDK. Furthermore, it emphasizes differences between 'traditional' algorithms for
43 graph canonization and our approach using dimensionality reduction algorithms and between
44 chemical reasonable connectivity and connectivity implied by spatial adjacency. These results
45 approve the novelty of our concept of using dimensionality reduction for molecule
46 linearization. Our findings indicate a difference in results obtained with molecules represented
47 in 2D and 3D. However, the observed differences again do not result in significant differences
48 of screening performance.

54 55 56 **PhAST data fusion**

57
58 PhAST 2D and PhAST 3D were combined in a data fusion approach based on compound
59 ranks, yielding PhAST DF. The averaged rank correlation of actives between results obtained
60 with PhAST 2D and PhAST 3D respectively is 0.69. Table 5 shows the averaged
retrospective performance per target, the percentages of screenings in which one method

performed significantly better than the other per target and the averaged rank correlation per target.

Judging from the averaged BEDROC scores per target, PhAST DF performs better than both original versions of PhAST only on THR and FXA. Taking into account the percentage of screenings one method outperforms the other significantly approves this result: Only on these two targets PhAST DF performs significantly better than PhAST 2D and PhAST 3D in more than 50% of all screenings at both significance levels. On other targets at least one of the other methods has at least the same percentage at 0.05. But on all targets and as consequence averaged over all targets retrospective performance did not increase enough (39% to baseline PhAST with 31% for the opposite case) to justify the computational cost of calculating 2D layouts and 3D single conformations for each molecule and conducting each screening twice, once in each dimensionality.

The averaged rank correlation between PhAST DF and PhAST 2D (PhAST 3D) is 0.85 (0.84). This shows that while PhAST DF ranks actives slightly better than PhAST 2D or PhAST 3D, it retains the order of actives relatively good with regard to these two PhAST versions and so does not introduce any novelty in chemotypes retrieved at higher ranks.

In general, the data fusion approach could not further improve the screening performance of PhAST and does not succeed in bringing any novelty to obtained screening results. As on the other hand it requires additional computational effort, we do not recommend this procedure for prospective application with the current versions of PhAST.

Multiple Conformations

We investigated the benefits of using PhAST with multiple 3D conformations (PhAST 3D MC) compared to 3D single conformations (PhAST 3D SC) in retrospective screenings with MVE RBF ($\sigma = 2^2$, kNN $k = 3$) as canonization algorithm. This variant of MVE performed best for 3D single conformations. Ten conformations were created for each molecule. BEDROC scores and results from the paired permutation test are presented in Table 6.

Averaged retrospective performance shows only a minor increase for PhAST 3D MC (0.41) compared to the SC variant (0.39) and PhAST 2D (0.40) in general. This is backed up by averaged significance estimations: At the 0.05 significance level PhAST 3D MC performs significantly better than PhAST 2D or PhAST 3D SC judged by the percentage of significantly better screenings averaged over all targets. These are 51% for PhAST 3D MC in both cases. At 0.01 the averaged percentages only slightly decrease to 48% and 47%. The opposite is true only in 35% (PhAST 2D) and 33% (PhAST 3D) at 0.05 significance level and 31% for both at 0.01. On particular targets on the other hand differences are more distinct, as the results for ACE and FXA indicate. The averaged BEDROC score for ACE increases from 0.37 for single conformers to 0.45, with that increase being significant in 79% (71%) at 0.05 (0.01) and the single conformer version being significantly better in 0% at both levels. For FXA retrospective performance is raised from 0.35 to 0.40 with significantly better results in 82% at both significance levels. On COX2 PhAST 3D MC performs worse than the SC variant, with this decrease in retrospective performance being significant in 65% at 0.05 and 61% at 0.01. So in some cases taking the maximum of all similarity values calculated in the comparison of all pairs of conformations as final similarity value is misleading and increases the number of false positives.

As an attempt to quantify the influence of different conformations on the corresponding PhAST-sequences we calculated Levenshtein- and Damerau-Levenshtein distances between PhAST-sequences of single conformations and their ten corresponding multiple conformations. PhAST-sequences generated from different conformations of the same molecule can be quite dissimilar with 4.98 (4.26) being the mean Levenshtein

(Damerou-Levenshtein) distance and standard deviation of 5.43 (5.32). As Table 7 shows this is most likely explained by the degrees of freedom each molecule has in three dimensions, measured by the averaged number of rotatable single bonds per target (descriptor b_1rotN in MOE). The Pearson correlation coefficient between these averages and the averaged Levenshtein (Damerou-Levenshtein) distance between each single conformation and its corresponding multiple conformation is 0.93 (0.87). Differences in retrospective performance between PhAST 3D SC and MC per target on the other hand do not correlate well with sequence differences (Pearson correlation coefficient of 0.31 for Levenshtein distance and 0.35 for Damerou-Levenshtein distance). This agrees with weak correlation between differences in retrospective performance and the averaged number of rotatable single bonds (0.28). So the generation process of PhAST-sequences is sensitive enough to changes in vertex placements to capture molecule flexibility and mirror it. At the same time, differences in retrospective performance between the SC and MC variant of PhAST 3D can not be explained by this behavior.

This analysis strongly suggests that the usage of multiple conformations can be beneficial on some targets. **This is in agreement with other studies evaluation screening methods with single and multiple conformations.⁶³ Still, it should be kept in mind that the increased computational cost (here: for 10 conformation per molecule, an approximately 100-fold increase) motivates 2D methods as a first choice.**

Comparison to other screening methods

We compared PhAST 2D and PhAST 3D to other virtual screening methods by their retrospective performance, significance of differences in retrospective results and active ranks measured by Kendall's rank correlation coefficient.

Table 8 gives the averaged retrospective performance per target for each method. PhAST 2D and PhAST 3D have the fourth and fifth highest averaged retrospective performance (0.40 for PhAST 2D, 0.39 for PhAST 3D) with 0.42 being the maximum performance in this comparison obtained with GpiDAPH3. The MOE fingerprints ESshape 3D and ESshape 3D HYD perform even worse than the SIMPLE vector of molecule properties (0.13 and 0.12 in contrast to 0.21). Calculating ranks for each method on each target based on retrospective performance and averaging these ranks results in placing PhAST 2D fifth and PhAST 3D eighth. Top-ranked according to this measure is LINGO that has second-highest retrospective performance (0.41). In general, PhAST succeeds in creating enrichment comparable to other established methods.

The significance between retrospective results was assessed using a paired permutation test. We calculated a symmetric distance matrix based on the average percentage of screenings one method performed significantly better than the other and used it to create a dendrogram using Ward's algorithm. The calculated symmetric distance matrix can be found in the supplemental material. The dendrograms received with significance levels 0.05 and 0.01 are shown in Figure 2A and Figure 2B. Clustering two methods together means the difference between these methods is significant in less cases than between any other pairs of methods, so these two methods show only insignificant differences in retrospective performance in most cases. The dendrogram created from significance estimation at level 0.05 mirrors the ranking order of methods by their averaged retrospective performance. From the seven top performing methods, six form a sub-tree with piDAPH4 (third-highest averaged retrospective performance) being excluded from this cluster. PhAST 2D is grouped with GpiDAPH3, the method with highest averaged retrospective performance, but ranked at fifth position by this measure. This shows that differences between methods implied by averaged performance are caused by the summation of small insignificant differences. The five worst

performing methods form a substructure of their own as well. The tree created from significance data at 0.01 is nearly identical. The only difference is clustering MACCS with the TAD / TGD / CATS sens sub-tree and placing the complete structure farther away from the best performing screening methods. PhAST 2D is grouped again with the best performing GpiDAPH3 method. So despite the fact that PhAST 2D has only fifth-highest averaged performance, it has least significant differences in retrospective screenings to the best performing method.

Motivated by these findings we analyzed in how many of the 689 screenings each method significantly outperforms each other method. The results per target are shown in Table 9. The results are nearly identical at both significance levels. At 0.01, LINGO has the highest percentage of screenings in which it significantly outperforms each other method (15%). The per-target-analysis reveals that this superiority mostly comes from screenings on FXA. PRPS dominates on DHFR and is ranked second best with 12% in total. Our method PhAST 2D is ranked third (9% at 0.01 significance level) with peak superiority on THR, whereat this significant higher performance is not as distinct compared to LINGO on FXA and PRPS on DHFR.

All significance analysis justify usage and further development of PhAST as it exhibits enrichment comparable or superior to established methods with these improvements being significant in a great number of cases.

Kendall's rank correlation coefficient calculated based on active ranks was used as distance measure to create hierarchical clusterings of the compared methods using Ward's algorithm. The calculated distance matrices can be found in the supplemental material, the dendrogram is presented in Figure 2C.

Using the inversed rank correlation the two methods based solely on structural features are grouped together (MACCS, LINGO). All our in-house implementations of the correlation vector concept (CATS, LIQUID, PRPS) are grouped with MOE pharmacophore based fingerprints (TGD, TAD, TGT, TAT) without mixing these two groups. For these methods, dimensionality of molecular representations seems to be of lesser importance than methodological differences between as in these groups variants applied to different dimensionalities are grouped together. This is also true for MOE fingerprints GpiDAPH3, piDAPH3 and piDAPH4 which use a more elaborate definitions of pharmacophoric points as TGD, TAD, TGT and TAT (eight potential pharmacophoric points instead of six). Using quadruplets (piDAPH4) instead of triangles (piDAPH3) seems to make a smaller difference in ranking actives than using only topological information (GpiDAPH3). The three worst performing methods (ESshape3D, ESshape3D HYD, SIMPLE) form their own sub-tree. All methods succeed in creating rankings of actives that are dissimilar from those. PhAST 2D and 3D are grouped together with methods solely based on structural features (MACCS, LINGO) and pharmacophoric points (piDAPH3, piDAPH4, GpiDAPH3). This high similarity between MACCS / LINGO and the piDAPH family of fingerprints is remarkable and surprising, as the general assumption is that pharmacophore methods create rankings different from structural methods. But the closeness of MACCS keys and pharmacophore methods can be explained by the fact that besides structural features (for example the presence of rings of different sizes) the substructures coded in a MACCS key represent functional groups responsible for certain interactions, *i.e.* determining the pharmacophore. LINGOs on the other hand are a flexible way of describing atom environments that as well describe functional groups responsible for interactions. The interaction information implicitly compared by these methods seems to be quite similar to the 8-point pharmacophore model of MOE and to each other. In the end, the dendrogram shows a clustering by complexity of pharmacophore models: The sparse models used in CATS, LIQUID and PRPS form one group, methods based on the five-point model of MOE another one, and the complex models with eight atom-types in MOE, nine atom-types in PhAST and even more types in MACCS and LINGO are grouped together as well.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Analysis on ranking behavior with regard to actives further justifies usage and development of PhAST as we succeeded in creating a method introducing new chemotypes at early ranks without diminishing enrichment capability.

Compared by runtime, other methods are superior to PhAST: The conversion of the complete COBRA library (8,311 compounds) to PhAST-sequences (30 minutes) is more than 250 times slower as substructure fingerprint generation, correlation vector calculation, and SMILES which need only about 20 seconds for this operation. There is no significant difference between the kernel functions employed by PhAST 2D and PhAST 3D with regard to calculation speed. On average, screening the COBRA library takes three times longer with PhAST (3 seconds) than with our in-house implementation of LINGO (1 second) or substructure fingerprints (about 1 second averaged over all fingerprints). But even fingerprints are outperformed 200-fold by screening correlation vector representations of molecules with Euclidean distance (0.005 seconds). As drastic as these differences appear, even with PhAST 10^6 compounds can be screened in approximately six minutes on a single core computer, once molecules have been converted to PhAST-sequences. The time-consuming step of molecule conversion is 'embarrassingly parallel' and can easily be distributed to several cores and computers.

Inter-method data fusion

Following the principle of data fusion in virtual screening, we combined PhAST 2D with a method based on three-dimensional molecule representations to see if we could further improve screening performance. We selected the second screening method by two criteria: The rank correlation with PhAST 2D should be low and the averaged retrospective performance high at the same time. Following these guidelines we selected PRPS ($\tau = 0.23$, averaged BEDROC 0.37) and combined both methods by ranking each compound with the minimum rank received with each method. Averaged retrospective performance per target, the percentages of significantly better screenings and the averaged rank correlation with each of the original methods per target are shown in Table 10.

The gain in screening performance is significant. Averaged retrospective performance increases from 0.40 (0.37) for PhAST 2D (PRPS) to 0.45. And even at the most rigorous significance level this improvement is significant in 50% (59%) of all cases for PhAST (PRPS). With averaged rank correlation of 0.53 to PRPS the ranking of actives is changed considerably, but relatively close to PhAST ($\tau = 0.70$). Hence we succeeded in selecting candidates for successful data fusion screenings based on easy to calculate properties.

In the comparison with other virtual screening methods PhAST exhibits comparable or superior screening capabilities and qualifies as a valuable tool in screening campaigns through the introduced novelty of chemotypes at good ranks and the distinct difference to other virtual screening method enabling successful data fusion.

Double Dynamic Programming

We adopted and parameterized the double dynamic programming (DDP) approach for calculating sequence alignments based on structural properties for the comparison of PhAST-sequences. It was applied to PhAST-sequences created using MVE DK applied to 2D layouts and MVE RBF ($\sigma = 2^2$, kNN with $k = 3$) applied to 3D single conformations in three different parameterizations. Table 11 presents the retrospective performance for each of these six combinations with that of PhAST 2D and PhAST 3D for comparison.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Regardless of the canonization algorithm, the simplest parameterization of DDP hand-built from a single example performs best in this comparison. With MVE DK for canonization before using DDP, PhAST 2D and 3D are outperformed in 1 of the 12 screenings (query S58), with MVE RBF in 2 screenings (queries DIF and S58). Enrichment better than random (BEDROC score 0.05) is achieved in most cases. We did not perform significance estimations because divergences between functional and structural versions of PhAST are distinct. Despite the fact that the faster versions of PhAST that score functional similarity and employ normal dynamic programming perform better in most screenings, these results show that the comparison of PhAST-sequences based on structural instead of functional properties is possible and succeeds in enrichment of actives, but the structural information available in the current implementation of DDP alone is not sufficient for general high enrichment. The implemented version of DDP is comparable to the first one applied to protein sequences¹² and has obvious flaws like the disability to regard differences in direction for vertices with the same distance. But since its first implementation, DDP for protein sequences was subject to numerous modifications and improvements addressing especially this deficit,^{12,45,46,49} and we are certain that these improvements will increase performance in the comparison of PhAST-sequences as well. As for DDP for protein sequences, the structural similarity score could be combined with the functional scores from our functional score matrix, or these functional scores could be calculated on the fly from pre-calculated properties as hydrogen-bond-donor and acceptor potentials, resulting in a holistic approach to molecular comparison, no longer dependent of a pharmacophoric point definition. Alignment speed could be improved by switching to the FSM algorithm as for PhAST 2D and PhAST 3D, but differences in alignments introduced this way would have to be carefully monitored. Reasons why the dynamic parameterization is inferior to fixed gap penalties remain unclear at this moment but will be the subject of further investigations.

The runtime of DDP is high compared to standard dynamic programming: On average, screening of the COBRA library takes 160 minutes on a single core of an Intel Xeon with 2.26 GHz. The query size has a strong impact on the computing time (asymptotic runtime of $O(n^4)$). With 14 symbols in the query sequence the complete COBRA set of 8,311 compounds is screened in 46 minutes. An increase in query size to 41 symbols increases screening time to 735 minutes. For the time being, this renders PhAST DDP only applicable to small, focused subsets that have been pre-selected by other methods.

We succeeded in the first time application of DDP to the calculation of structural similarity scores of small molecules. Adopting existing improvements of this approach to the comparison of textual representations of small molecules will be part of future studies.

CONCLUSIONS AND OUTLINE

In this study, we investigated the impact of using three-dimensional conformations instead of two-dimensional layouts of molecular graphs for our PhAST screening method. Canonizing CORINA single conformations with MVE combining radial basis function kernel ($\sigma = 2^2$) and k nearest neighbor connectivity ($k = 3$) has retrospective performance only slightly below the application of MVE with a diffusion kernel ($\beta = 0.4$) and covalent connectivity to molecular graphs, but requires conformer generation. **Despite of an observable difference in the retrieved actives (complementarity of PhAST 2D and PhAST 3D)**, combining these two methods through data fusion could not further improve overall screening performance. Nevertheless, using multiple conformations in PhAST significantly increases screening performance for individual targets.

We could show that, for non-linear dimensionality reduction, the applied connectivity algorithm has high impact on screening performance. A further method for the definition of neighborhoods not yet investigated is b -matching.⁶⁶ There, exactly b neighbors are assigned to each vertex. **This technique has been already been shown to be beneficial to other applications of MVE.**⁶⁷ We showed that our approach using dimensionality reduction for graph canonization yields PhAST-sequences and active rankings different from those obtained with canonization algorithms developed for molecular graphs. Further we demonstrated that PhAST ranks actives dissimilar to other methods without sacrificing screening performance, introducing novel chemotypes at earlier ranks. This proves that our approach of text-based virtual screening is worthy of further investigation and development. As PhAST MVE DK is still the best performing variant of PhAST and this difference is evidentially significant, we advise the usage of this method for prospective application.

We successfully applied double dynamic programming to the comparison of PhAST-sequences, calculating structural similarity scores. In most of our test cases calculating functional similarity resulted in better screening performance, but compared to PhAST DDP, PhAST 2D and PhAST 3D are highly optimized and their superior performance had to be expected. But these first results with PhAST DDP are very promising. There are many known improvements to DDP we are confident will improve performance and speed of PhAST DDP as well. **As the second layer of dynamic programming in DDP is only used to assess structural equivalence, a faster method for this purpose would speed up the calculation of sequence alignments based on structural similarity. One could, for example, adopt of the idea of ultrafast shape recognition.**⁶⁸ There, distance distributions are characterized by their first three moments. **The difference between these values can be used to assess the similarity of distributions, and as a consequence, structural similarity.** The combination of structural and functional similarity will lead to a holistic approach of molecule comparison based on one-dimensional textual representations.

Retrospective comparison of methods with regard to significance of performance differences revealed that in some cases a ranking of methods based solely on averaged performance is misleading. These differences may be caused by the summation of insignificant differences. Because of this discovery we highly encourage the calculation of significance estimates.

Results from the comparison of PhAST 3D applied to single and multiple conformations of molecules support our findings from a previous analysis of the robustness of canonization algorithms against topological modifications of molecular graphs,² that PhAST can not be used as additive scoring function for *de novo* design of compounds. Even small changes in the spatial arrangement of potential pharmacophoric points cause measurable changes in corresponding PhAST-sequences. As the distance measures we used to assess these differences are insensitive to the exchange of positions between equal symbol types, these changes might be more severe than observed. So small changes in graph topology as

well in spatial arrangement of potential pharmacophoric points cause changes in the PhAST-sequence as representation of a molecule, which makes them non-additive.

Graph canonization through dimensionality reduction showed at the example of PCA that the projection on one single straight axis does not yield good results. An alternative could be the projection on space-filling curves like Hilbert-,⁶⁹ Peano-,⁷⁰ or Koch-curves⁷¹. As these curves are space-filling they have the ability to de-skew the projection created by PCA. Of course, this method would again depend on the generation of vertex coordinates.

ACKNOWLEDGEMENTS

V.H. is grateful for a Ph.D. scholarship granted by Merck KGaA. **The authors thank the reviewers for helpful suggestions.**

REFERENCES

- 1 Hähnke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. *J Comput Chem* 2009, 30, 761.
- 2 Hähnke, V.; Rupp, M.; Krier, M.; Rippmann, F.; Schneider, G. *J Comput Chem* 31, 2810.
- 3 Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis*; Cambridge University Press: Cambridge, 1998.
- 4 Shaw, B.; Jebara, T. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*; Omnipress: Madison, 2007.
- 5 Kondor, R. I.; Lafferty, J. D. *Proceedings of the Nineteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, 2002.
- 6 Schneider, P.; Schneider, G. *QSAR Comb Sci* 2003, 22, 713.
- 7 Truchon, J. F.; Bayly, C. I. *J Chem Inf Model* 2007, 47, 488.
- 8 Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. *BMC Bioinformatics* 2009, 10, 225.
- 9 Kendall, M. *Biometrika* 1938, 30, 81.
- 10 Levenshtein, V. I. *Soviet Physics Doklady* 1966, 10, 707.
- 11 Damerau, F. J. *Commun ACM* 1964, 7, 171.
- 12 Taylor, W. R.; Orengo, C. A. *J Mol Biol* 1989, 208, 1.
- 13 Pearson, K. *Philos Mag* 1901, 2, 559.
- 14 Belkin, M.; Niyogi, P. *Neural Comput* 2003, 15, 1373.
- 15 Tenenbaum, J. B.; de Silva, V.; Langford, J. C. *Science* 2000, 290, 2319.
- 16 Tsuda, K.; Noble, W. S. *Bioinformatics* 2004, 20, 326.
- 17 Smola, A. J.; Kondor, R. I. *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*; Springer: Berlin and Heidelberg, 2003.
- 18 Schölkopf, B.; Smola, A. J. *Learning with Kernels*; The MIT Press: Cambridge (Massachusetts), London (England), 2002.
- 19 Cover, T.; Hart, P. *IEEE Trans Inform Theor* 1967, 13, 21.
- 20 Agrafiotis, D. K. *J Comput Chem* 2003, 24, 1215.
- 21 Jochum, C.; Gasteiger, J. *J Chem Inf Comput Sci* 1977, 17, 113.
- 22 Weininger, D.; Weininger, A.; Weininger, J. L. *J Chem Inf Comput Sci* 1989, 29, 97.
- 23 Prabhakar, Y. S.; Balasubramanian, K. *J Chem Inf Model* 2006, 46, 52.
- 24 Needleman, S. B.; Wunsch, C. D. *J Mol Biol* 1970, 48, 443.
- 25 Klenner, A.; Weisel, M.; Reisen, F.; Proschak, E.; Schneider, G. *Mol Inf* 2010, 29, 189.

- 1
2
3
4 26 Rohrer, S. G.; Baumann, K. *J Chem Inf Model* 2009, 49, 169.
5 27 Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.;
6 Kallioniemi, O. *J Chem Inf Model* 2009, 49, 2168.
7 28 Huang, N.; Shoichet, B. K.; Irwin, J. J. *J Med Chem* 2006, 49, 6789.
8 29 Good, A. C.; Oprea, T. I. *J Comput Aided Mol Des* 2008, 22, 169.
9 30 Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. *J Cheminf* 2009, 1, 14.
10 31 Bemis, G.; Murcko, M. A. *J Med Chem* 1996, 39, 2887.
11 32 Bemis, G.; Murcko, M. A. *J Med Chem* 1999, 42, 5095.
12 33 Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. *Bioinformatics* 2010, 26, 1348.
13 34 Ward, J. H. *J Amer Statistical Assoc* 1963, 58, 236.
14 35 Ginn, C. M. R.; Willett, P.; Bradshaw, J. *Perspect Drug Discov* 2000, 20, 1.
15 36 Lipton, M.; Still, W. C. *J Comput Chem* 1988, 9, 343.
16 37 Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. *J Chem Inf Comput Sci* 2002,
17 42, 1273.
18
19 38 Jaccard, P. *Bull Soc Vaudoise Sci Nat* 1901, 37, 241.
20 39 Vidal, D.; Thormann, M.; Pons, M. *J Chem Inf Model* 2005, 45, 386.
21 40 Vidal, D.; Thormann, M.; Pons, M. *J Chem Inf Model* 2006, 46, 836.
22 41 Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew Chem Int Ed* 1999, 38,
23 2894.
24 42 Tanrikulu, Y.; Nietert, M.; Scheffer, U.; Proschak, E.; Grabowski, T.; Schneider, P.;
25 Weidlich, M.; Karas, M.; Göbel, M.; Schneider, G. *ChemBioChem* 2007, 8, 1932.
26 43 Tanrikulu, Y.; Proschak, E.; Werner, T.; Geppert, T.; Todoroff, N.; Klenner, A.;
27 Kottke, T.; Sander, K.; Schneider, E.; Seifert, R.; Stark, H.; Clark, T.; Schneider, G.
28 *ChemMedChem* 2009, 4, 820.
29 44 Moreau, G.; Broto, P. *Nouveau J Chimie* 1980, 4, 757.
30 45 Taylor, W. R.; Orengo, C. A. *Protein Eng* 1989, 2, 505.
31 46 Orengo, C. A.; Taylor, W. T. *Theor Biol* 1990, 147, 517.
32 47 Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. *Atlas of Protein Sequence and*
33 *Structure*; National Biomedical Research Foundation: Washington, DC, 1978.
34 48 Henikoff, S.; Henikoff, J. G. *Proc Natl Acad Sci* 1992, 89, 10915.
35 49 Eidhammer, I.; Jonassen, I.; Taylor, W. R. *Protein Bioinformatics*; John Wiley & Sons
36 Ltd: West Sussex (England), 2004.
37 50 Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.;
38 Shindyalov, I.N.; Bourne, P.E. *Nucleic Acids Res* 2000, 28, 235.
39 51 Natesh, R.; Schwager, S.L.U.; Sturrock, E.D.; Acharya, K.R. *Nature* 2003, 421, 551.
40 52 Natesh, R.; Schwager, S.L.U.; Evans, H.R.; Sturrock, E.D.; Acharya, K.R.
41 *Biochemistry* 2004, 43, 8718.
42 53 Rowlinson, S.W.; Kiefer, J.R.; Prusakiewicz, J.J.; Pawlitz, J.L.; Kozak, K.R.;
43 Kalgutkar, A.S.; Stallings, W.C.; Kurumbail, R.G.; Marnett, L.J. *J Biol Chem* 2003,
44 278, 45763.
45 54 Kurumbail, R.G.; Stevens, A.M.; Gierse, J.K.; McDonald, J.J.; Stegeman, R.A.; Pak,
46 J.Y.; Gildehaus, D.; Miyashiro, J.M.; Penning, T.D.; Seibert, K.; Isakson, P.C.;
47 Stallings, W.C. *Nature* 1996, 384, 644.
48 55 Li, R.; Sirawaraporn, R.; Chitnumsub, P.; Sirawaraporn, W.; Wooden, J.; Athappilly,
49 F.; Turley, S.; Hol, W.G. *J Mol Biol* 2000, 295, 307.
50 56 Cody, V.; Galitsky, N.; Luft, J.R.; Pangborn, W.; Blakley, R.L.; Gangjee, A. *Anti-*
51 *Cancer Drug Des* 1998, 13, 307.
52 57 Maignan, S.; Guilloteau, J.P.; Pouzieux, S.; Choi-Sledeski, Y.M.; Becker, M.R.; Klein,
53 S.I.; Ewing, W.R.; Pauls, H.W.; Spada, A.P.; Mikol, V. *J Med Chem* 43, 3226.
54 58 Adler, M.; Davey, D.D.; Phillips, G.B.; Kim, S.H.; Jancarik, J.; Rumennik, G.; Light,
55 D.R.; Whitlow, M. *Biochemistry* 2000, 39, 12534.
56
57
58
59
60

Hähnke et al.

22

- 1
2
3
4 59 Gampe Jr., R.T.; Montana, V.G.; Lambert, M.H.; Miller, A.B.; Bledsoe, R.K.;
5 Milburn, M.V.; Kliewer, S.A.; Willson, T.M.; Xu, H.E. *Mol Cell* 2000, 5, 545.
6 60 Li, Y.; Choi, M.; Suino, K.; Kovach, A.; Daugherty, J.; Kliewer, S.A.; Xu, H.E. *Proc*
7 *Natl Acad Sci Usa* 2005, 102, 9505.
8 61 Baum, B.; Steinmetzer, T.; Heine, A.; Klebe, G.; to be published.
9 62 Lange, U.E.; Bauke, D.; Hornberger, W.; Mack, H.; Seitz, W.; Hoeffken, H.W. *Bioorg*
10 *Med Chem Lett* 2003, 19, 2029
11 63 McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.;
12 Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. *J Chem Inf Model* 2007, 47,
13 1504.
14 64 Matter, H.; Pötter, T. *J Chem Inf Comput Sci* 1999, 39, 1211.
15 65 Pearson, K. *Phil Trans R Soc* 1896, 187, 253.
16 66 Müller-Hannemann, M.; Schwartz, A. *J Exp Algor* 2000, 5.
17 67 Jebara, T.; Wang, J.; Chang, S. F. *Proceedings of the 26th International Conference on*
18 *Machine Learning*, Omnipress: Madison, 2009.
19 68 Ballester, P. J.; Richards, W. G. *Proc R Soc A* 2007, 463, 1307.
20 69 Hilbert, D. *Math Ann* 1891, 38, 459.
21 70 Peano, G. *Math Ann* 1890, 36, 157.
22 71 Koch, H. *Acta Math* 1906, 30, 145.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Legends to the figures

Figure 1 Dendrograms of canonization algorithms created using Ward's algorithm. A) distance measure: $(1-\tau)$ where τ is the averaged Kendall rank correlation of active ranks in retrospective virtual screenings, B) distance measure: averaged Levenshtein distance between PhAST-sequences generated for molecules in the complete COBRA library using different canonization algorithms. Using Damerau-Levenshtein distance results in the same dendrogram as B). Lengths of edges are solely for visualization, with no respect to actual distances. Distances matrices are available in the supplemental material.

Figure 2 Dendrograms of virtual screening methods created using Ward's algorithm. A) distance measure: percentage of significant differences in retrospective screenings at level 0.05, b) distance measure: percentage of significant differences in retrospective screenings at level 0.01, C) distance measure: $(1-\tau)$ where τ is the averaged Kendall rank correlation of active ranks in retrospective virtual screenings. Lengths of edges are solely for visualization, with no respect to actual distances. Distances matrices are available in the supplemental material.

Hähnke et al.

24

Text for Graphical Abstract

We applied non-linear dimensionality reduction to the problem of linearizing three-dimensional conformations of molecules. These linear representations are compared by functional and structural properties using our virtual screening method PhAST.

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1. Targets in the COBRA library version 6.1 used for retrospective virtual screenings. Shown are abbreviations used in this study as well as the number of active compounds. The total number of molecules in the COBRA library is 8311.

Target	Abbreviation	No. Actives
Angiotensine-converting Enzyme	ACE	34
Cyclooxygenase 2	COX2	136
Dihydrofolat-reductase	DHFR	64
Factor Xa	FXA	228
Peroxisome-proliferator activated receptor γ	PPAR γ	44
Thrombin	THR	183
Total		689

Table 2. Targets, query structure IDs in the PDB⁵⁰ and the identifier of the PDB structures they were taken from used for single screening evaluations.

	PDB Code	Ligand ID
ACE	1o86 ⁵¹	LPR
	1ufz ⁵²	MCO
COX2	1pxx ⁵³	DIF
	6cox ⁵⁴	S58
DHFR	1dg5 ⁵⁵	TOP
	1hfr ⁵⁶	MOT
FXA	1ezq ⁵⁷	RPR
	1fjs ⁵⁸	Z34
PPAR γ	1fm9 ⁵⁹	570
	1zgy ⁶⁰	BRL
THR	3eq0 ⁶¹	2TS
	1o0d ⁶²	163

Table 3. Comparison of canonization algorithms applied to two-dimensional layouts and three-dimensional conformations of molecular graphs. Differences between algorithms are quantified by the averaged Levenshtein and Damerau-Levenshtein distance between generated PhAST-sequences for the complete COBRA library and the averaged Kendall rank correlation coefficient calculated pairwise from ranks of actives between 689 virtual screenings performed with each algorithm.

	Levenshtein Distance		Damerau-Levenshtein Distance		Kendall's τ	
	\emptyset	σ	\emptyset	σ	\emptyset	σ
Centroid Linearization	9.90	6.54	9.32	6.62	0.52	0.15
Isomap kNN k = 2	11.64	7.74	11.31	7.82	0.36	0.16
Isomap kNN k = 3	8.29	5.46	7.55	5.47	0.60	0.12
Isomap kNN k = 4	8.60	5.57	8.00	5.57	0.60	0.11
Isomap kNN k = 5	8.04	5.78	7.41	5.76	0.62	0.13
Laplacian Eigenmaps kNN k = 2	10.77	7.49	10.40	7.57	0.38	0.16
Laplacian Eigenmaps kNN k = 3	7.46	5.29	6.60	5.24	0.63	0.12
Laplacian Eigenmaps kNN k = 4	7.47	5.43	6.70	5.42	0.64	0.12
Laplacian Eigenmaps kNN k = 5	6.80	5.32	6.01	5.26	0.66	0.12
PCA	7.65	6.50	7.05	6.48	0.57	0.16
MVE EDK covalent	5.57	4.88	4.69	4.75	0.70	0.11
MVE EDK kNN k = 2	7.92	6.20	7.35	6.21	0.58	0.14
MVE EDK kNN k = 3	6.55	5.51	5.83	5.47	0.65	0.12
MVE EDK kNN k = 4	6.46	5.62	5.78	5.55	0.65	0.13
MVE EDK kNN k = 5	6.74	5.98	6.13	5.92	0.62	0.14
MVE RBF $\sigma = 2^{-1}$ covalent	7.66	6.67	7.03	6.76	0.52	0.17
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	11.22	7.43	10.84	7.50	0.42	0.15
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	9.45	6.60	8.85	6.71	0.49	0.17
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	9.71	6.77	9.14	6.87	0.47	0.16
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	10.28	7.36	9.79	7.51	0.41	0.19
MVE RBF $\sigma = 2^0$ covalent	6.47	5.90	5.69	5.90	0.62	0.14
MVE RBF $\sigma = 2^0$ kNN k = 2	9.83	7.22	9.37	7.27	0.45	0.16
MVE RBF $\sigma = 2^0$ kNN k = 3	7.77	6.18	7.05	6.23	0.60	0.14
MVE RBF $\sigma = 2^0$ kNN k = 4	8.09	6.19	7.38	6.23	0.57	0.14
MVE RBF $\sigma = 2^0$ kNN k = 5	7.71	6.30	7.00	6.34	0.58	0.15
MVE RBF $\sigma = 2^1$ covalent	5.44	4.82	4.55	4.67	0.70	0.11
MVE RBF $\sigma = 2^1$ kNN k = 2	8.57	6.64	8.06	6.67	0.54	0.15
MVE RBF $\sigma = 2^1$ kNN k = 3	6.51	5.53	5.77	5.48	0.65	0.13
MVE RBF $\sigma = 2^1$ kNN k = 4	6.61	5.56	5.90	5.51	0.66	0.12
MVE RBF $\sigma = 2^1$ kNN k = 5	6.48	5.67	5.75	5.61	0.65	0.13
MVE RBF $\sigma = 2^2$ covalent	5.42	4.65	4.53	4.48	0.71	0.11
MVE RBF $\sigma = 2^2$ kNN k = 2	8.05	6.31	7.49	6.31	0.58	0.14
MVE RBF $\sigma = 2^2$ kNN k = 3	6.36	5.38	5.61	5.29	0.65	0.12
MVE RBF $\sigma = 2^2$ kNN k = 4	6.26	5.46	5.56	5.39	0.66	0.11
MVE RBF $\sigma = 2^2$ kNN k = 5	6.10	5.61	5.40	5.52	0.66	0.13
MVE RBF $\sigma = 2^3$ covalent	5.51	4.81	4.63	4.67	0.70	0.11
MVE RBF $\sigma = 2^3$ kNN k = 2	7.77	6.20	7.18	6.20	0.60	0.13
MVE RBF $\sigma = 2^3$ kNN k = 3	6.47	5.47	5.74	5.41	0.65	0.12
MVE RBF $\sigma = 2^3$ kNN k = 4	6.27	5.56	5.59	5.49	0.66	0.12
MVE RBF $\sigma = 2^3$ kNN k = 5	6.22	5.76	5.57	5.69	0.65	0.13
SPE	6.82	6.19	6.18	6.13	0.60	0.14

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

MVEPE	8.37	6.41	7.91	6.40	0.52	0.13
-------	------	------	------	------	------	------

For Peer Review

Table 4. Number of canonization algorithms that perform significantly better on two-dimensional layouts (three-dimensional conformations) compared to three-dimensional conformations (two-dimensional layouts) in more than 50% of the performed retrospective virtual screenings. The total number of algorithms compared is 42. Results are shown for significance levels 0.05 and 0.01 per target and for the averaged percentages for all targets.

	p < 0.05		p < 0.01	
	2D	3D	2D	3D
ACE	2	24	2	18
COX2	19	19	19	19
DHFR	19	15	13	12
FXA	18	15	16	11
PPAR γ	3	12	0	4
THR	5	26	3	25
Ø	1	14	0	4

Table 5. Retrospective results attained from data fusion between PhAST 2D (canonization: MVE DK, covalent connectivity) and PhAST 3D (canonization: MVE RBF $\sigma = 2^2$, kNN connectivity with $k = 3$). Shown are the averaged retrospective performance measured by averaged BEDROC scores per target and averaged over all targets, the percentage of significantly improved screenings at significance level 0.05 (0.01) and the averaged rank correlation of active ranks between each of the original methods and the method resulting from rank-based data fusion per target (PhAST DF).

		ACE	COX2	DHFR	FXA	PPAR γ	THR	\emptyset
BEDROC	PhAST DF	0.40	0.43	0.55	0.38	0.26	0.43	0.41
	PhAST 2D	0.40	0.40	0.57	0.36	0.25	0.42	0.40
	PhAST 3D	0.37	0.43	0.51	0.35	0.27	0.41	0.39
% p < 0.05 (% p < 0.01)	PhAST DF	35 (21)	66 (62)	5 (5)	65 (61)	36 (34)	55 (51)	44 (39)
	PhAST 2D	35 (24)	21 (17)	86 (78)	26 (24)	27 (18)	30 (27)	38 (31)
	PhAST DF	74 (65)	31 (27)	70 (79)	81 (79)	14 (14)	63 (61)	55 (53)
	PhAST 3D	9 (9)	46 (43)	16 (9)	12 (11)	52 (41)	27 (25)	27 (23)
Kendall's τ	PhAST 2D / PhAST DF	0.82	0.83	0.93	0.82	0.86	0.85	0.85
	PhAST 3D / PhAST DF	0.77	0.86	0.87	0.82	0.88	0.84	0.84

Table 6. Retrospective Comparison of PhAST with canonization algorithm MVE RBF $\sigma = 2^2$, kNN connectivity with $k = 3$ (PhAST 3D) applied to single (SC) and multi conformations (MC) generated for the COBRA library. Using multiple conformations, each molecule was represented by 10 conformations. Shown are the averaged retrospective performance measured by averaged BEDROC scores per target and averaged over all targets and the percentage of significantly improved screenings at significance levels 0.05 and 0.01.

		ACE	COX2	DHFR	PPAR γ	THR	FXA	\emptyset
BEDROC	PhAST 3D MC	0.45	0.41	0.51	0.28	0.42	0.40	0.41
	PhAST 2D	0.40	0.40	0.57	0.25	0.42	0.36	0.40
	PhAST 3D SC	0.37	0.43	0.51	0.27	0.41	0.35	0.39
% p < 0.05 (% p < 0.01)	PhAST 3D MC	74 (62)	49 (48)	9 (9)	57 (52)	45 (44)	71 (71)	51 (48)
	PhAST 2D	3 (0)	38 (38)	78 (75)	23 (11)	48 (45)	20 (18)	35 (31)
	PhAST 3D MC	79 (71)	19 (16)	30 (27)	43 (36)	49 (48)	82 (82)	51 (47)
	PhAST 3D SC	0 (0)	65 (63)	48 (47)	32 (20)	44 (42)	12 (11)	33 (31)

Table 7. Retrospective comparison of PhAST 3D (canonization: MVE RBF $\sigma = 2^2$, kNN connectivity with $k = 3$) applied to single conformations (SC) and 10 conformations (MC). Shown are the retrospective performance of each method measured by averaged BEDROC scores per target and averaged over all targets, the difference between retrospective performance, the averaged Levenshtein and Damerau-Levenshtein distance between PhAST-sequences generated from single conformations and the 10 conformations generated for the same molecule for each target as well as the averaged number of rotatable single bonds in molecules per target.

	ACE	COX2	DHFR	PPAR γ	THR	FXA
PhAST 3D MC	0.45	0.41	0.51	0.28	0.42	0.40
PhAST 3D SC	0.37	0.43	0.51	0.27	0.41	0.35
Δ BEDROC	0.08	-0.02	0.00	0.02	0.01	0.05
\emptyset Levenshtein distance	6.24	2.32	3.62	3.54	8.61	6.25
\emptyset Damerau-Levenshtein distance	5.43	2.04	2.66	2.85	7.72	6.25
\emptyset No. Rotatable Bonds	9.06	3.91	6.78	7.00	10.51	7.63

Table 8. Retrospective comparison of virtual screening methods. Shown are averaged BEDROC scores per target, the averaged BEDROC score averaged over all targets and the averaged rank of each method based on per target method rankings.

	BEDROC							
	ACE	COX2	DHFR	FXA	PPAR γ	THR	\emptyset	\emptyset Rank
CATS 2D raw	0.43	0.25	0.20	0.18	0.17	0.24	0.24	14.17
CATS 2D sens	0.51	0.27	0.44	0.26	0.34	0.39	0.37	6.83
Esshape 3D	0.11	0.23	0.10	0.17	0.07	0.12	0.13	17.33
Esshape 3D HYD	0.09	0.23	0.11	0.14	0.07	0.10	0.12	17.33
GpiDAPH3	0.61	0.55	0.49	0.26	0.29	0.32	0.42	5.50
LINGO	0.59	0.47	0.36	0.39	0.25	0.38	0.41	5.50
LIQUID	0.41	0.22	0.19	0.19	0.16	0.35	0.26	14.00
MACCS	0.48	0.47	0.44	0.29	0.26	0.33	0.38	7.50
PhAST 2D	0.40	0.40	0.57	0.42	0.25	0.36	0.40	6.67
PhAST 3D	0.37	0.43	0.51	0.41	0.27	0.35	0.39	7.17
piDAPH3	0.45	0.50	0.51	0.20	0.28	0.25	0.37	8.67
piDAPH4	0.49	0.56	0.55	0.28	0.28	0.29	0.41	6.17
PRPS	0.33	0.54	0.70	0.25	0.20	0.20	0.37	9.83
SIMPLE	0.24	0.30	0.20	0.21	0.12	0.20	0.21	14.50
TAD	0.59	0.38	0.28	0.28	0.33	0.34	0.37	7.50
TAT	0.56	0.37	0.35	0.33	0.34	0.38	0.39	5.83
TGD	0.60	0.39	0.28	0.26	0.32	0.37	0.37	6.83
TGT	0.52	0.32	0.35	0.24	0.25	0.36	0.34	9.67

Table 9. Number of screenings a screening method performs significantly better than every other method in the comparison. Results for significance levels 0.05 and 0.01, the latter in parentheses. Per target the number of screenings is presented, the last column shows the percentage of all 689 screenings performed for comparison.

	ACE	COX2	DHFR	FXA	PPAR γ	THR	\emptyset %
MACCS	1 (0)	10 (10)	4 (3)	35 (35)	0 (0)	0 (0)	7 (7)
CATS2D raw	0 (0)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
CATS2D sens	0 (0)	0 (0)	0 (0)	1 (1)	9 (7)	26 (25)	5 (5)
ESshape3D	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ESshape3D HYD	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)	5 (5)	1 (1)
GpiDAPH3	2 (1)	41 (39)	1 (1)	3 (2)	4 (4)	4 (4)	8 (7)
LINGO	1 (1)	2 (2)	0 (0)	94 (89)	0 (0)	12 (12)	16 (15)
LIQUID	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)	10 (10)	2 (2)
PRPS	0 (0)	22 (21)	50 (50)	4 (4)	2 (2)	5 (5)	12 (12)
PhAST_2D	0 (0)	0 (0)	1 (1)	20 (19)	1 (1)	45 (41)	10 (9)
PhAST_3D	0 (0)	1 (0)	0 (0)	12 (12)	0 (0)	28 (25)	6 (5)
SIMPLE	0 (0)	3 (3)	0 (0)	6 (4)	2 (2)	1 (1)	2 (1)
TAD	2 (2)	0 (0)	0 (0)	13 (12)	3 (2)	1 (1)	3 (2)
TAT	1 (0)	0 (0)	0 (0)	11 (11)	5 (3)	12 (11)	4 (4)
TGD	4 (1)	2 (2)	0 (0)	2 (2)	0 (0)	7 (7)	2 (2)
TGT	2 (2)	9 (9)	0 (0)	0 (0)	0 (0)	7 (7)	3 (3)
piDAPH3	0 (0)	5 (5)	1 (1)	1 (1)	1 (0)	0 (0)	1 (1)
piDAPH4	0 (0)	2 (2)	4 (4)	1 (1)	0 (0)	0 (0)	1 (1)

Table 10. Retrospective results attained from data fusion between PhAST 2D (canonization: MVE DK, covalent connectivity) and PRPS. Shown are the averaged retrospective performance measured by averaged BEDROC scores per target and averaged over all targets, the percentage of significantly improved screenings at significance level 0.05 (0.01) and the averaged rank correlation between each of the original methods and the method resulting from rank-based data fusion per target (referred to as 'Fused').

		ACE	COX2	DHFR	FXA	PPAR γ	THR	\emptyset	
BEDROC	Fused	0.44	0.53	0.72	0.36	0.27	0.37	0.45	
	PhAST 2D	0.40	0.40	0.57	0.42	0.25	0.36	0.40	
	PRPS	0.33	0.54	0.70	0.25	0.20	0.20	0.37	
% p < 0.05 (% p < 0.01)	Fused	59 (44)	79 (79)	81 (81)	39 (38)	45 (39)	17 (17)	54 (50)	
	PhAST 2D	38 (38)	18 (18)	19 (19)	57 (56)	30 (27)	78 (76)	40 (39)	
	Fused	68 (65)	40 (38)	27 (23)	92 (91)	57 (50)	91 (89)	62 (59)	
	PRPS	3 (3)	49 (46)	45 (45)	8 (8)	27 (20)	7 (7)	23 (22)	
	Kendall's τ	Fused / PhAST 2D	0.76	0.56	0.67	0.79	0.65	0.79	0.70
	Fused / PRPS	0.42	0.63	0.71	0.48	0.59	0.34	0.53	

Table 11. Retrospective results of PhAST employing double dynamic programming (DDP) as scoring function for aligned residues instead of a function-based score matrix. Query compounds are named according to their PDB identifier. Each screening was evaluated by its BEDROC score. Retrospective results of PhAST 2D (canonization: MVE DK, covalent connectivity) and PhAST 3D (canonization: MVE RBF $\sigma = 2^2$, kNN connectivity with $k = 3$) are shown for comparison. For PhAST DDP, the 2D canonization algorithm is MVE DK with covalent connectivity, the 3D canonization algorithm is MVE RBF $\sigma = 2^2$ with kNN connectivity and $k = 3$.

		PhAST DDP						PhAST 2D	PhAST 3D
Query		Canonization 2D			Canonization 3D				
		static	flexible	dynamic	static	flexible	dynamic		
ACE	LPR	0.28	0.07	0.13	0.14	0.08	0.16	0.47	0.49
	MCO	0.06	0.05	0.03	0.13	0.08	0.05	0.28	0.43
COX2	DIF	0.12	0.11	0.07	0.07	0.14	0.12	0.09	0.08
	S58	0.66	0.17	0.16	0.57	0.26	0.24	0.53	0.51
DHFR	TOP	0.19	0.09	0.16	0.17	0.17	0.12	0.70	0.69
	MOT	0.30	0.12	0.17	0.26	0.30	0.12	0.73	0.66
FXA	RPR	0.23	0.18	0.17	0.23	0.17	0.19	0.57	0.51
	Z34	0.19	0.23	0.13	0.18	0.12	0.15	0.44	0.52
PPAR γ	570	0.17	0.10	0.14	0.13	0.16	0.17	0.20	0.22
	BRL	0.30	0.25	0.32	0.28	0.31	0.29	0.52	0.51
THR	2TS	0.23	0.13	0.16	0.19	0.13	0.17	0.54	0.58
	163	0.24	0.08	0.14	0.10	0.08	0.14	0.34	0.30
	\emptyset	0.25	0.13	0.15	0.20	0.17	0.16	0.45	0.46

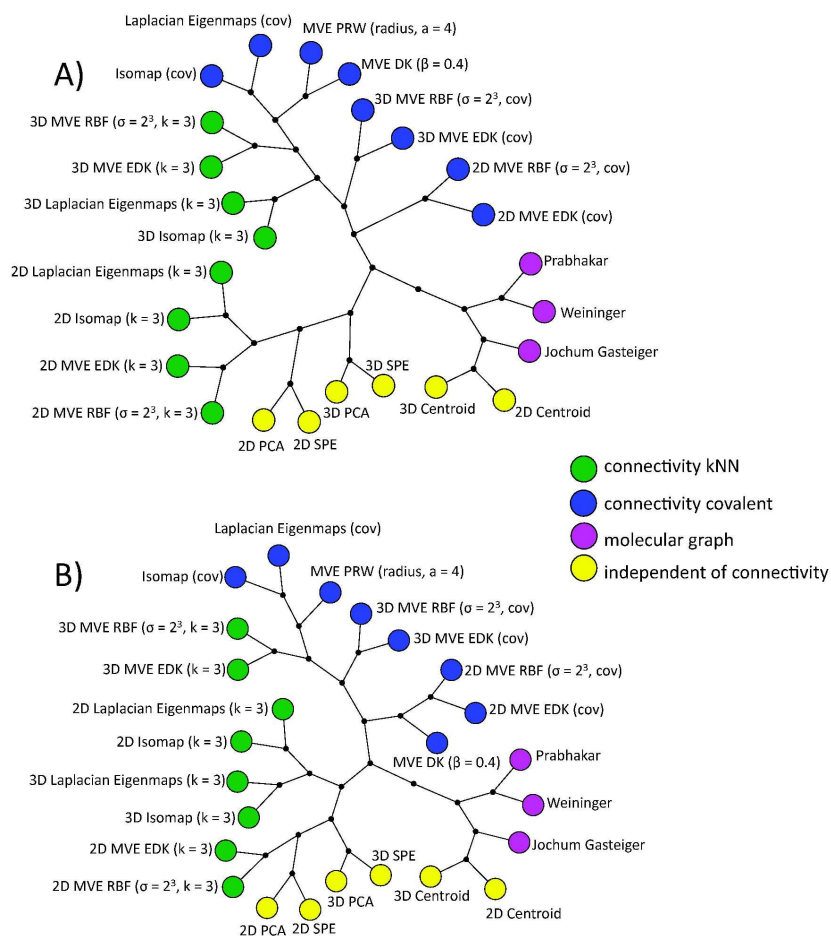


Figure 1
 154x170mm (600 x 600 DPI)

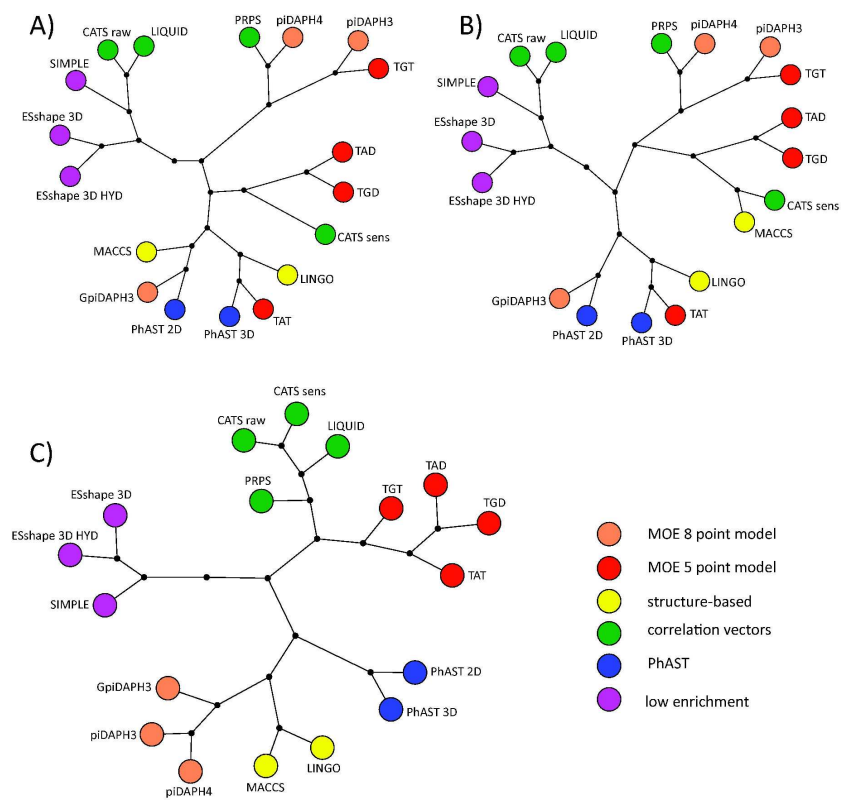


Figure 2
160x152mm (600 x 600 DPI)

Canonization Algorithms

The atom-typing step of PhAST yields a graph of potential pharmacophoric points. It has the same topology as the original molecular graph without hydrogen atoms yielding a total of n vertices. Each vertex is colored with a symbol corresponding to one out of nine potential pharmacophoric features. Edges correspond to covalent bonds. Canonization is the labeling of the vertices with the natural numbers $1, 2, 3, \dots, n$. In the following description of canonization methods, vertices are referred to as v_i with $1 \leq i \leq n$ and their coordinates in Euclidean space as x_{v_i} .

Centroid Linearization. Centroid linearization uses the distance of each vertex to the geometric centre of the complete graph as a prioritization criterion. The vertex with the lowest distance has highest priority and received the smallest canonical label. Vertices were labeled in ascending order. Centroid linearization was applied to 2D layouts and 3D conformations of molecular graphs.

Principal Component Analysis. Principal component analysis (PCA) is a linear dimensionality reduction method often used to visualize high-dimensional data.¹ We used PCA to compute one-dimensional (1D) coordinates from 2D graph layouts generated by the 2D depiction algorithm of MOE (Molecular Operating Environment, v2010.06, Chemical Computing Group, Montreal, Canada). For PCA coordinates of vertices are mean-centered. Eigenvectors of the covariance matrix calculated from position vectors x_{v_i} are used to compute new coordinates for all vertices. The dot product between the original position vector of a vertex and the eigenvector with highest eigenvalue yields a 1D coordinate. We assigned canonical labels in ascending order starting from the vertex with lowest 1D coordinate in principal component space. PCA was applied to 2D layouts and 3D conformations of molecular graphs.

Laplacian Eigenmaps. Laplacian Eigenmaps rely on a neighborhood definition between vertices generated through a connectivity algorithm (see section Connectivity Algorithms).² The canonization process starts by calculating three matrices from the neighborhood graph: (i) the weight matrix W with (Eq. 1)

$$W_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{else} \end{cases}, \quad (1)$$

(ii) the degree weight matrix D with the column sums of W as entries (Eq. 2),

$$D_{ii} = \sum_j W_{ij}, \quad (2)$$

and (iii) the positive semidefinite Laplacian matrix L (Eq. 3) with

$$L = D - W. \quad (3)$$

Then, the eigenvalues and eigenvectors of the generalized eigenvector problem (Eq. 4) are calculated.

$$Lf = \lambda Df \quad (4)$$

Eigenvectors (f_i) are sorted according to their eigenvalues (l_i) in ascending order. Eigenvector f_0 with $l_0 = 0$ is omitted. The next d eigenvectors are used for embedding. In our

case, the second eigenvector contains the coordinates for the 1D embedding. Laplacian eigenmaps were applied to 2D layouts and 3D conformations of molecular graphs.

Isomap. Isomap needs a neighborhood definition between vertices.³ The algorithm uses the neighborhood graph to estimate geodesic distances between vertices. A matrix D of shortest distances between all vertices was computed using the Floyd-Warshall algorithm.^{4,5} Using D , the matrix $\tau(D)$ is calculated (Eq. 5):

$$\tau(D) = -\frac{1}{2} * HSH, \quad (5)$$

where S is the matrix of squared distances (Eq. 6)

$$S_{ij} = D_{ij}^2, \quad (6)$$

and H the centering matrix (Eq. 7)

$$H_{ij} = \delta_{ij} - \frac{1}{n} \quad (7)$$

with δ_{ij} the Kronecker delta and n the number of vertices. The eigenvectors and eigenvalues of $\tau(D)$ are computed. To embed in d dimensions, the first d eigenvectors sorted according to their eigenvalues in decreasing order are used. If λ_p is the p^{th} eigenvalue of $\tau(D)$ and f_p^i is the i^{th} component of the p^{th} eigenvector, then the p^{th} component of the d -dimensional coordinate vector of a vertex is equal to $f_p^i \sqrt{\lambda_p}$. Isomap was applied to 2D layouts and 3D conformations of molecular graphs.

Minimum Volume Embedding. Minimum volume embedding (MVE) is a non-linear dimensionality reduction algorithm.⁶ It minimizes information loss during embedding in d dimensions. MVE requires two representations of the set of vertices: The affinity matrix A calculated using a kernel function, and a symmetric binary connectivity matrix C constructed through the application of a connectivity algorithm to A . Dimensionality reduction is achieved by an iterative process based on semidefinite programming (SDP). A third matrix K is set equal to A and the following procedure is repeated until convergence: (i) calculate the eigenvectors f_i and eigenvalues λ_i of K and sort the f_i descending to their corresponding λ_i . (ii) calculate the matrix B using Eq. (8),

$$B = -\sum_{i=1}^d f_i f_i^T + \sum_{i=d+1}^N f_i f_i^T \quad (8)$$

(iii) use SDP to solve Eq. (9)

$$K = \underset{K \in \mathcal{K}}{\operatorname{argmin}} \operatorname{tr}(KB) \quad (9)$$

under constraints \mathcal{K} defined by Shaw and Jebara,⁶ tr denotes the matrix trace (sum of the diagonal elements). After convergence, kernel PCA⁷ is performed with K to get the d eigenvectors used for embedding.

We will now describe the kernel functions and connectivity algorithms we chose for our study.

1. Diffusion Kernel

The currently best-performing version of PhAST uses a diffusion kernel.⁸ It is solely based on topological information, thus independent from spatial vertex coordinates. For each pair of vertices (v_i, v_j) the diffusion kernel calculates the probability of a random walk starting in v_i ending in v_j after an infinite number of steps, with only a low probability of leaving the current vertex in each step. The diffusion kernel matrix is calculated according to Eq. (10)⁹

$$K^{diffusion} = e^{(-\beta L)} \quad (10)$$

with β the diffusion parameter and L the Laplacian matrix (Eq. 3). The best performing version of PhAST so far uses $\beta = 0.4$. The combination of MVE and the diffusion kernel will be referred to as ‘MVE DK’. MVE DK was applied only to 2D layouts of molecular graphs.

2. *P-Step Random Walk Kernel*

A kernel function that only depends on graph topology is the p -step random walk kernel.¹⁰ It is calculated using the normalized Laplacian matrix \tilde{L} with entries (Eq. 11)

$$\tilde{L}_{ij} = \begin{cases} 1 & \text{if } i = j \\ \frac{1}{\sqrt{\deg(v_i) \times \deg(v_j)}} & \text{if } v_i \text{ adjacent } v_j \\ 0 & \text{else} \end{cases} \quad (11)$$

where deg is the degree of a vertex. The kernel matrix is calculated as Eq. (12)

$$K^{pstep} = (aI - \tilde{L})^p \quad (12)$$

with $a \geq 2$.¹⁰ We investigate p -step random walk kernels with a assuming values of 2, 4, 6, 8 and 10. For p we chose two values that automatically adjust to the current graph: We measure the distance between two vertices as the number of bonds along the shortest path. The eccentricity of a vertex is its distance to the farthest vertex in the graph.¹¹ Our first choice for p was the smallest eccentricity of a vertex in the graph, referred to as the ‘graph radius’.¹¹ The second choice was the largest eccentricity of a vertex in the graph, referred to as the graph ‘diameter’.¹¹ Both were determined using the Floyd Warshall algorithm.^{4,5} The combination of MVE and the p -step random walk kernel will be referred to as ‘MVE PRW’. MVE PRW was applied only to 2D layouts of molecular graphs.

3. *Inner Products from Euclidean Coordinates*

As a first kernel function depending on Euclidean vertex coordinates we employed a method that calculates inner products from Euclidean coordinates of vertices referred to as ‘Euclidean distance kernel’ (Eq. 13).

$$K_{ij}^{euclidean} = \frac{1}{2} \times \left(\|x_{v_i}\|^2 + \|x_{v_j}\|^2 - \|x_{v_i} - x_{v_j}\|^2 \right) \quad (13)$$

The combination of MVE and the Euclidean distance kernel will be referred to as ‘MVE EDK’. MVE EDK was applied to 2D layouts and 3D conformations of molecular graphs.

4. RBF Kernel

As second kernel function depending on Euclidean coordinates is the Gaussian radial basis function (RBF) kernel (Eq. 14).¹²

$$K_{if}^{rbf} = \exp\left(-\frac{\|x_{v_i} - x_{v_j}\|^2}{2\sigma^2}\right), \quad (14)$$

with σ the standard deviation. We parameterized the RBF kernel in a grid search with σ chosen according to Eq. (15)

$$\sigma = 2^k, \quad -1 \leq k \leq 3 \quad (15)$$

where k was incremented in steps of 1. For $k < -1$ the kernel matrix was mostly filled with zeros, and the eigenvalue problem is degenerated. For $k > 3$ all matrix entries approached 1 for small molecules. The combination of MVE and the Gaussian radial basis function kernel will be referred to as ‘MVE RBF’. MVE RBF was applied to 2D layouts and 3D conformations of molecular graphs.

Proximity Embedding. Proximity embedding (PE) utilizes pairwise distances between points to embed a dataset in arbitrary dimensions conserving the given distances. Given a pair of points (x_{v_i}, x_{v_j}) the algorithm calculates their Euclidean distance $d(t x_{v_i}, t x_{v_j})$ in the target dimension t (in our case: 1). This distance is compared to the corresponding distance $d(s x_{v_i}, s x_{v_j})$ in the starting dimension s . If $d(t x_{v_i}, t x_{v_j}) \neq d(s x_{v_i}, s x_{v_j})$, the position of v_i and v_j in the target dimension is updated using Eq. (16) and Eq. (17), respectively.

$$t x_{v_i}^{new} = t x_{v_i}^{old} + \frac{\lambda(d(s x_{v_i}, s x_{v_j}) - d(t x_{v_i}^{old}, t x_{v_j}^{old}))}{d(t x_{v_i}^{old}, t x_{v_j}^{old})} * (t x_{v_i}^{old} - t x_{v_j}^{old}) \quad (16)$$

$$t x_{v_j}^{new} = t x_{v_j}^{old} + \frac{\lambda(d(s x_{v_i}, s x_{v_j}) - d(t x_{v_i}^{old}, t x_{v_j}^{old}))}{d(t x_{v_i}^{old}, t x_{v_j}^{old})} * (t x_{v_i}^{old} - t x_{v_j}^{old}) \quad (17)$$

where l is a linear learning rate that controls the update step-size of the algorithm with $l = 1$ for the first iteration and $l = 0$ for the last iteration. We used PE for the embedding from two and three dimensions in one dimension with 5000n iterations. After termination the order of vertices in the embedding dimension starting from the lowest coordinate defines the canonical order. We used PE in two different variants: In *stochastic proximity embedding* (SPE)¹³ pairs of vertices are chosen randomly. We also evaluated a systematic algorithmic variation in which all pairs of vertices were chosen 5000 times. To ensure the invariant ordering of vertices independent from molecule input, the molecular graph was canonically labeled with MVE DK before the application of PE. This version is referred to as ‘MVEPE’. SPE and MVEPE were applied to 2D layouts and 3D conformations of molecular graphs.

Connectivity Algorithms

Laplacian eigenmaps, Isomap and MVE in variants EDK and RBF depend on neighborhood definitions for each vertex. In this study, we compared results obtained with two different connectivity algorithms: (i) covalent bonds, and (ii) k nearest neighbors.

Covalent Bonds. The graph of potential pharmacophoric points has the same topology as the original molecular graph with suppressed hydrogen atoms. Edges represent covalent bonds. Using only this information the binary connectivity matrix C corresponds to the adjacency matrix of the graph.

k Nearest Neighbors. Using symmetric k nearest neighbors (kNN)¹⁴ the binary connectivity matrix C was initialized with all entries 0. For each vertex v_i with $1 \leq i \leq n$, C_{ij} and C_{ji} were set to 1 if the distance calculated from A_{ij} is one of the top k values for $1 \leq j \leq n$. For MVE, the k nearest neighbor algorithm was applied to distances the affinity matrix, not the original space. Due to the symmetry condition vertices can end up having more than k neighbors. For k we used 2, 3, 4 and 5.

MVE DK and MVE PRW were used only with covalent connectivity. This was due to the fact that in MVE the connectivity algorithm is applied to the kernel matrix, which in these cases is calculated using already defined neighborhoods.

References

- 1 Pearson, K. *Philos Mag* 1901, 2, 559.
- 2 Belkin, M.; Niyogi, P. *Neural Comput* 2003, 15, 1373.
- 3 Tenenbaum, J. B.; de Silva, V.; Langford, J. C. *Science* 2000, 290, 2319.
- 4 Floyd, R. W. *Comm ACM* 1962, 5, 345.
- 5 Warshall, S. *J ACM* 1962, 9, 11.
- 6 Shaw, B.; Jebara, T. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*; Omnipress: Madison, 2007.
- 7 Schölkopf, B.; Smola, A.; Müller, K. *Neural Comput* 1998, 10, 1299.
- 8 Kondor, R. I.; Lafferty, J. D. *Proceedings of the Nineteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, 2002.
- 9 Tsuda, K.; Noble, W. S. *Bioinformatics* 2004, 20, 326.
- 10 Smola, A. J.; Kondor, R. I. *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*; Springer: Berlin and Heidelberg, 2003.
- 11 Petitjean, M. *J Chem Inf Comput Sci* 1992, 32, 331-337.
- 12 Schölkopf, B.; Smola, A. J. *Learning with Kernels*; The MIT Press: Cambridge (Massachusetts), London (England), 2002.
- 13 Agrafiotis, D. K. *J Comput Chem* 2003, 24, 1215.
- 14 Cover, T.; Hart, P. *IEEE Trans Inform Theor* 1967, 13, 21.

Parameterization of Double Dynamic Programming

In order to use DDP as sequence comparison step in PhAST we switched back to the slower Needleman-Wunsch⁻¹ instead of the FSM² algorithm because it calculates the exact optimal global pairwise sequence alignment. Using DDP, scores for the alignment of particular symbols are calculated by a second level of dynamic programming based on structural instead of functional similarity. We will refer to these two levels as ‘residue level’ for the dynamic programming level equal to the normal dynamic programming and ‘distance level’ for the dynamic programming level calculating the scores for the residue level. The simplest approach for proteins is to consider only C_α atoms in these calculations.

When sequences $X = x_1x_2\dots x_n$ and $Y = y_1y_2\dots y_m$ are aligned and the score for aligning residues x_i and y_j on residue level have to be calculated, a position-specific distance score matrix D^{ij} with entries [Eq. (1)]

$$D_{kl}^{ij} = \frac{a}{|X_{d_{i,k}} - Y_{d_{j,l}}| + b} \quad (1)$$

Because of experiences with DDP for protein comparison³ we implemented a slightly modified approach and modified Eq. (1) yielding Eq. (2)

$$D_{k,l}^{i,j} = \begin{cases} a^2b^2 & \text{if } k = i, l = j \\ \frac{a}{|X_{d_{i,k}} - Y_{d_{j,l}}|} \times \log(|k - i| + |l - j| + 1) & \text{else} \end{cases} \quad (2)$$

To ensure that the alignment of x_i and y_j is included in the distance level alignment the score of this event was set to a^2b^2 . The first part of the term in the ‘else’ case is the actual structural component of the score, expressing the structural similarity of x_k and y_l under the assumption that x_i and y_j are structurally equivalent. The second part is a sequence distance component that damps the contribution from near neighbors in the sequence. The idea is that it is likely that the structural similarity is high if x_k (y_l) is close to x_i (y_j).³

With two levels of dynamic programming there are two sets of gap penalties accompanied by the new parameters a and b . In addition, all scores calculated during DDP are positive, because of the absolute value used in Eq. (20) and Eq. (21), respectively. As a result, gap penalties have to be low positive scores. Compared to sequence alignment with dynamic programming the scores calculated with DDP turned out to be huge (not shown). We addressed all these problems as follows:

To have an origin for parameterization we set $a = 100$ and $b = 2$ in the calculation of distance level score matrices. After calculation of distance level alignment scores, a^2b^2 was subtracted as this value was only used to attract the alignment algorithm to the alignment of x_i and y_j in the calculation of the distance level alignment, but it dominates the calculated alignment score.

For distance level gap penalties we implemented three different solutions: i) static gap penalties that are used in every distance level alignment for every sequence comparison, ii) flexible gap penalties that are determined for each sequence pair but used in every distance level alignment, and iii) dynamic gap penalties that are determined for each distance level alignment. At the same time we determined a ‘correction value’ that is subtracted from each distance level alignment score, translating these scores partially in the

negative spectrum to make the scoring system comparable to functional score matrices. Gap penalties and correction values were calculated under some constraints deduced from the original PhAST (functional) score matrix⁴ as template: 24% (64%) of the scores were below the gap open (gap extension) penalty of -5 (-1). Again, 64% of all scores were below 0, so we used this ratio as guideline for the correction value.

Static Gap Penalties. We determined a gap open and a gap extension penalty using two molecules as example. We chose two PPAR γ agonists from the COBRA library (Figure 1).⁵ They exhibit partial structural similarity and differ only slightly in size. We calculated all distance level score matrices and merged them in one distribution. With gap open (gap extension) penalty 30 (53) the constraints of 24% (64%) of the scores in this distribution being smaller than the corresponding penalty were fulfilled. The constraint for the correction value was satisfied with a value of $m^2n^2*2.11$. With this model, gap penalties are static, but the correction value depends on sequence lengths. We chose a non-static correction value instead of one suitable for our toy example because preliminary results indicated higher retrospective performance with this choice.

Flexible Gap Penalties. To make the scoring system more flexible, we implemented a second penalty model. It is similar to the construction of the static model but applied to each sequence pair before its actual comparison. Before DDP was applied to a sequence pair, we constructed all distance level score matrices and merged all distance scores into a single distribution. Gap penalties were chosen to fulfill constraints. The correction value was optimized in a binary search approach: starting from 10,000 the fraction of scores below 0 obtained using this correction value was determined. If the correction value was too high (low), it was scaled with 0.5 (1.5). Starting from 10,000 proved to be sufficient for all our test cases.

Dynamic Gap Penalties. For each distance score matrix the gap penalties were set to fulfill constraints. So each of the mn distance score matrices created in the comparison of two sequences used a different set of gap penalties. The correction value was calculated as described for flexible gap penalties.

To further reduce distance level alignment scores after subtracting a^2b^2 and the correction value, each score $s_{i,j}$ was transformed according to Eq. (22).

$$s_{i,j}^* = \text{sign}(s_{i,j}) * \log_{\Phi}(|s_{i,j}|), \quad (22)$$

where sign is the signum function and Φ is the golden ratio (1.618). Applying the logarithm to scores was already reported for DDP for protein sequence comparison⁴⁷ We used the golden ratio because we needed a small base, and preliminary results were promising. Resulting scores spanned a range comparable to those calculated by PhAST using the original functional score matrix. Because of that fact we used gap open penalty -5 and gap extension penalty -1, and as final similarity measure between sequences the residue level alignment score normalized to alignment length.

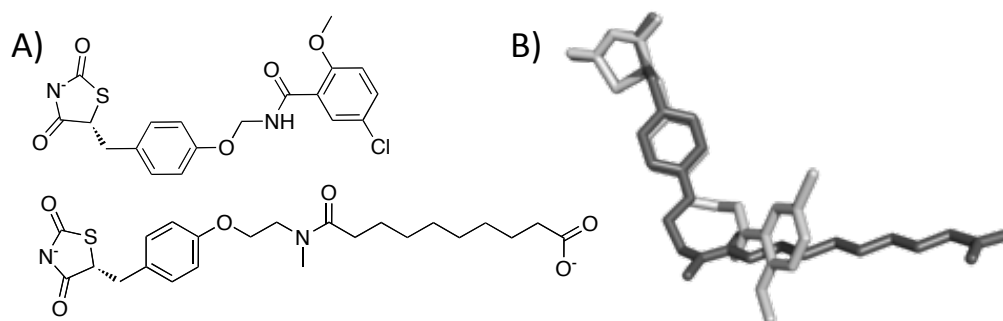


Figure 1. PPAR γ agonists used for parameterization of double dynamic programming. A) 2D depiction of the molecular graphs revealing parts of structural identity, depictions generated with MOE (Molecular Operating Environment, v2010.06, Chemical Computing Group Inc., Montreal, Canada), B) MOE rigid body alignment of CORINA (v3.46, Molecular Networks GmbH, Erlangen, Germany) 3D conformations. Molecules were taken from the COBRA collection of drugs and lead compounds.

References

- 1 Needleman, S. B.; Wunsch, C. D. *J Mol Biol* 1970, 48, 443.
- 2 Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis*; Cambridge University Press: Cambridge, 1998.
- 3 Eidhammer, I.; Jonassen, I.; Taylor, W. R. *Protein Bioinformatics*; John Wiley & Sons Ltd: West Sussex (England), 2004.
- 4 Hähnke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. *J Comput Chem* 2009, 30, 761.
- 5 Schneider, P.; Schneider G. *QSAR Comb Sci* 2003, 22, 713.

Discussion Canonization Algorithms

PhAST employing minimum volume embedding utilizing a diffusion kernel in combination with covalent connectivity exhibited highest retrospective screening performance so far. Due to this fact, this particular version of PhAST will be referred to as ‘baseline PhAST’.

We compared the retrospective performance of PhAST employing different canonization approaches applied to 2D layouts and 3D conformations of molecules in a series of virtual screenings. Table 1 presents the results evaluated using the BEDROC metric ($\alpha = 20$). Corresponding p -values attained from the paired permutation test for significance assessment are given in Table 2 (significance level 0.05) and Table 3 (significance level 0.01). Table 4 to Table 7 give p -values for baseline PhAST compared to the other canonization algorithms listed in Table 3. In the following comparison, the application of a canonization algorithm to 2D layouts is referred to as ‘2D version’, the application to 3D conformations as ‘3D version’.

With an averaged BEDROC score of 0.3 centroid linearization performs significantly worse than baseline PhAST. This is true for both dimensionalities of molecular representations. Results obtained from the application to 2D layouts and 3D conformations are not identical but not significantly different either: At significance level 0.05 retrospective performance with 2D layouts (3D conformations) performs significantly better in 43% (46%) of all screenings. At 0.01 these percentages decrease to 39% and 43%, respectively. Centroid linearization is clearly outperformed by baseline PhAST in 77% and 75% of the performed screenings at significance levels of 0.05 and 0.01, respectively.

For Isomap the retrospective performance strongly depends on the value of k used in determination of neighborhood relations. Worst performance results from $k = 2$ with an averaged BEDROC of 0.28 for both dimensionalities. Best performance is observed with $k = 3$, with an averaged BEDROC of 0.34 for 2D and 0.35 for 3D. Only with $k = 2$ the 2D version performs significantly better than 3D, admittedly in less than 50% of all screenings. For all other k the 3D version significantly outperforms the 2D version at both significance levels, at 0.05 in more than 50% of all screenings. The same is true for laplacian eigenmaps: $k = 2$ yields lowest BEDROC scores for both dimensionalities, with increasing performance for increasing k with a slight drop again at $k = 4$ for the 2D version. For both significance levels the 3D version outperforms the 2D version significantly in more cases than *vice versa* for all k except $k = 2$. Baseline PhAST outperforms all versions of Isomap and laplacian eigenmaps in more than 50% of all screenings in both dimensionalities at both significance levels.

The mediocre performance achieved using PCA for canonization we already knew to be true for 2D layouts (0.35) was affirmed for the application to 3D conformations as well (0.34).

With MVE EDK for canonization the best retrospective performance is achieved using covalent connectivity (0.38 for 2D, 0.39 for 3D) with significant differences in 54% (45%) of all screenings. As for Isomap and Laplacian Eigenmaps, starting from $k = 2$ in k nearest neighbors as connectivity algorithm, retrospective performance increases with increasing k . In all parameterizations tested, the 3D version has higher averaged performance. For k nearest neighbors, highest performance is achieved for both dimensionalities with $k = 3$. At both significance levels baseline PhAST performs significantly better than any variant of MVE EDK in more than 50% of all screenings.

For MVE RBF we evaluated five s in combination with five connectivity variants. The question is: Which one of these variables has greater influence on screening performance with MVE RBF as canonization algorithm? To address this problem we analyzed retrospective results for both dimensionalities separately. We calculated mean and standard deviations of averaged retrospective performance for each σ with varying connectivity and each

connectivity with varying s . From these values we calculated the corresponding coefficient of variation (standard deviation / mean) (CV). For fixed σ (connectivity) with varying connectivity (s) the mean CV is 0.038 (0.049) for the 2D version. For 3D the corresponding values are 0.045 and 0.065. These results identify s to have a (slightly) bigger impact on retrospective performance, because divergence in retrospective results is higher if s varies. The best performing parameter combination for the application to 2D layouts is $s = 2^2$ and covalent connectivity, for 3D conformations $s = 2^2$ k nearest neighbors with $k = 3$. The corresponding averaged BEDROC scores are 0.37 for the 2D version and 0.39 for 3D. For the 25 variants of MVE RBF tested, the 2D (3D) version performs better than the 3D (2D) version in 10 (15) cases. If we use the number of times a dimensionality outperforms the other in over 50% of the performed screenings at the chosen significance level as superiority criterion instead of just the averaged performance, these numbers further decrease for 2D (3D) to 1 (6) at 0.05 and only 0 (1) at 0.01. These results indicate that the assumed superiority in averaged retrospective performance for the usage of 3D single conformations is caused by the summation of insignificant differences. Compared to all variants of MVE RBF, baseline PhAST performs significantly better in at least 57% (49%) of all screenings in 2D (3D) at 0.05 and 53% (45%) at 0.01 with the opposite being true in only 29 (33%) and 26% (33%) of all cases. Baseline PhAST performs superior to PhAST with MVE RBF, and no advantage of using a certain dimensionality of molecular representation could be established.

Both variants of proximity embedding have lower averaged retrospective performance than baseline PhAST and are outperformed in at least 70% (65%) of all screenings at 0.05 (0.01) significance level. For both methods of choosing vertex pairs, the application on 2D layouts performs slightly better, but always in fewer than 50% of all screenings.

No parameterization of MVE PRW performs better in PhAST than baseline PhAST. The best performing variant ($p = \text{radius}$, $a = 4$) is outperformed in 51% (47) of all screenings at 0.05 (0.01), but performs better than baseline PhAST in only 25% (27%). So despite the fact that both kernels are based on random walks, this new variant of MVE using only topological information is no improvement.

Table 1. Comparison of different canonization algorithms by their averaged retrospective performance. Algorithms were applied to molecular representations in two and three dimensions. Screenings were evaluated by their BEDROC score. As MVE DK and MVE PRW only use topological information, the application to 3D conformations yielded identical results and is not shown.

	Ø BEDROC			Ø BEDROC	
	2D	3D		2D	3D
Centroid Linearization	0.30	0.30	MVE RBF $\sigma = 2^1$ kNN k = 3	0.37	0.39
Isomap kNN k = 2	0.28	0.28	MVE RBF $\sigma = 2^1$ kNN k = 4	0.36	0.37
Isomap kNN k = 3	0.34	0.35	MVE RBF $\sigma = 2^1$ kNN k = 5	0.36	0.37
Isomap kNN k = 4	0.32	0.34	MVE RBF $\sigma = 2^2$ covalent	0.37	0.38
Isomap kNN k = 5	0.33	0.35	MVE RBF $\sigma = 2^2$ kNN k = 2	0.35	0.35
Laplacian Eigenmaps kNN k = 2	0.29	0.27	MVE RBF $\sigma = 2^2$ kNN k = 3	0.37	0.39
Laplacian Eigenmaps kNN k = 3	0.35	0.36	MVE RBF $\sigma = 2^2$ kNN k = 4	0.37	0.37
Laplacian Eigenmaps kNN k = 4	0.34	0.36	MVE RBF $\sigma = 2^2$ kNN k = 5	0.37	0.36
Laplacian Eigenmaps kNN k = 5	0.34	0.37	MVE RBF $\sigma = 2^3$ covalent	0.37	0.39
PCA	0.35	0.34	MVE RBF $\sigma = 2^3$ kNN k = 2	0.35	0.36
MVE EDK covalent	0.38	0.39	MVE RBF $\sigma = 2^3$ kNN k = 3	0.37	0.39
MVE EDK kNN k = 2	0.35	0.35	MVE RBF $\sigma = 2^3$ kNN k = 4	0.37	0.37
MVE EDK kNN k = 3	0.37	0.39	MVE RBF $\sigma = 2^3$ kNN k = 5	0.37	0.37
MVE EDK kNN k = 4	0.37	0.37	SPE	0.36	0.35
MVE EDK kNN k = 5	0.36	0.36	MVEPE	0.32	0.32
MVE RBF $\sigma = 2^{-1}$ covalent	0.36	0.34	MVE DK (b = 0.4)	0.40	
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	0.32	0.31	MVE PRW diameter a = 2	0.37	
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	0.34	0.33	MVE PRW diameter a = 4	0.38	
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	0.32	0.32	MVE PRW diameter a = 6	0.39	
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	0.30	0.29	MVE PRW diameter a = 8	0.39	
MVE RBF $\sigma = 2^0$ covalent	0.36	0.37	MVE PRW diameter a = 10	0.39	
MVE RBF $\sigma = 2^0$ kNN k = 2	0.33	0.33	MVE PRW radius a = 2	0.38	
MVE RBF $\sigma = 2^0$ kNN k = 3	0.36	0.38	MVE PRW radius a = 4	0.39	
MVE RBF $\sigma = 2^0$ kNN k = 4	0.34	0.34	MVE PRW radius a = 6	0.39	
MVE RBF $\sigma = 2^0$ kNN k = 5	0.34	0.35	MVE PRW radius a = 8	0.39	
MVE RBF $\sigma = 2^1$ covalent	0.37	0.38	MVE PRW radius a = 10	0.38	
MVE RBF $\sigma = 2^1$ kNN k = 2	0.35	0.35			

Table 2. Comparison of different canonization algorithms by estimated significance of differences in retrospective performance. Algorithms were applied to molecular representations in two and three dimensions. For each algorithm applied to 2D and 3D representations of molecules the percentage of screenings is listed where this combination of algorithm and dimensionality significantly outperforms the same algorithm on the other dimensionality. Percentages may not add up to 100 because in some cases differences are not significant. Significance level: 0.05.

	∅		ACE		COX2		DHFR		FXA		PPAR γ		THR	
	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
Centroid Linearization	43	46	68	18	20	76	28	61	79	18	41	32	21	69
Isomap kNN k = 2	45	40	9	74	44	41	56	31	66	25	39	34	57	37
Isomap kNN k = 3	31	52	12	62	24	74	61	19	29	60	30	39	27	62
Isomap kNN k = 4	30	55	24	56	24	70	50	41	33	60	16	50	32	56
Isomap kNN k = 5	28	55	12	56	28	65	25	53	48	42	25	50	28	65
Laplacian Eigenmaps kNN k = 2	51	37	12	79	64	32	61	30	81	16	30	36	59	31
Laplacian Eigenmaps kNN k = 3	33	53	15	71	21	70	53	34	57	32	27	48	26	65
Laplacian Eigenmaps kNN k = 4	25	57	15	65	26	64	28	50	42	50	20	39	19	73
Laplacian Eigenmaps kNN k = 5	25	57	18	68	26	62	27	56	38	46	16	45	23	66
PCA	46	38	26	47	60	30	50	39	76	21	20	39	40	52
MVE EDK covalent	29	54	12	62	32	60	25	67	33	55	18	43	54	38
MVE EDK kNN k = 2	43	42	18	71	60	32	59	22	44	50	43	23	33	55
MVE EDK kNN k = 3	35	50	15	59	62	28	38	41	52	40	18	66	26	67
MVE EDK kNN k = 4	38	44	21	56	68	19	34	33	48	47	23	50	33	60
MVE EDK kNN k = 5	42	42	24	47	77	17	48	27	53	44	18	57	31	59
MVE RBF $\sigma = 2^{-1}$ covalent	50	36	71	9	26	60	50	31	78	19	43	34	29	62
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	44	45	26	56	82	15	58	36	57	33	14	64	31	64
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	45	43	32	53	66	28	61	30	75	20	14	59	23	68
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	41	47	44	41	18	77	25	64	79	20	32	39	49	40
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	43	41	32	15	26	67	55	38	57	37	16	70	72	21
MVE RBF $\sigma = 2^0$ covalent	38	44	32	26	32	60	22	70	42	45	48	25	50	37
MVE RBF $\sigma = 2^0$ kNN k = 2	44	44	12	65	84	11	23	61	46	52	55	20	43	54
MVE RBF $\sigma = 2^0$ kNN k = 3	33	51	18	68	63	27	19	73	38	48	20	41	38	49
MVE RBF $\sigma = 2^0$ kNN k = 4	38	46	24	53	30	61	30	53	47	44	43	27	54	38
MVE RBF $\sigma = 2^0$ kNN k = 5	37	44	32	32	29	66	38	38	57	32	23	50	42	48
MVE RBF $\sigma = 2^1$ covalent	29	51	18	44	30	58	20	70	26	59	39	32	43	44
MVE RBF $\sigma = 2^1$ kNN k = 2	36	50	15	74	38	52	58	30	40	57	18	45	46	44
MVE RBF $\sigma = 2^1$ kNN k = 3	32	49	12	62	63	24	19	69	39	51	30	30	31	61
MVE RBF $\sigma = 2^1$ kNN k = 4	35	43	29	41	56	34	50	30	29	61	11	43	34	52
MVE RBF $\sigma = 2^1$ kNN k = 5	39	44	18	50	64	26	45	42	50	41	20	50	35	56
MVE RBF $\sigma = 2^2$ covalent	32	51	24	38	27	65	22	70	27	63	52	25	40	46
MVE RBF $\sigma = 2^2$ kNN k = 2	40	40	6	59	49	35	64	25	47	46	30	27	46	46
MVE RBF $\sigma = 2^2$ kNN k = 3	30	53	12	62	46	38	42	42	41	49	16	59	25	66
MVE RBF $\sigma = 2^2$ kNN k = 4	39	44	26	47	68	22	56	28	39	55	18	50	27	60
MVE RBF $\sigma = 2^2$ kNN k = 5	46	34	24	35	76	18	66	25	46	43	30	30	36	55
MVE RBF $\sigma = 2^3$ covalent	26	55	26	38	26	65	19	73	22	64	18	48	46	42
MVE RBF $\sigma = 2^3$ kNN k = 2	40	41	21	65	56	35	52	27	40	51	34	20	40	48
MVE RBF $\sigma = 2^3$ kNN k = 3	31	49	12	50	46	35	28	56	53	39	20	48	23	67
MVE RBF $\sigma = 2^3$ kNN k = 4	42	37	35	29	77	18	38	36	41	52	36	27	26	61
MVE RBF $\sigma = 2^3$ kNN k = 5	43	34	26	41	70	23	48	27	53	41	30	18	32	57
SPE	45	34	18	44	57	31	50	31	51	38	50	16	45	44
MVEPE	44	41	41	41	41	54	52	38	54	40	36	23	42	50

Table 3. Comparison of different canonization algorithms by estimated significance of differences in retrospective performance. Algorithms were applied to molecular representations in two and three dimensions. For each algorithm applied to 2D and 3D representations of molecules the percentage of screenings is listed where this combination of algorithm and dimensionality significantly outperforms the same algorithm on the other dimensionality. Percentages may not add up to 100 because in some cases differences are not significant. Significance level: 0.01.

	∅		ACE		COX2		DHFR		FXA		PPAR γ		THR	
	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
Centroid Linearization	39	43	59	18	18	76	28	58	79	18	30	20	20	69
Isomap kNN k = 2	43	39	6	74	44	40	55	31	65	24	32	30	55	34
Isomap kNN k = 3	28	47	9	50	23	74	59	14	28	59	25	25	26	59
Isomap kNN k = 4	28	52	21	53	23	67	45	38	31	57	16	43	30	54
Isomap kNN k = 5	24	48	12	47	26	64	22	48	48	39	14	30	24	63
Laplacian Eigenmaps kNN k = 2	49	36	12	74	64	32	59	28	81	15	23	36	57	29
Laplacian Eigenmaps kNN k = 3	30	48	12	53	19	68	50	30	53	30	20	41	23	63
Laplacian Eigenmaps kNN k = 4	23	53	9	59	26	63	28	45	40	49	16	32	18	69
Laplacian Eigenmaps kNN k = 5	22	54	18	68	24	59	19	55	36	44	14	34	21	66
PCA	44	33	26	35	60	26	44	34	74	20	20	34	39	51
MVE EDK covalent	25	45	3	35	31	56	25	61	30	51	16	32	48	34
MVE EDK kNN k = 2	40	39	21	65	59	31	58	20	41	47	32	18	31	51
MVE EDK kNN k = 3	32	44	12	41	56	26	36	34	50	38	14	57	23	66
MVE EDK kNN k = 4	33	40	15	50	67	18	30	30	46	46	16	39	27	59
MVE EDK kNN k = 5	38	38	15	47	75	15	45	23	51	43	14	45	27	56
MVE RBF $\sigma = 2^{-1}$ covalent	46	33	65	9	25	57	48	27	78	18	34	27	28	60
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	42	41	18	44	80	15	58	36	56	33	11	52	28	63
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	42	40	24	50	64	26	59	28	73	19	9	48	21	68
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	39	44	38	32	17	75	25	61	77	20	27	36	48	40
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	40	39	32	12	25	65	52	38	57	37	7	61	69	20
MVE RBF $\sigma = 2^0$ covalent	33	41	24	24	32	57	19	69	39	43	36	18	48	34
MVE RBF $\sigma = 2^0$ kNN k = 2	39	42	0	56	84	11	19	59	46	51	48	20	40	53
MVE RBF $\sigma = 2^0$ kNN k = 3	30	48	15	65	63	25	17	69	36	46	14	34	36	48
MVE RBF $\sigma = 2^0$ kNN k = 4	34	41	15	41	30	60	27	48	46	42	36	20	50	37
MVE RBF $\sigma = 2^0$ kNN k = 5	33	41	26	24	29	65	34	34	55	32	16	45	39	46
MVE RBF $\sigma = 2^1$ covalent	24	46	12	38	29	53	19	67	25	59	20	20	42	40
MVE RBF $\sigma = 2^1$ kNN k = 2	33	45	12	62	35	51	56	27	39	56	11	32	44	43
MVE RBF $\sigma = 2^1$ kNN k = 3	29	45	9	59	61	21	19	61	37	48	16	18	30	60
MVE RBF $\sigma = 2^1$ kNN k = 4	32	37	26	29	52	30	44	25	26	58	11	32	33	50
MVE RBF $\sigma = 2^1$ kNN k = 5	35	39	15	41	63	26	38	41	47	38	16	36	32	55
MVE RBF $\sigma = 2^2$ covalent	27	47	15	29	25	63	20	70	24	58	41	18	39	43
MVE RBF $\sigma = 2^2$ kNN k = 2	38	36	6	56	47	32	61	20	45	44	27	20	45	44
MVE RBF $\sigma = 2^2$ kNN k = 3	29	48	12	53	44	35	39	36	39	46	14	55	23	65
MVE RBF $\sigma = 2^2$ kNN k = 4	37	39	18	38	65	21	55	25	39	53	18	34	26	60
MVE RBF $\sigma = 2^2$ kNN k = 5	43	30	21	26	74	18	66	23	46	43	20	18	34	51
MVE RBF $\sigma = 2^3$ covalent	23	50	15	26	24	63	17	72	20	60	18	41	42	39
MVE RBF $\sigma = 2^3$ kNN k = 2	38	37	18	56	54	31	52	23	39	50	27	16	39	44
MVE RBF $\sigma = 2^3$ kNN k = 3	28	45	9	50	46	32	25	56	50	36	16	30	22	66
MVE RBF $\sigma = 2^3$ kNN k = 4	40	34	32	26	76	18	31	34	40	50	32	18	26	58
MVE RBF $\sigma = 2^3$ kNN k = 5	39	33	21	41	68	23	42	27	51	40	18	11	32	57
SPE	41	31	15	35	57	29	47	31	51	37	34	11	43	42
MVEPE	40	37	35	41	38	51	48	31	52	38	27	14	41	50

Table 4. Percentage of significantly better screenings per target for baseline PhAST and any other canonization algorithm. Dimensionality of molecular representation: 2D. Significance level 0.05. B = baseline PhAST MVE DK ($\beta = 0.4$), C = candidate canonization algorithm named in first column.

	\emptyset		ACE		COX2		DHFR		FXA		PPAR γ		THR	
	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Centroid Linearization	77	13	62	18	69	24	88	9	78	18	77	2	90	6
Isomap kNN k = 2	87	6	85	0	85	9	92	6	93	6	75	5	90	8
Isomap kNN k = 3	72	15	74	9	57	35	89	3	78	15	52	9	81	16
Isomap kNN k = 4	80	13	82	3	67	29	84	11	87	9	70	14	86	10
Isomap kNN k = 5	76	15	79	3	54	38	88	11	82	15	68	11	85	14
Laplacian Eigenmaps kNN k = 2	82	10	82	3	79	17	92	8	90	8	64	14	85	8
Laplacian Eigenmaps kNN k = 3	67	22	79	12	29	61	91	3	64	28	55	16	83	14
Laplacian Eigenmaps kNN k = 4	71	18	74	9	54	40	89	3	75	18	57	20	80	15
Laplacian Eigenmaps kNN k = 5	69	21	76	18	36	55	91	6	72	21	55	14	85	13
PCA	71	21	85	6	29	64	81	11	79	15	70	16	79	16
MVE EDK covalent	59	26	74	12	47	34	81	11	61	31	20	43	72	24
MVE EDK kNN k = 2	66	21	76	0	40	49	86	8	76	17	34	36	83	14
MVE EDK kNN k = 3	60	25	82	6	26	66	80	11	63	25	43	11	67	28
MVE EDK kNN k = 4	61	26	79	6	21	70	72	11	75	19	43	27	76	20
MVE EDK kNN k = 5	65	25	74	12	18	77	80	9	80	14	59	20	78	18
MVE RBF $\sigma = 2^{-1}$ covalent	67	20	82	9	62	28	75	17	71	21	30	36	83	10
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	76	14	85	0	60	33	89	6	82	13	52	23	87	8
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	68	22	88	3	21	72	89	8	76	19	48	18	85	10
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	75	12	74	6	71	21	89	9	88	8	39	18	87	10
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	78	13	79	6	52	37	92	8	92	6	64	11	87	10
MVE RBF $\sigma = 2^0$ covalent	65	23	76	15	57	27	77	17	69	22	32	45	81	11
MVE RBF $\sigma = 2^0$ kNN k = 2	69	18	85	3	49	43	86	6	83	11	25	39	87	8
MVE RBF $\sigma = 2^0$ kNN k = 3	57	27	71	6	15	80	81	9	65	24	32	27	80	14
MVE RBF $\sigma = 2^0$ kNN k = 4	71	18	79	9	49	40	88	9	83	11	41	25	85	12
MVE RBF $\sigma = 2^0$ kNN k = 5	68	20	82	9	32	53	88	8	71	24	50	20	87	9
MVE RBF $\sigma = 2^1$ covalent	61	26	71	12	48	38	80	16	59	35	27	43	80	14
MVE RBF $\sigma = 2^1$ kNN k = 2	69	18	85	3	45	42	88	11	76	14	39	30	84	10
MVE RBF $\sigma = 2^1$ kNN k = 3	57	29	74	6	20	73	77	6	64	29	34	36	74	22
MVE RBF $\sigma = 2^1$ kNN k = 4	63	25	74	12	24	69	86	14	76	17	41	20	75	16
MVE RBF $\sigma = 2^1$ kNN k = 5	63	26	82	6	19	74	78	17	76	17	39	25	81	14
MVE RBF $\sigma = 2^2$ covalent	59	25	59	12	49	30	84	11	61	32	25	45	75	19
MVE RBF $\sigma = 2^2$ kNN k = 2	68	20	85	0	38	48	86	11	78	14	36	34	82	13
MVE RBF $\sigma = 2^2$ kNN k = 3	62	28	82	9	29	67	80	8	61	32	50	23	67	30
MVE RBF $\sigma = 2^2$ kNN k = 4	62	25	76	6	20	72	83	9	74	20	45	23	75	21
MVE RBF $\sigma = 2^2$ kNN k = 5	61	28	82	9	23	70	70	16	75	20	43	27	71	24
MVE RBF $\sigma = 2^3$ covalent	62	24	68	15	57	27	86	8	61	31	23	43	76	18
MVE RBF $\sigma = 2^3$ kNN k = 2	64	21	76	3	34	55	88	6	72	19	36	30	81	14
MVE RBF $\sigma = 2^3$ kNN k = 3	61	27	82	6	29	63	80	11	55	36	52	20	66	29
MVE RBF $\sigma = 2^3$ kNN k = 4	59	29	76	15	20	72	80	13	76	18	30	36	74	20
MVE RBF $\sigma = 2^3$ kNN k = 5	57	28	62	6	23	71	75	14	76	18	34	36	71	22
SPE	69	20	88	3	40	53	80	14	76	18	52	18	78	16
MVEPE	74	16	76	9	66	27	84	6	74	20	64	18	81	15
MVE PRW diameter a = 2	59	23	56	12	63	22	84	8	40	48	36	32	73	16
MVE PRW diameter a = 4	52	25	35	18	54	29	84	9	54	35	23	34	63	23
MVE PRW diameter a = 6	52	25	41	21	52	31	83	11	46	37	32	27	57	24
MVE PRW diameter a = 8	49	27	41	18	47	32	78	9	48	39	23	36	55	29
MVE PRW diameter a = 10	51	28	53	21	51	29	78	9	46	40	20	41	58	27
MVE PRW radius a = 2	55	25	50	15	52	32	86	8	52	36	27	32	62	25
MVE PRW radius a = 4	51	27	38	24	47	37	83	9	49	39	30	30	61	24
MVE PRW radius a = 6	53	29	44	21	54	29	80	11	52	38	27	45	58	27
MVE PRW radius a = 8	51	26	41	32	53	26	78	9	48	38	30	20	57	28
MVE PRW radius a = 10	54	25	53	12	50	33	77	9	47	38	39	30	57	29

Table 5. Percentage of significantly better screenings per target for baseline PhAST and any other canonization algorithm. Dimensionality of molecular representation: 3D. Significance level 0.05. B = baseline PhAST MVE DK ($\beta = 0.4$), C = candidate canonization algorithm named in first column.

	\emptyset		ACE		COX2		DHFR		FXA		PPAR γ		THR	
	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Centroid Linearization	77	14	85	6	44	43	83	16	85	11	80	0	87	8
Isomap kNN k = 2	77	13	41	24	83	10	86	13	90	8	68	20	91	4
Isomap kNN k = 3	62	25	62	21	15	76	92	6	72	22	50	11	82	14
Isomap kNN k = 4	65	23	59	21	25	71	91	6	85	11	52	18	81	13
Isomap kNN k = 5	65	22	50	24	30	59	91	8	84	12	61	11	77	20
Laplacian Eigenmaps kNN k = 2	80	12	71	18	83	15	89	8	96	3	55	18	89	8
Laplacian Eigenmaps kNN k = 3	58	27	47	24	15	77	91	8	69	22	52	14	74	19
Laplacian Eigenmaps kNN k = 4	61	26	62	21	24	65	91	9	69	22	50	14	68	24
Laplacian Eigenmaps kNN k = 5	55	30	38	32	25	65	88	8	67	26	41	32	74	19
PCA	72	18	85	6	35	56	78	11	88	10	68	11	78	16
MVE EDK covalent	49	33	44	18	38	53	83	6	44	43	14	57	73	19
MVE EDK kNN k = 2	64	22	50	24	50	38	88	8	69	25	45	30	84	10
MVE EDK kNN k = 3	52	32	56	15	34	55	81	9	66	25	27	41	47	46
MVE EDK kNN k = 4	61	27	71	15	41	51	84	11	77	15	27	41	64	29
MVE EDK kNN k = 5	59	26	53	26	38	49	81	11	87	10	32	32	66	28
MVE RBF $\sigma = 2^{-1}$ covalent	66	24	82	12	44	46	83	16	89	9	32	36	68	23
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	72	20	71	18	87	11	89	11	82	14	25	52	81	13
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	60	31	59	26	31	63	86	13	87	11	25	52	75	21
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	67	22	68	15	30	59	88	13	94	4	41	30	84	14
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	69	23	82	6	33	59	91	8	92	6	27	52	90	6
MVE RBF $\sigma = 2^0$ covalent	58	25	68	21	44	48	61	14	65	30	32	25	81	11
MVE RBF $\sigma = 2^0$ kNN k = 2	67	20	56	21	82	15	67	17	64	30	48	27	89	7
MVE RBF $\sigma = 2^0$ kNN k = 3	53	32	59	15	22	68	67	22	66	24	27	45	75	17
MVE RBF $\sigma = 2^0$ kNN k = 4	65	20	68	18	24	66	86	13	86	8	41	7	86	11
MVE RBF $\sigma = 2^0$ kNN k = 5	62	28	65	15	27	69	84	9	79	17	34	41	81	17
MVE RBF $\sigma = 2^1$ covalent	49	29	50	21	32	54	69	6	41	48	25	30	74	14
MVE RBF $\sigma = 2^1$ kNN k = 2	58	27	35	35	42	48	88	5	60	32	36	32	85	8
MVE RBF $\sigma = 2^1$ kNN k = 3	50	32	65	21	24	62	56	20	58	33	27	41	72	17
MVE RBF $\sigma = 2^1$ kNN k = 4	59	27	74	6	28	65	86	9	66	28	25	39	77	18
MVE RBF $\sigma = 2^1$ kNN k = 5	58	27	65	12	26	64	78	11	76	17	27	36	75	19
MVE RBF $\sigma = 2^2$ covalent	54	30	59	18	32	59	86	6	40	52	36	25	73	19
MVE RBF $\sigma = 2^2$ kNN k = 2	62	23	47	21	51	38	84	3	67	25	39	43	86	11
MVE RBF $\sigma = 2^2$ kNN k = 3	49	36	59	15	21	66	83	6	57	34	23	57	50	39
MVE RBF $\sigma = 2^2$ kNN k = 4	57	28	68	12	28	58	86	9	71	23	23	39	67	26
MVE RBF $\sigma = 2^2$ kNN k = 5	62	26	68	12	40	50	86	9	75	20	36	36	67	26
MVE RBF $\sigma = 2^3$ covalent	51	32	56	18	33	56	86	6	43	45	16	52	74	16
MVE RBF $\sigma = 2^3$ kNN k = 2	63	23	50	24	43	50	88	5	62	28	50	25	84	8
MVE RBF $\sigma = 2^3$ kNN k = 3	52	33	59	18	24	58	78	9	65	29	39	39	48	43
MVE RBF $\sigma = 2^3$ kNN k = 4	58	26	71	9	43	45	83	9	63	27	30	39	62	29
MVE RBF $\sigma = 2^3$ kNN k = 5	58	27	59	18	33	57	84	6	81	13	25	41	63	30
SPE	70	19	82	9	49	46	75	11	76	20	59	14	78	16
MVEPE	76	14	71	9	68	27	86	13	79	17	75	7	79	14

Table 6. Percentage of significantly better screenings per target for baseline PhAST and any other canonization algorithm. Dimensionality of molecular representation: 2D. Significance level 0.01. B = baseline PhAST MVE DK ($\beta = 0.4$), C = candidate canonization algorithm named in first column.

	\emptyset		ACE		COX2		DHFR		FXA		PPAR γ		THR	
	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Centroid Linearization	75	11	56	9	68	22	84	9	75	18	75	2	89	4
Isomap kNN k = 2	82	5	74	0	85	8	92	6	92	5	64	5	87	7
Isomap kNN k = 3	69	13	68	6	55	35	89	3	77	15	48	7	80	15
Isomap kNN k = 4	77	11	74	3	66	27	84	11	86	8	64	9	85	7
Isomap kNN k = 5	74	13	76	3	51	34	88	9	81	14	61	5	84	12
Laplacian Eigenmaps kNN k = 2	81	8	79	0	78	17	92	5	89	7	61	14	85	8
Laplacian Eigenmaps kNN k = 3	63	21	74	12	26	58	89	3	61	27	48	11	80	13
Laplacian Eigenmaps kNN k = 4	69	15	74	6	51	38	89	3	73	17	50	14	79	12
Laplacian Eigenmaps kNN k = 5	66	19	68	15	33	51	91	3	70	19	50	14	84	11
PCA	69	20	82	3	28	62	80	11	79	13	66	16	78	16
MVE EDK covalent	55	20	59	6	45	31	80	9	60	29	18	23	71	22
MVE EDK kNN k = 2	63	19	71	0	37	49	84	6	73	17	30	30	82	11
MVE EDK kNN k = 3	58	22	76	6	24	62	78	6	61	24	39	9	67	26
MVE EDK kNN k = 4	58	24	76	6	20	68	67	9	75	17	34	23	74	18
MVE EDK kNN k = 5	63	22	65	9	18	73	78	8	79	12	59	14	77	16
MVE RBF $\sigma = 2^{-1}$ covalent	64	18	79	9	59	26	73	13	70	21	23	27	82	10
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	75	13	88	0	56	33	89	6	81	13	48	18	87	8
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	63	20	82	3	20	71	88	8	75	18	30	14	85	9
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	72	11	68	6	68	19	89	9	88	8	30	16	87	8
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	75	12	74	6	51	32	92	6	92	6	52	9	87	10
MVE RBF $\sigma = 2^0$ covalent	60	19	56	15	54	23	75	14	68	21	25	30	80	10
MVE RBF $\sigma = 2^0$ kNN k = 2	69	16	85	0	48	42	86	6	82	10	25	34	87	6
MVE RBF $\sigma = 2^0$ kNN k = 3	53	24	68	6	13	79	81	9	62	23	16	16	79	13
MVE RBF $\sigma = 2^0$ kNN k = 4	68	15	74	9	46	39	86	8	83	9	32	18	85	10
MVE RBF $\sigma = 2^0$ kNN k = 5	65	18	79	3	30	53	86	6	71	23	36	18	87	8
MVE RBF $\sigma = 2^1$ covalent	58	21	68	12	40	35	80	9	57	33	25	23	79	13
MVE RBF $\sigma = 2^1$ kNN k = 2	68	15	82	0	45	40	86	9	73	13	36	20	84	9
MVE RBF $\sigma = 2^1$ kNN k = 3	54	26	74	6	18	71	72	6	62	28	23	25	73	21
MVE RBF $\sigma = 2^1$ kNN k = 4	58	23	68	12	21	67	81	13	75	15	27	14	74	15
MVE RBF $\sigma = 2^1$ kNN k = 5	59	24	82	6	16	72	78	13	74	17	25	25	79	13
MVE RBF $\sigma = 2^2$ covalent	56	20	50	12	45	29	84	8	60	32	23	20	74	16
MVE RBF $\sigma = 2^2$ kNN k = 2	62	17	68	0	37	45	78	9	77	13	30	25	81	13
MVE RBF $\sigma = 2^2$ kNN k = 3	56	25	71	6	26	64	78	8	58	30	34	16	67	27
MVE RBF $\sigma = 2^2$ kNN k = 4	58	23	68	3	19	71	83	9	72	18	32	20	75	19
MVE RBF $\sigma = 2^2$ kNN k = 5	55	25	68	9	21	70	66	13	74	20	34	18	70	21
MVE RBF $\sigma = 2^3$ covalent	58	19	59	9	51	24	86	8	60	30	18	27	75	15
MVE RBF $\sigma = 2^3$ kNN k = 2	63	18	74	3	34	51	88	6	71	17	32	23	81	11
MVE RBF $\sigma = 2^3$ kNN k = 3	56	25	79	6	24	63	78	8	53	33	36	11	65	28
MVE RBF $\sigma = 2^3$ kNN k = 4	56	25	71	6	19	71	75	9	75	16	25	27	73	18
MVE RBF $\sigma = 2^3$ kNN k = 5	53	26	53	3	22	71	73	13	75	17	25	30	70	21
SPE	65	19	85	0	39	53	78	13	73	18	39	18	78	15
MVEPE	71	14	62	3	63	27	81	6	74	18	64	16	81	15
MVE PRW diameter a = 2	55	21	47	15	62	19	83	8	37	46	27	23	72	14
MVE PRW diameter a = 4	49	22	26	18	53	25	84	9	51	32	20	25	61	21
MVE PRW diameter a = 6	47	23	29	21	48	29	77	6	44	36	27	23	55	22
MVE PRW diameter a = 8	45	23	32	15	46	29	75	5	45	38	16	27	54	26
MVE PRW diameter a = 10	47	23	47	15	46	29	70	8	45	39	16	20	57	26
MVE PRW radius a = 2	53	22	47	9	51	29	83	6	50	32	25	30	60	23
MVE PRW radius a = 4	47	23	32	18	43	32	78	5	47	36	20	25	60	21
MVE PRW radius a = 6	48	25	38	18	52	26	70	9	47	36	23	34	56	25
MVE PRW radius a = 8	45	22	26	18	47	24	75	8	47	36	20	18	55	26
MVE PRW radius a = 10	50	22	50	9	46	28	73	8	46	36	27	27	56	26

Table 7. Percentage of significantly better screenings per target for baseline PhAST and any other canonization algorithm. Dimensionality of molecular representation: 3D. Significance level 0.01. B = baseline PhAST MVE DK ($\beta = 0.4$), C = candidate canonization algorithm named in first column.

	\emptyset		ACE		COX2		DHFR		FXA		PPAR γ		THR	
	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Centroid Linearization	76	13	85	6	43	43	81	16	84	10	77	0	87	7
Isomap kNN k = 2	75	12	38	15	82	10	86	13	90	8	64	20	91	4
Isomap kNN k = 3	60	24	59	21	14	72	92	6	71	21	43	9	81	13
Isomap kNN k = 4	64	21	56	15	24	68	91	6	83	11	50	16	81	12
Isomap kNN k = 5	63	20	50	21	27	55	91	8	83	12	52	5	77	19
Laplacian Eigenmaps kNN k = 2	77	11	59	18	82	14	89	8	95	3	52	16	87	8
Laplacian Eigenmaps kNN k = 3	56	24	41	12	15	76	89	8	68	21	48	9	73	17
Laplacian Eigenmaps kNN k = 4	56	24	50	15	22	65	89	6	68	21	41	11	67	23
Laplacian Eigenmaps kNN k = 5	53	27	35	32	24	63	86	6	66	25	36	20	73	17
PCA	69	18	76	6	34	55	78	9	86	10	61	11	76	15
MVE EDK covalent	45	30	32	15	35	50	78	6	43	43	9	50	73	18
MVE EDK kNN k = 2	59	16	32	9	47	32	88	5	67	23	39	20	83	9
MVE EDK kNN k = 3	49	29	56	9	30	52	80	9	62	24	23	36	44	44
MVE EDK kNN k = 4	58	23	65	15	38	49	78	9	75	14	27	25	64	28
MVE EDK kNN k = 5	58	23	53	21	37	48	81	8	86	9	25	25	65	27
MVE RBF $\sigma = 2^{-1}$ covalent	64	22	82	12	39	44	83	16	88	8	23	27	68	22
MVE RBF $\sigma = 2^{-1}$ kNN k = 2	70	19	65	18	86	11	89	11	80	14	20	48	80	13
MVE RBF $\sigma = 2^{-1}$ kNN k = 3	59	25	56	15	29	59	86	11	87	11	20	34	75	21
MVE RBF $\sigma = 2^{-1}$ kNN k = 4	65	21	68	15	29	58	88	11	93	4	30	23	84	14
MVE RBF $\sigma = 2^{-1}$ kNN k = 5	66	19	74	6	30	56	91	6	92	5	20	36	90	4
MVE RBF $\sigma = 2^0$ covalent	55	21	56	15	43	46	56	9	64	29	30	20	81	9
MVE RBF $\sigma = 2^0$ kNN k = 2	64	17	47	18	81	13	67	16	62	29	43	20	86	7
MVE RBF $\sigma = 2^0$ kNN k = 3	49	29	56	12	20	65	66	19	62	23	16	43	75	14
MVE RBF $\sigma = 2^0$ kNN k = 4	64	19	62	12	24	65	86	11	85	8	41	7	86	10
MVE RBF $\sigma = 2^0$ kNN k = 5	59	24	62	9	26	66	84	6	78	15	25	32	80	15
MVE RBF $\sigma = 2^1$ covalent	45	25	41	15	32	54	63	6	40	45	18	20	74	13
MVE RBF $\sigma = 2^1$ kNN k = 2	56	23	35	24	38	45	86	5	59	31	32	27	84	8
MVE RBF $\sigma = 2^1$ kNN k = 3	45	29	50	15	20	59	52	17	55	32	25	36	70	16
MVE RBF $\sigma = 2^1$ kNN k = 4	56	24	65	6	26	63	86	9	64	25	18	23	75	17
MVE RBF $\sigma = 2^1$ kNN k = 5	56	24	59	6	25	64	77	11	76	17	27	27	75	18
MVE RBF $\sigma = 2^2$ covalent	51	28	56	15	31	57	84	5	39	48	23	25	71	17
MVE RBF $\sigma = 2^2$ kNN k = 2	57	20	32	18	49	34	84	3	67	23	27	32	84	10
MVE RBF $\sigma = 2^2$ kNN k = 3	47	33	56	15	19	64	83	6	56	33	20	41	49	37
MVE RBF $\sigma = 2^2$ kNN k = 4	56	24	65	6	28	55	86	8	68	22	20	27	66	25
MVE RBF $\sigma = 2^2$ kNN k = 5	56	23	50	9	35	49	86	9	73	19	27	27	67	23
MVE RBF $\sigma = 2^3$ covalent	48	29	47	18	29	52	84	6	42	43	11	39	74	15
MVE RBF $\sigma = 2^3$ kNN k = 2	58	21	38	18	39	48	84	5	61	27	43	23	82	7
MVE RBF $\sigma = 2^3$ kNN k = 3	47	30	50	12	22	57	78	9	62	28	27	30	44	42
MVE RBF $\sigma = 2^3$ kNN k = 4	55	23	65	3	37	43	78	8	61	26	30	32	62	28
MVE RBF $\sigma = 2^3$ kNN k = 5	55	23	56	12	30	55	83	6	80	11	18	25	63	28
SPE	66	18	71	9	48	43	73	9	74	18	55	11	77	15
MVEPE	72	12	59	6	67	25	86	9	77	15	64	5	77	12

Distance Matrix of Canonization Algorithms based on inverted Rank Correlation

2d Centroid	0	0.496980497	0.510772104	0.488228082
	0.503957085	0.511185162	0.486235448	0.50032099
	0.491802748	0.488833117	0.502616403	0.501889473
	0.488701195	0.521807802	0.490530671	0.515900596
	0.478422992	0.511689047	0.511130136	0.488504634
	0.501088578	0.490956754	0.493545762	0.52416884
	0.522395745			
Isomap covalent	0.496980497	0	0.378996396	0.458508794
	0.246463172	0.354690934	0.298089549	0.29707767
	0.341457998	0.264956194	0.297522189	0.342163988
	0.381314454	0.483057401	0.349722565	0.481012662
	0.520040054	0.32251361	0.321978721	0.283620418
	0.30289588	0.28159221	0.276057175	0.402156969
	0.383604759			
2d Isomap knn 3	0.510772104	0.378996396	0	0.50509923
	0.390453052	0.344200968	0.39413458	0.396625144
	0.365666641	0.393984201	0.395130681	0.353021478
	0.385407356	0.50197285	0.374118161	0.498911832
	0.53851896	0.401731712	0.403302545	0.395917283
	0.393374415	0.394182237	0.392053442	0.43105407
	0.421828856			
Jochum Gasteiger	0.488228082	0.458508794	0.50509923	
	0	0.466898568	0.496897665	0.444838698
	0.485940856	0.457296208	0.463170683	0.48557129
	0.514654573	0.516804985	0.48965819	0.497611635
	0.515062613	0.481384974	0.483999782	0.449841614
	0.463545655	0.447167491	0.462009664	0.516716131
	0.500214114			
Laplacian Eigenmaps	0.503957085	0.246463172	0.390453052	
	0.466898568	0	0.360324991	0.318617623
	0.35841598	0.297405711	0.316473848	0.357940476
	0.395055621	0.478400315	0.359093316	0.476726113
	0.527684371	0.335252183	0.305636537	0.312378743
	0.321750917	0.314060631	0.300440164	0.421110406
	0.398477665			
2d Laplacian Eigenmaps knn 30	0.511185162	0.354690934	0.344200968	
	0.496897665	0.360324991	0	0.377680509
	0.346579722	0.371429636	0.373372588	0.347079866
	0.386332689	0.492175069	0.360090857	0.481706574
	0.541434638	0.372199231	0.36649396	0.363805329
	0.365680239	0.365558541	0.361819427	0.416818391
	0.402874735			
MVE Diffsuion Kernel	0.486235448	0.298089549	0.39413458	
	0.444838698	0.318617623	0.377680509	0
	0.34756358	0.275944392	0.283907676	0.353361281
	0.388034421	0.473502037	0.359481728	0.467546257
	0.514024982	0.363033772	0.358910958	0.291309425
	0.311596497	0.289507096	0.307253763	0.410749693

0.385917511
 2d MVE E covalent 0.50032099 0.29707767 0.396625144
 0.465717264 0.317358196 0.374550085 0.281884397
 0 0.340435326 0.293995859 0.075205278 0.352693546
 0.384055568 0.476890438 0.356042535 0.482288526
 0.523531446 0.358570326 0.352631735 0.291306318
 0.308966392 0.290366532 0.297266446 0.418154633
 0.392746408
 2d MVE E knn 3 0.491802748 0.341457998 0.365666641
 0.485940856 0.35841598 0.346579722 0.34756358
 0.340435326 0 0.348759598 0.344187713 0.278482182
 0.345593931 0.481393518 0.333423843 0.485614487
 0.520896856 0.385874746 0.375320943 0.346030902
 0.348914189 0.347204214 0.348060262 0.405732552
 0.393774533
 MVE Pra 4 0.488833117 0.264956194 0.393984201 0.457296208
 0.297405711 0.371429636 0.275944392 0.293995859
 0.348759598 0 0.294860332 0.34456258 0.391480541
 0.480634089 0.356574024 0.481999326 0.5241785
 0.356414369 0.354115285 0.282399112 0.311183387
 0.278234298 0.285554203 0.417881759 0.396704183
 2d MVE R 2 pow 3 covalent 0.502616403 0.297522189 0.395130681
 0.463170683 0.316473848 0.373372588 0.283907676
 0.075205278 0.344187713 0.294860332 0 0.350627616
 0.386192487 0.480407488 0.352695855 0.482994891
 0.526049914 0.357410259 0.350761671 0.292126606
 0.30929494 0.290203785 0.298881935 0.419339625
 0.389994258
 2d MVE R 2 pow 1 knn 3 0.501889473 0.342163988 0.353021478
 0.48557129 0.357940476 0.347079866 0.353361281
 0.352693546 0.278482182 0.34456258 0.350627616
 0 0.365287401 0.494520575 0.339374166 0.496803235
 0.532033454 0.387406004 0.377500675 0.343135159
 0.355364276 0.342630566 0.344812567 0.415376467
 0.398496446
 2d PCA 0.488701195 0.381314454 0.385407356 0.514654573
 0.395055621 0.386332689 0.388034421 0.384055568
 0.345593931 0.391480541 0.386192487 0.365287401
 0 0.499496446 0.308119927 0.498859058 0.533686158
 0.408578971 0.402843106 0.387300633 0.393122484
 0.390761287 0.38032514 0.421714835 0.419402161
 Prabhakar 0.521807802 0.483057401 0.50197285 0.516804985
 0.478400315 0.492175069 0.473502037 0.476890438
 0.481393518 0.480634089 0.480407488 0.494520575
 0.499496446 0 0.489634046 0.4320807 0.539480505
 0.495862021 0.491699522 0.471603939 0.477466895
 0.473223855 0.476877867 0.508450763 0.4975327
 2d SPE 0.490530671 0.349722565 0.374118161 0.48965819
 0.359093316 0.360090857 0.359481728 0.356042535
 0.333423843 0.356574024 0.352695855 0.339374166

0.308119927	0.489634046	0	0.485350132	0.529010549
0.382005305	0.376835955	0.354812862	0.354488483	
0.357648292	0.349046544	0.415798623	0.391027612	
Weininger	0.515900596	0.481012662	0.498911832	0.497611635
0.476726113	0.481706574	0.467546257	0.482288526	
0.485614487	0.481999326	0.482994891	0.496803235	
0.498859058	0.4320807	0.485350132	0	0.531811959
0.481911828	0.482451904	0.473164284	0.480668831	
0.472455797	0.472609061	0.497674349	0.490617	
3d Centroid	0.478422992	0.520040054	0.53851896	0.515062613
0.527684371	0.541434638	0.514024982	0.523531446	
0.520896856	0.5241785	0.526049914	0.532033454	
0.533686158	0.539480505	0.529010549	0.531811959	
0	0.547794871	0.536869229	0.510902101	0.514817961
0.515160777	0.522375432	0.504979745	0.500332951	
3d Isomap knn 3	0.511689047	0.32251361	0.401731712	
0.481384974	0.335252183	0.372199231	0.363033772	
0.358570326	0.385874746	0.356414369	0.357410259	
0.387406004	0.408578971	0.495862021	0.382005305	
0.481911828	0.547794871	0	0.314889421	0.354632308
0.348373592	0.350040681	0.326446516	0.422770886	
0.40267234				
3d Laplacian Eigenmaps knn 30	0.511130136	0.321978721	0.403302545	
0.483999782	0.305636537	0.36649396	0.358910958	
0.352631735	0.375320943	0.354115285	0.350761671	
0.377500675	0.402843106	0.491699522	0.376835955	
0.482451904	0.536869229	0.314889421	0	0.345297011
0.342828314	0.345736237	0.325577557	0.421624943	
0.396249839				
3d MVE E covalent	0.488504634	0.283620418	0.395917283	
0.449841614	0.312378743	0.363805329	0.291309425	
0.291306318	0.346030902	0.282399112	0.292126606	
0.343135159	0.387300633	0.471603939	0.354812862	
0.473164284	0.510902101	0.354632308	0.345297011	
0	0.278786695	0.062982505	0.280807416	0.399403378
0.374900179				
3d MVE E knn 3	0.501088578	0.30289588	0.393374415	
0.463545655	0.321750917	0.365680239	0.311596497	
0.308966392	0.348914189	0.311183387	0.30929494	
0.355364276	0.393122484	0.477466895	0.354488483	
0.480668831	0.514817961	0.348373592	0.342828314	
0.278786695	0	0.280908984	0.273267078	0.388621288
0.36003185				
3d MVE R 2 pow 3 covalent	0.490956754	0.28159221	0.394182237	
0.447167491	0.314060631	0.365558541	0.289507096	
0.290366532	0.347204214	0.278234298	0.290203785	
0.342630566	0.390761287	0.473223855	0.357648292	
0.472455797	0.515160777	0.350040681	0.345736237	
0.062982505	0.280908984	0	0.279500807	0.400397698
0.373790654				

3d MVE R 2 pow 1 knn 3 0.493545762 0.276057175 0.392053442
0.462009664 0.300440164 0.361819427 0.307253763
0.297266446 0.348060262 0.285554203 0.298881935
0.344812567 0.38032514 0.476877867 0.349046544
0.472609061 0.522375432 0.326446516 0.325577557
0.280807416 0.273267078 0.279500807 0 0.408098123
0.372169743
3d PCA 0.52416884 0.402156969 0.43105407 0.516716131
0.421110406 0.416818391 0.410749693 0.418154633
0.405732552 0.417881759 0.419339625 0.415376467
0.421714835 0.508450763 0.415798623 0.497674349
0.504979745 0.422770886 0.421624943 0.399403378
0.388621288 0.400397698 0.408098123 0 0.323786464
3d SPE 0.522395745 0.383604759 0.421828856 0.500214114
0.398477665 0.402874735 0.385917511 0.392746408
0.393774533 0.396704183 0.389994258 0.398496446
0.419402161 0.4975327 0.391027612 0.490617 0.500332951
0.40267234 0.396249839 0.374900179 0.36003185
0.373790654 0.372169743 0.323786464 0

Distance Matrix of Canonization Algorithms based on Levenshtein Distance

2d Centroid	0	15.07111058	15.14306341	11.44555408
	15.1767537	15.15016244	14.95487908	14.9951871
	14.93045362	15.00553483	15.00072193	14.96739261
	14.81217663	15.02911804	14.86162917	15.26386716
	9.904223318	15.21333173	15.2517146	14.99735291
	15.01347612	14.99843581	15.00048129	14.99254001
	15.00360967			
Isomap covalent	15.07111058	0	8.36866803	15.19492239
	4.192034653	7.704367705	7.022740946	5.985200337
	6.595475875	5.065575743	5.970280351	6.736493803
	7.816508242	14.34749128	7.175069185	15.72012995
	15.0749609	7.063169294	6.930212971	5.741427024
	5.724581879	5.718084466	5.468655998	8.14161954
	7.401395741			
2d Isomap knn 3	15.14306341	8.36866803	0	15.33954999
	8.514739502	6.602454578	9.403200578	8.835037902
	7.938274576	8.843340152	8.787751173	7.704969318
	8.337624835	14.80327277	8.199975936	15.94224522
	15.1553363	8.28660811	8.537721093	8.996991938
	8.787871496	8.977379377	8.77150764	9.20069787
	8.941162315			
Jochum Gasteiger	11.44555408	15.19492239	15.33954999	
	0	15.31223679	15.38154253	15.05763446
	15.22079172	15.18529659	15.13825051	15.23017687
	15.28059199	14.68583805	15.24377331	14.54782818
	11.91613524	15.31319937	15.36397545	15.17386596
	15.18168692	15.17663338	15.19660691	15.31753098
	15.27168812			
Laplacian Eigenmaps	15.1767537	4.192034653	8.514739502	
	15.31223679	0	7.318252918	7.434003128
	6.865359163	5.577908796	6.642281314	7.051137047
	8.049452533	14.22704849	7.210804957	15.61593069
	15.20743593	7.01612321	6.086511852	6.39549994
	6.121766334	6.393213813	5.869811094	8.35988449
	7.323667429			
2d Laplacian Eigenmaps knn 3	15.15016244	7.704367705	6.602454578	
	15.38154253	7.318252918	0	8.959090362
	7.415232824	8.315605824	8.299723258	7.303814222
	8.133437613	14.55023463	7.607869089	15.73950186
	15.18132595	8.082781855	7.461316328	8.416436049
	8.145951149	8.389724462	8.162315004	8.831307905
	8.262062327			
MVE Diffsuion Kernel	14.95487908	7.022740946	9.403200578	
	15.05763446	7.434003128	8.959090362	0
	7.450126339	6.299482613	5.655035495	7.70508964
	8.583202984	14.18481531	8.081337986	15.64673325
	14.9657081	8.91601492	8.7568283	6.272891349
	7.065816388	6.241607508	7.129106004	8.952592949

8.303573577

2d MVE E covalent 14.9951871 5.985200337 8.835037902
15.13548309 6.650222597 8.349897726 5.609433281
0 6.370954157 5.620141981 0.829142101 6.894116232
7.820719528 14.27553844 7.349416436 15.6759716
14.99205872 8.403922512 8.224642041 5.568282998
6.23583203 5.532065937 6.282396823 8.52785465
7.843701119

2d MVE E knn 3 14.93045362 6.595475875 7.938274576
15.22079172 6.865359163 7.415232824 7.450126339
6.370954157 0 6.673926122 6.387077367 4.429069907
6.534953676 14.42750572 6.222115269 15.71904705
15.00890386 8.359643846 8.109613765 6.827337264
6.548911082 6.793045362 6.705691253 8.077487667
7.471784382

MVE Pra 4 15.00553483 5.065575743 8.843340152 15.18529659
5.577908796 8.315605824 6.299482613 5.620141981
6.673926122 0 5.58693298 6.79833955 8.050655757
14.28889424 7.522079172 15.6845145 15.02827578
7.911803634 7.7295151 5.386957045 5.695223198
5.353627722 5.416556371 8.555769462 7.826134039

2d MVE R 2 pow 3 covalent 15.00072193 5.970280351 8.787751173
15.13825051 6.642281314 8.299723258 5.655035495
0.829142101 6.387077367 5.58693298 0 6.807002767
7.805438575 14.2801107 7.289736494 15.67452773
14.99627 8.374684154 8.197328841 5.574780412 6.22704849
5.513776922 6.233425581 8.508001444 7.807484057

2d MVE R 2 pow 1 knn 3 14.96739261 6.736493803 7.704969318
15.23017687 7.051137047 7.303814222 7.70508964
6.894116232 4.429069907 6.79833955 6.807002767
0 7.124172783 14.47334857 6.758753459 15.77391409
14.98147034 8.477800505 8.251594273 7.062086392
6.905185898 7.029960294 6.514739502 8.320178077
7.742750572

2d PCA 14.81217663 7.816508242 8.337624835 15.28059199
8.049452533 8.133437613 8.583202984 7.820719528
6.534953676 8.050655757 7.805438575 7.124172783
0 14.57261461 5.363253519 15.77283119 14.94068103
8.827216941 8.668150644 8.089880881 7.794368909
8.072073156 7.990374203 7.652388401 7.657201299

Prabhakar 15.02911804 14.34749128 14.80327277 14.68583805
14.22704849 14.55023463 14.18481531 14.27553844
14.42750572 14.28889424 14.2801107 14.47334857
14.57261461 0 14.44146312 12.50607628 15.04812899
14.66923355 14.48887017 14.27734328 14.3427987
14.28372037 14.38130189 14.60871135 14.43231861

2d SPE 14.86162917 7.175069185 8.199975936 15.24377331
7.210804957 7.607869089 8.081337986 7.349416436
6.222115269 7.522079172 7.289736494 6.758753459
5.363253519 14.44146312 0 15.69401997 14.97774034

	8.459631813	8.186259175	7.556371074	7.173865961
	7.520635303	7.365178679	7.834797257	6.816147275
Weininger	15.26386716	15.72012995	15.94224522	14.54782818
	15.61593069	15.73950186	15.64673325	15.6759716
	15.71904705	15.6845145	15.67452773	15.77391409
	15.77283119	12.50607628	15.69401997	0 15.29394778
	15.84983756	15.71952834	15.71940801	15.68403321
	15.72783059	15.74864637	15.72337865	15.68030321
3d Centroid	9.904223318	15.0749609	15.1553363	11.91613524
	15.20743593	15.18132595	14.9657081	14.99205872
	15.00890386	15.02827578	14.99627 14.98147034	14.94068103
	15.04812899	14.97774034	15.29394778	0 15.20021658
	15.26579232	14.99578871	14.98905066	15.00625677
	15.02635062	14.82120082	14.86367465	
3d Isomap knn 3	15.21333173	7.063169294	8.28660811	
	15.31319937	7.01612321	8.082781855	8.91601492
	8.403922512	8.359643846	7.911803634	8.374684154
	8.477800505	8.827216941	14.66923355	8.459631813
	15.84983756	15.20021658	0 5.788112141	8.217663338
	7.913969438	8.196125617	7.586451691	8.969317772
	8.461316328			
3d Laplacian Eigenmaps knn 3	15.2517146	6.930212971	8.537721093	
	15.36397545	6.086511852	7.461316328	8.7568283
	8.224642041	8.109613765	7.7295151	8.197328841
	8.251594273	8.668150644	14.48887017	8.186259175
	15.71952834	15.26579232	5.788112141	0 8.009024185
	7.70605222	7.983515822	7.505594995	8.830104681
	8.160510167			
3d MVE E covalent	14.99735291	5.741427024	8.996991938	
	15.17386596	6.39549994	8.416436049	6.272891349
	5.568282998	6.827337264	5.386957045	5.574780412
	7.062086392	8.089880881	14.27734328	7.556371074
	15.71940801	14.99578871	8.217663338	8.009024185
	0 4.947659728	0.524004332	5.57273493	8.074359283
	7.4199254			
3d MVE E knn 3	15.01347612	5.724581879	8.787871496	
	15.18168692	6.121766334	8.145951149	7.065816388
	6.23583203	6.548911082	5.695223198	6.22704849
	6.905185898	7.794368909	14.3427987	7.173865961
	15.68403321	14.98905066	7.913969438	7.70605222
	4.947659728	0 5.009505475	4.4365299	7.729274456
	6.999157743			
3d MVE R 2 pow 3 covalent	14.99843581	5.718084466	8.977379377	
	15.17663338	6.393213813	8.389724462	6.241607508
	5.532065937	6.793045362	5.353627722	5.513776922
	7.029960294	8.072073156	14.28372037	7.520635303
	15.72783059	15.00625677	8.196125617	7.983515822
	0.524004332	5.009505475	0 5.536517868	8.056431236
	7.391288654			
3d MVE R 2 pow 1 knn 3	15.00048129	5.468655998	8.77150764	

	15.19660691	5.869811094	8.162315004	7.129106004	
	6.282396823	6.705691253	5.416556371	6.233425581	
	6.514739502	7.990374203	14.38130189	7.365178679	
	15.74864637	15.02635062	7.586451691	7.505594995	
	5.57273493	4.4365299	5.536517868	0	8.231981711
	7.454939237				
3d PCA	14.99254001	8.14161954	9.20069787	15.31753098	
	8.35988449	8.831307905	8.952592949	8.52785465	
	8.077487667	8.555769462	8.508001444	8.320178077	
	7.652388401	14.60871135	7.834797257	15.72337865	
	14.82120082	8.969317772	8.830104681	8.074359283	
	7.729274456	8.056431236	8.231981711	0	5.890265913
3d SPE	15.00360967	7.401395741	8.941162315	15.27168812	
	7.323667429	8.262062327	8.303573577	7.843701119	
	7.471784382	7.826134039	7.807484057	7.742750572	
	7.657201299	14.43231861	6.816147275	15.68030321	
	14.86367465	8.461316328	8.160510167	7.4199254	
	6.999157743	7.391288654	7.454939237	5.890265913	
	0				

Distance Matrix of Canonization Algorithms based on Damerau Levenshtein Distance

2d Centroid	0	15.07111058	15.14306341	11.44555408
	15.1767537	15.15016244	14.95487908	14.9951871
	14.93045362	15.00553483	15.00072193	14.96739261
	14.81217663	15.02911804	14.86162917	15.26386716
	9.904223318	15.21333173	15.2517146	14.99735291
	15.01347612	14.99843581	15.00048129	14.99254001
	15.00360967			
Isomap covalent	15.07111058	0	8.36866803	15.19492239
	4.192034653	7.704367705	7.022740946	5.985200337
	6.595475875	5.065575743	5.970280351	6.736493803
	7.816508242	14.34749128	7.175069185	15.72012995
	15.0749609	7.063169294	6.930212971	5.741427024
	5.724581879	5.718084466	5.468655998	8.14161954
	7.401395741			
2d Isomap knn 3	15.14306341	8.36866803	0	15.33954999
	8.514739502	6.602454578	9.403200578	8.835037902
	7.938274576	8.843340152	8.787751173	7.704969318
	8.337624835	14.80327277	8.199975936	15.94224522
	15.1553363	8.28660811	8.537721093	8.996991938
	8.787871496	8.977379377	8.77150764	9.20069787
	8.941162315			
Jochum Gasteiger	11.44555408	15.19492239	15.33954999	
	0	15.31223679	15.38154253	15.05763446
	15.22079172	15.18529659	15.13825051	15.23017687
	15.28059199	14.68583805	15.24377331	14.54782818
	11.91613524	15.31319937	15.36397545	15.17386596
	15.18168692	15.17663338	15.19660691	15.31753098
	15.27168812			
Laplacian Eigenmaps	15.1767537	4.192034653	8.514739502	
	15.31223679	0	7.318252918	7.434003128
	6.865359163	5.577908796	6.642281314	7.051137047
	8.049452533	14.22704849	7.210804957	15.61593069
	15.20743593	7.01612321	6.086511852	6.39549994
	6.121766334	6.393213813	5.869811094	8.35988449
	7.323667429			
2d Laplacian Eigenmaps knn 3	15.15016244	7.704367705	6.602454578	
	15.38154253	7.318252918	0	8.959090362
	7.415232824	8.315605824	8.299723258	7.303814222
	8.133437613	14.55023463	7.607869089	15.73950186
	15.18132595	8.082781855	7.461316328	8.416436049
	8.145951149	8.389724462	8.162315004	8.831307905
	8.262062327			
MVE Diffsuion Kernel	14.95487908	7.022740946	9.403200578	
	15.05763446	7.434003128	8.959090362	0
	7.450126339	6.299482613	5.655035495	7.70508964
	8.583202984	14.18481531	8.081337986	15.64673325
	14.9657081	8.91601492	8.7568283	6.272891349

7.065816388 6.241607508 7.129106004 8.952592949
 8.303573577
 2d MVE E covalent 14.9951871 5.985200337 8.835037902
 15.13548309 6.650222597 8.349897726 5.609433281
 0 6.370954157 5.620141981 0.829142101 6.894116232
 7.820719528 14.27553844 7.349416436 15.6759716
 14.99205872 8.403922512 8.224642041 5.568282998
 6.23583203 5.532065937 6.282396823 8.52785465
 7.843701119
 2d MVE E knn 3 14.93045362 6.595475875 7.938274576
 15.22079172 6.865359163 7.415232824 7.450126339
 6.370954157 0 6.673926122 6.387077367 4.429069907
 6.534953676 14.42750572 6.222115269 15.71904705
 15.00890386 8.359643846 8.109613765 6.827337264
 6.548911082 6.793045362 6.705691253 8.077487667
 7.471784382
 MVE Pra 4 15.00553483 5.065575743 8.843340152 15.18529659
 5.577908796 8.315605824 6.299482613 5.620141981
 6.673926122 0 5.58693298 6.79833955 8.050655757
 14.28889424 7.522079172 15.6845145 15.02827578
 7.911803634 7.7295151 5.386957045 5.695223198
 5.353627722 5.416556371 8.555769462 7.826134039
 2d MVE R 2 pow 3 covalent 15.00072193 5.970280351 8.787751173
 15.13825051 6.642281314 8.299723258 5.655035495
 0.829142101 6.387077367 5.58693298 0 6.807002767
 7.805438575 14.2801107 7.289736494 15.67452773
 14.99627 8.374684154 8.197328841 5.574780412 6.22704849
 5.513776922 6.233425581 8.508001444 7.807484057
 2d MVE R 2 pow 1 knn 3 14.96739261 6.736493803 7.704969318
 15.23017687 7.051137047 7.303814222 7.70508964
 6.894116232 4.429069907 6.79833955 6.807002767
 0 7.124172783 14.47334857 6.758753459 15.77391409
 14.98147034 8.477800505 8.251594273 7.062086392
 6.905185898 7.029960294 6.514739502 8.320178077
 7.742750572
 2d PCA 14.81217663 7.816508242 8.337624835 15.28059199
 8.049452533 8.133437613 8.583202984 7.820719528
 6.534953676 8.050655757 7.805438575 7.124172783
 0 14.57261461 5.363253519 15.77283119 14.94068103
 8.827216941 8.668150644 8.089880881 7.794368909
 8.072073156 7.990374203 7.652388401 7.657201299
 Prabhakar 15.02911804 14.34749128 14.80327277 14.68583805
 14.22704849 14.55023463 14.18481531 14.27553844
 14.42750572 14.28889424 14.2801107 14.47334857
 14.57261461 0 14.44146312 12.50607628 15.04812899
 14.66923355 14.48887017 14.27734328 14.3427987
 14.28372037 14.38130189 14.60871135 14.43231861
 2d SPE 14.86162917 7.175069185 8.199975936 15.24377331
 7.210804957 7.607869089 8.081337986 7.349416436
 6.222115269 7.522079172 7.289736494 6.758753459

5.363253519	14.44146312	0	15.69401997	14.97774034
8.459631813	8.186259175	7.556371074	7.173865961	
7.520635303	7.365178679	7.834797257	6.816147275	
Weininger	15.26386716	15.72012995	15.94224522	14.54782818
15.61593069	15.73950186	15.64673325	15.6759716	
15.71904705	15.6845145	15.67452773	15.77391409	
15.77283119	12.50607628	15.69401997	0	15.29394778
15.84983756	15.71952834	15.71940801	15.68403321	
15.72783059	15.74864637	15.72337865	15.68030321	
3d Centroid	9.904223318	15.0749609	15.1553363	11.91613524
15.20743593	15.18132595	14.9657081	14.99205872	
15.00890386	15.02827578	14.99627	14.98147034	14.94068103
15.04812899	14.97774034	15.29394778	0	15.20021658
15.26579232	14.99578871	14.98905066	15.00625677	
15.02635062	14.82120082	14.86367465		
3d Isomap knn 3	15.21333173	7.063169294	8.28660811	
15.31319937	7.01612321	8.082781855	8.91601492	
8.403922512	8.359643846	7.911803634	8.374684154	
8.477800505	8.827216941	14.66923355	8.459631813	
15.84983756	15.20021658	0	5.788112141	8.217663338
7.913969438	8.196125617	7.586451691	8.969317772	
8.461316328				
3d Laplacian Eigenmaps knn 3	15.2517146	6.930212971	8.537721093	
15.36397545	6.086511852	7.461316328	8.7568283	
8.224642041	8.109613765	7.7295151	8.197328841	
8.251594273	8.668150644	14.48887017	8.186259175	
15.71952834	15.26579232	5.788112141	0	8.009024185
7.70605222	7.983515822	7.505594995	8.830104681	
8.160510167				
3d MVE E covalent	14.99735291	5.741427024	8.996991938	
15.17386596	6.39549994	8.416436049	6.272891349	
5.568282998	6.827337264	5.386957045	5.574780412	
7.062086392	8.089880881	14.27734328	7.556371074	
15.71940801	14.99578871	8.217663338	8.009024185	
0	4.947659728	0.524004332	5.57273493	8.074359283
7.4199254				
3d MVE E knn 3	15.01347612	5.724581879	8.787871496	
15.18168692	6.121766334	8.145951149	7.065816388	
6.23583203	6.548911082	5.695223198	6.22704849	
6.905185898	7.794368909	14.3427987	7.173865961	
15.68403321	14.98905066	7.913969438	7.70605222	
4.947659728	0	5.009505475	4.4365299	7.729274456
6.999157743				
3d MVE R 2 pow 3 covalent	14.99843581	5.718084466	8.977379377	
15.17663338	6.393213813	8.389724462	6.241607508	
5.532065937	6.793045362	5.353627722	5.513776922	
7.029960294	8.072073156	14.28372037	7.520635303	
15.72783059	15.00625677	8.196125617	7.983515822	
0.524004332	5.009505475	0	5.536517868	8.056431236
7.391288654				

3d MVE R 2 pow 1 knn 3 15.00048129 5.468655998 8.77150764
15.19660691 5.869811094 8.162315004 7.129106004
6.282396823 6.705691253 5.416556371 6.233425581
6.514739502 7.990374203 14.38130189 7.365178679
15.74864637 15.02635062 7.586451691 7.505594995
5.57273493 4.4365299 5.536517868 0 8.231981711
7.454939237
3d PCA 14.99254001 8.14161954 9.20069787 15.31753098
8.35988449 8.831307905 8.952592949 8.52785465
8.077487667 8.555769462 8.508001444 8.320178077
7.652388401 14.60871135 7.834797257 15.72337865
14.82120082 8.969317772 8.830104681 8.074359283
7.729274456 8.056431236 8.231981711 0 5.890265913
3d SPE 15.00360967 7.401395741 8.941162315 15.27168812
7.323667429 8.262062327 8.303573577 7.843701119
7.471784382 7.826134039 7.807484057 7.742750572
7.657201299 14.43231861 6.816147275 15.68030321
14.86367465 8.461316328 8.160510167 7.4199254
6.999157743 7.391288654 7.454939237 5.890265913
0

Distance Matrix of Screening Methods based on inverted Rank Correlation

MACCS	0	0.691787486	0.604234896	0.81633558	0.825335532
	0.524528856	0.545776535	0.716839538	0.763979141	
	0.621370817	0.626834628	0.758220312	0.596611071	
	0.549738521	0.586082362	0.609575465	0.566874258	
	0.609301341				
CATS2D_raw	0.691787486	0	0.527810006	0.717152297	
	0.71301454	0.704920407	0.753227192	0.770079192	
	0.838101571	0.695314426	0.708013029	0.658030738	
	0.684947671	0.676120852	0.681653649	0.724987478	
	0.74030281	0.763610184			
CATS2D_sens	0.604234896	0.527810006	0	0.802025701	
	0.813246498	0.586017085	0.657385092	0.644284574	
	0.748130893	0.637517198	0.647883523	0.762647571	
	0.508850369	0.498999261	0.496792499	0.567839584	
	0.628762922	0.660251865			
ESshape3D	0.81633558	0.717152297	0.802025701	0	
	0.45126512	0.798631989	0.850255382	0.854053843	
	0.86119084	0.767022521	0.772664169	0.385778005	
	0.725827117	0.742129876	0.710755251	0.771803281	
	0.848821186	0.862117695			
ESshape3D_HYD	0.825335532	0.71301454	0.813246498		
	0.45126512	0	0.803605862	0.848908492	0.884103682
	0.870408442	0.740194649	0.739312919	0.510523939	
	0.791839743	0.793789164	0.765704661	0.798810694	
	0.834886817	0.850504202			
GpiDAPH3	0.524528856	0.704920407	0.586017085	0.798631989	
	0.803605862	0	0.562163742	0.708450488	0.760633679
	0.579464726	0.585005577	0.749366733	0.571611393	
	0.521768278	0.558088067	0.575014163	0.378608491	
	0.440989906				
LINGO	0.545776535	0.753227192	0.657385092	0.850255382	
	0.848908492	0.562163742	0	0.756939017	0.76533775
	0.592743103	0.593401818	0.792181501	0.636293415	
	0.598485771	0.622629691	0.664174062	0.591468277	
	0.610469183				
LIQUID	0.716839538	0.770079192	0.644284574	0.854053843	
	0.884103682	0.708450488	0.756939017	0	0.852400719
	0.8011805	0.813879654	0.828564188	0.639968054	
	0.628385508	0.673973703	0.644158009	0.71465588	
	0.7353891				
PRPS	0.763979141	0.838101571	0.748130893	0.86119084	
	0.870408442	0.760633679	0.76533775	0.852400719	
	0	0.769186741	0.761877445	0.848115451	0.716245886
	0.729228484	0.732664089	0.767157307	0.772588217	
	0.79839223				
PhAST_2D	0.621370817	0.695314426	0.637517198	0.767022521	
	0.740194649	0.579464726	0.592743103	0.8011805	
	0.769186741	0	0.308005447	0.735167705	0.642120383

	0.608697167	0.616099789	0.663466906	0.635162206
	0.651219016			
PhAST_3D	0.626834628	0.708013029	0.647883523	0.772664169
	0.739312919	0.585005577	0.593401818	0.813879654
	0.761877445	0.308005447	0	0.745893869
	0.614684901	0.61840541	0.669957062	0.646797904
	0.661973613			
SIMPLE_DESC	0.758220312	0.658030738	0.762647571	
	0.385778005	0.510523939	0.749366733	0.792181501
	0.828564188	0.848115451	0.735167705	0.745893869
	0	0.71258691	0.721853104	0.693940407
	0.797894918	0.825367109		0.742437411
TAD	0.596611071	0.684947671	0.508850369	0.725827117
	0.791839743	0.571611393	0.636293415	0.639968054
	0.716245886	0.642120383	0.640048008	0.71258691
	0	0.276732262	0.246919091	0.464929319
	0.644651874			0.623813501
TAT	0.549738521	0.676120852	0.498999261	0.742129876
	0.793789164	0.521768278	0.598485771	0.628385508
	0.729228484	0.608697167	0.614684901	0.721853104
	0.276732262	0	0.331133564	0.471311977
	0.593722361			0.568275256
TGD	0.586082362	0.681653649	0.496792499	0.710755251
	0.765704661	0.558088067	0.622629691	0.673973703
	0.732664089	0.616099789	0.61840541	0.693940407
	0.246919091	0.331133564	0	0.415864882
	0.65368489			0.630170409
TGT	0.609575465	0.724987478	0.567839584	0.771803281
	0.798810694	0.575014163	0.664174062	0.644158009
	0.767157307	0.663466906	0.669957062	0.742437411
	0.464929319	0.471311977	0.415864882	0
	0.652180369			0.614969762
piDAPH3	0.566874258	0.74030281	0.628762922	0.848821186
	0.834886817	0.378608491	0.591468277	0.71465588
	0.772588217	0.635162206	0.646797904	0.797894918
	0.623813501	0.568275256	0.630170409	0.614969762
	0	0.338718486		
piDAPH4	0.609301341	0.763610184	0.660251865	0.862117695
	0.850504202	0.440989906	0.610469183	0.7353891
	0.79839223	0.651219016	0.661973613	0.825367109
	0.644651874	0.593722361	0.65368489	0.652180369
	0.338718486	0		

Distance Matrix of Screening Methods based on Significance (0.01)

MACCS	0	0.5838641	0.010640128	0.790030577	0.811569432
	0.163569129	0.127220374	0.443175522	0.150660678	
	0.066407638	0.027893327	0.722048465	0.032994218	
	0.08631428	0.01950191	0.110819333	0.154356922	
	0.251801684				
CATS2D_raw	0.5838641	0	0.617480655	0.521846705	
	0.529398508	0.653498078	0.721918757	0.029298078	
	0.191686759	0.610583805	0.612417706	0.140256315	
	0.629480539	0.686522234	0.675337269	0.518477949	
	0.381158711	0.372293701			
CATS2D_sens	0.010640128	0.617480655	0	0.657625428	
	0.656375518	0.150387592	0.19588545	0.526971282	
	0.142119346	0.220661013	0.215237205	0.518072139	
	0.056287607	0.171161625	0.109878752	0.170175596	
	0.094799979	0.111727937			
ESshape3D	0.790030577	0.521846705	0.657625428	0	
	0.138635168	0.785230728	0.841314612	0.533563414	
	0.589554053	0.883092222	0.878850764	0.631539072	
	0.743691863	0.802970466	0.75125772	0.718059727	
	0.717283001	0.616297035			
ESshape3D_HYD	0.811569432	0.529398508	0.656375518		
	0.138635168	0	0.818500503	0.844444623	0.512442405
	0.583905141	0.882895071	0.884042316	0.614352418	
	0.747835595	0.805892011	0.761991677	0.733352957	
	0.725446724	0.651993887			
GpiDAPH3	0.163569129	0.653498078	0.150387592	0.785230728	
	0.818500503	0	0.074525561	0.512853713	0.276930409
	8.66E-04	0.056036478	0.661594818	0.169238744	0.108601371
	0.196269563	0.289832144	0.44853586	0.464429728	
LINGO	0.127220374	0.721918757	0.19588545	0.841314612	
	0.844444623	0.074525561	0	0.665361409	0.269062341
	0.100234157	0.073858189	0.746756514	0.272109157	
	0.117602705	0.226519556	0.391439674	0.18185819	
	0.314421884				
LIQUID	0.443175522	0.029298078	0.526971282	0.533563414	
	0.512442405	0.512853713	0.665361409	0	0.158389413
	0.57420793	0.539300877	0.225108875	0.567155145	
	0.690576822	0.603324937	0.498448003	0.285869984	
	0.245238165				
PRPS	0.150660678	0.191686759	0.142119346	0.589554053	
	0.583905141	0.276930409	0.269062341	0.158389413	
	0	0.105925259	0.092664066	0.404327319	0.208341965
	0.251938705	0.169450538	0.084676808	0.095371278	
	0.039781639				
PhAST_2D	0.066407638	0.610583805	0.220661013	0.883092222	
	0.882895071	8.66E-04	0.100234157	0.57420793	0.105925259
	0	0.13690565	0.754991296	0.132313222	0.055691443
	0.147568579	0.316434124	0.151444885	0.254436578	

PhAST_3D	0.027893327	0.612417706	0.215237205	0.878850764
	0.884042316	0.056036478	0.073858189	0.539300877
	0.092664066	0.13690565	0	0.692774034
	0.071998586	0.17469046	0.327335398	0.108870375
	0.245285266			
SIMPLE_DESC	0.722048465	0.140256315	0.518072139	
	0.631539072	0.614352418	0.661594818	0.746756514
	0.225108875	0.404327319	0.754991296	0.692774034
	0	0.619272767	0.664631312	0.598881402
	0.521349357	0.459981098		
TAD	0.032994218	0.629480539	0.056287607	0.743691863
	0.747835595	0.169238744	0.272109157	0.567155145
	0.208341965	0.132313222	0.149946669	0.619272767
	0	0.28972976	0.036647118	0.219710379
	0.162089831			0.084167012
TAT	0.08631428	0.686522234	0.171161625	0.802970466
	0.805892011	0.108601371	0.117602705	0.690576822
	0.251938705	0.055691443	0.071998586	0.664631312
	0.28972976	0	0.262943161	0.352678713
	0.264410096			0.212732023
TGD	0.01950191	0.675337269	0.109878752	0.75125772
	0.761991677	0.196269563	0.226519556	0.603324937
	0.169450538	0.147568579	0.17469046	0.598881402
	0.036647118	0.262943161	0	0.290615754
	0.189966805			0.058684846
TGT	0.110819333	0.518477949	0.170175596	0.718059727
	0.733352957	0.289832144	0.391439674	0.498448003
	0.084676808	0.316434124	0.327335398	0.531870547
	0.219710379	0.352678713	0.290615754	0
	0.035279019			0.014665438
piDAPH3	0.154356922	0.381158711	0.094799979	0.717283001
	0.725446724	0.44853586	0.18185819	0.285869984
	0.095371278	0.151444885	0.108870375	0.521349357
	0.084167012	0.212732023	0.058684846	0.014665438
	0	0.041746778		
piDAPH4	0.251801684	0.372293701	0.111727937	0.616297035
	0.651993887	0.464429728	0.314421884	0.245238165
	0.039781639	0.254436578	0.245285266	0.459981098
	0.162089831	0.264410096	0.189966805	0.035279019
	0.041746778	0		

Distance Matrix of Screening Methods based on Significance (0.05)

MACCS	0	0.58122645	0.034658251	0.791172755	0.809738037
	0.1708456	0.109176271	0.445693149	0.146374522	
	0.048483634	0.020804981	0.720631868	0.027922818	
	0.093327197	0.025623974	0.120515872	0.154447319	
	0.255419391				
CATS2D_raw	0.58122645	0	0.639036612	0.527416382	
	0.52982844	0.64879737	0.731433048	0.046705853	
	0.202057622	0.616744663	0.612708867	0.133947825	
	0.625414384	0.668634378	0.680928446	0.528387647	
	0.375176152	0.363295715			
CATS2D_sens	0.034658251	0.639036612	0	0.680729989	
	0.659787395	0.130285253	0.189451626	0.532436083	
	0.162200186	0.226751997	0.205458564	0.539399969	
	0.071499172	0.177012329	0.104099696	0.168219112	
	0.098714478	0.138340815			
ESshape3D	0.791172755	0.527416382	0.680729989	0	
	0.095913596	0.787681709	0.837646512	0.536578765	
	0.60649768	0.884970461	0.88477488	0.650745311	
	0.739214524	0.798220161	0.752977706	0.716954015	
	0.717025003	0.62430756			
ESshape3D_HYD	0.809738037	0.52982844	0.659787395		
	0.095913596	0	0.821872131	0.851525885	0.534540988
	0.585245821	0.88511462	0.886604704	0.607094388	
	0.751928316	0.803261278	0.760766187	0.735392753	
	0.715668083	0.638077763			
GpiDAPH3	0.1708456	0.64879737	0.130285253	0.787681709	
	0.821872131	0	0.06085882	0.521340538	0.280358782
	0.005497824	0.046232877	0.667786833	0.159663438	
	0.110424394	0.188261854	0.297306394	0.440678522	
	0.464001878				
LINGO	0.109176271	0.731433048	0.189451626	0.837646512	
	0.851525885	0.06085882	0	0.667923798	0.267230947
	0.105021725	0.075954731	0.74392116	0.254850787	
	0.127768459	0.197411349	0.396806749	0.205808137	
	0.326512489				
LIQUID	0.445693149	0.046705853	0.532436083	0.536578765	
	0.534540988	0.521340538	0.667923798	0	0.179335685
	0.574634082	0.542629441	0.238201112	0.56510222	
	0.695811722	0.605281422	0.501282071	0.289546454	
	0.250251534				
PRPS	0.146374522	0.202057622	0.162200186	0.60649768	
	0.585245821	0.280358782	0.267230947	0.179335685	
	0	0.109107233	0.08172248	0.403253486	0.216998641
	0.258094253	0.172507423	0.063859175	0.113380444	
	0.027391746				
PhAST_2D	0.048483634	0.616744663	0.226751997	0.884970461	
	0.88511462	0.005497824	0.105021725	0.574634082	
	0.109107233	0	0.13188391	0.764139165	0.126674467

	0.061893835	0.124167058	0.31962978	0.14248208
	0.242564764			
PhAST_3D	0.020804981	0.612708867	0.205458564	0.88477488
	0.886604704	0.046232877	0.075954731	0.542629441
	0.08172248	0.13188391	0	0.687335799
	0.087425562	0.175345228	0.336621241	0.088492255
	0.233133401			
SIMPLE	0.720631868	0.133947825	0.539399969	0.650745311
	0.607094388	0.667786833	0.74392116	0.238201112
	0.403253486	0.764139165	0.687335799	0
	0.672920113	0.605203573	0.535570048	0.511031458
	0.459238505			
TAD	0.027922818	0.625414384	0.071499172	0.739214524
	0.751928316	0.159663438	0.254850787	0.56510222
	0.216998641	0.126674467	0.139842007	0.615481904
	0	0.300333985	0.043709142	0.253342771
	0.169737714			
TAT	0.093327197	0.668634378	0.177012329	0.798220161
	0.803261278	0.110424394	0.127768459	0.695811722
	0.258094253	0.061893835	0.087425562	0.672920113
	0.300333985	0	0.274446302	0.347312603
	0.262665317			
TGD	0.025623974	0.680928446	0.104099696	0.752977706
	0.760766187	0.188261854	0.197411349	0.605281422
	0.172507423	0.124167058	0.175345228	0.605203573
	0.043709142	0.274446302	0	0.294093956
	0.191973438			
TGT	0.120515872	0.528387647	0.168219112	0.716954015
	0.735392753	0.297306394	0.396806749	0.501282071
	0.063859175	0.31962978	0.336621241	0.535570048
	0.253342771	0.347312603	0.294093956	0
	0.039568479			
piDAPH3	0.154447319	0.375176152	0.098714478	0.717025003
	0.715668083	0.440678522	0.205808137	0.289546454
	0.113380444	0.14248208	0.088492255	0.511031458
	0.086677722	0.21438495	0.068430324	0.009482217
	0	0.030590502		
piDAPH4	0.255419391	0.363295715	0.138340815	0.62430756
	0.638077763	0.464001878	0.326512489	0.250251534
	0.027391746	0.242564764	0.233133401	0.459238505
	0.169737714	0.262665317	0.191973438	0.039568479
	0.030590502	0		

Appendix C

Authors: Hähnke, V.
Schneider, G.

Title: Pharmacophore Alignment Search Tool: Influence of Scoring Systems on Text-based Similarity Searching

Journal: Journal of Computational Chemistry
Accepted for publication (see letter from the editor)

Letter from the Editor:

Date:03-Dec-2010
Ref.: JCC-10-0499.R1

Dear Prof. Schneider:

I am pleased to inform you that your manuscript, Pharmacophore Alignment Search Tool: Influence of Scoring Systems on Text-based Similarity Searching, is acceptable for publication in the Journal of Computational Chemistry. Thank you for publishing your work in our journal.

We must receive a completed Copyright Transfer Agreement, which can be downloaded at www.wiley.com/go/ctapsus.

Please fax this form, with a cover page bearing the JCC manuscript number, directly to the attention of the Production Staff at (USA) 717-738-9478 or (USA) 717-738-9479.

Thank you for your support of the Journal of Computational Chemistry. I look forward to seeing more of your work in the future.

Sincerely,

Prof. Gernot Frenking
Editor, Journal of Computational Chemistry



Pharmacophore Alignment Search Tool: Influence of Scoring Systems on Text-based Similarity Searching

Journal:	<i>Journal of Computational Chemistry</i>
Manuscript ID:	JCC-10-0499.R1
Wiley - Manuscript type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Haehnke, Volker; ETH Zurich, CHAB Schneider, Gisbert; ETH, DCHAB
Key Words:	Global alignment, Virtual screening, Similarity, Pharmacophore elucidation, Line notation

SCHOLARONE™
Manuscripts

Hähnke et al.

1

Pharmacophore Alignment Search Tool: Influence of Scoring Systems on Text-based Similarity Searching

Volker Hähnke, Gisbert Schneider*

Swiss Federal Institute of Technology (ETH), Institute of Pharmaceutical Sciences,
Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland

* author to whom correspondence should be sent:

Prof. Dr. Gisbert Schneider, Swiss Federal Institute of Technology (ETH), Department of
Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, HCI H411,
Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland

Email: gisbert.schneider@pharma.ethz.ch

Phone: +41 44 633 7327

Hähnke et al.

2

ABSTRACT

The text-based similarity searching method PhAST is grounded on pairwise comparisons of potential pharmacophoric points between a query and screening compounds. The underlying scoring matrix is of critical importance for successful virtual screening and hit retrieval from large compound libraries. Here, we compare three conceptually different computational methods for systematic deduction of scoring matrices: assignment-based, alignment-based, and stochastic optimization. All three methods resulted in optimized pharmacophore scoring matrices with significantly superior retrospective performance in comparison to simplistic scoring schemes. Computer-generated similarity matrices of pharmacophoric features turned out to agree well with a manually constructed matrix. We introduce the concept of position-specific scoring to text-based similarity searching so that knowledge about specific ligand-receptor binding patterns can be included, and demonstrate its benefit for hit retrieval. The approach was also used for automated pharmacophore elucidation in agonists of peroxisome proliferator activated receptor (PPAR) gamma, successfully identifying key interactions for receptor activation.

KEYWORDS

Global alignment, Virtual screening, Similarity, Pharmacophore elucidation, Line notation

INTRODUCTION

The Pharmacophore Alignment Search Tool (PhAST) is a string-based approach to virtual screening utilizing topological molecule information.¹⁻³ It reduces each molecule to an unambiguous linear representation by describing its potential pharmacophore in three steps: i) each non-hydrogen atom of the molecular graph is replaced by a potential pharmacophoric point (PPP) symbol, and hydrogen atoms are removed, ii) vertices of this 'pharmacophore graph' are canonically labeled, and iii) vertex symbols are concatenated into a string in increasing order according to their canonic labels. For virtual screening, both the screening compound collection (compound 'library') and the query molecule(s) are converted as described, and the resulting 'PhAST-sequences' are compared using pairwise global sequence alignment⁴. Molecular similarity is calculated as the ratio of the alignment score and the alignment length for the purpose of retrieving pharmacophorically similar molecules from compound libraries.

Previously, we analyzed the impact of structure canonization algorithms,¹⁻³ sequence alignment evaluation,^{1,2} and the dimensionality of molecular representation³ on the virtual screening performance of PhAST. Here, we investigate the effect of the PPP scoring system for matches and mismatches in the alignment on virtual screening performance. For this purpose we adapted methods that are related to the approach used for score calculations in the Point-Accepted-Mutation (PAM) matrix⁵ and BLOcks SUBstitution Matrix (BLOSUM)⁶ used for protein sequence alignments. Scores for matches and mismatches are determined from i) kernel-based assignments of potential pharmacophoric points⁷ as well as from ii) global pairwise sequence alignments,⁴ both created from bioactive reference ligands. As reference datasets we employed a dataset collected by Krier & Hutter⁸ for the construction of a score matrix. In addition to systematic determination of scores from reference alignments we performed stochastic optimization of match- and mismatch-scores. The overall aim was to quantify the influence of PPP scoring on similarity searching with PhAST.

A second aim was to assess the usefulness and effect of position-specific scoring. PhAST employs a general score matrix for matches and mismatches of PPPs. However, sequence alignment allows for using a position-specific scoring matrix⁹ that scores the same matches and mismatches differently depending on PPP symbol positions. For protein sequences, it is common practice to use explicit position-specific scoring matrices with a specific set of match- and mismatch-scores for each residue position.^{9,10} In analogy, we have implemented a weighting scheme based on an implicit definition of a position-specific score matrix: Here, positional specificity is achieved through a weighting factor associated with a particular position of the query sequence. This way it is possible to incorporate knowledge about the relative importance of pharmacophoric features into a PhAST similarity search. A weighting factor > 1 increases the influence of this position on alignment generation and the alignment score used for similarity assessment, potentially resulting in better contrasting between compounds with and without this particular feature.

We compared the different score matrices and the effects of positional weighting of PhAST-sequences by retrospective virtual screening of a collection of drugs and lead compounds (COBRA).¹¹ For statistical evaluation we used the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC)¹² in combination with a paired permutation test for significance assessment.¹³

METHODS

Sequence Alignment

Sequence alignment is used in bioinformatics to estimate the phylogenetic relationship between two sequences (DNA, RNA, amino acid sequences). To create the alignment of two sequences $X = x_1x_2\dots x_n$ and $Y = y_1y_2\dots y_m$, their symbols are matched. Thereby the symbol order is retained and gaps may be inserted to improve the matching (insertion of paired gaps is forbidden). Three cases exist: (i) x_i is aligned to y_j and $x_i = y_j$ (match), (ii) x_i is aligned to y_j and $x_i \neq y_j$ (mismatch), (iii) x_i is aligned to a gap in Y , or y_j is aligned to a gap in X . In protein sequence alignment, matches represent 'conserved' residues. Mismatches may arise from mutations, and gaps from insertions or deletions in an assumed evolutionary process of the compared sequences. Consequently, matches are rewarded with a positive score, mismatches are -- depending on the specific case -- either rewarded with a positive score or penalized by a negative score contribution, and gaps are always penalized by a negative score. The optimal alignment is the one with the highest score (summed over the whole alignment). It can be computed using dynamic programming¹⁴. Instead of the original Needleman-Wunsch algorithm¹⁴ we employed a faster method described by Durbin *et al.*⁴ It can be derived from a simple finite state machine and therefore will be referred to as 'FSM algorithm'. We previously demonstrated that it runs 60% faster than the Needleman-Wunsch algorithm and the calculated alignments are nearly identical.²

Alignment Evaluation

Sequence alignments are used for similarity assessment between PhAST-sequences. In a previous study,² we identified the alignment score normalized to the alignment length to be the best performing alignment evaluation method for our purpose so far. In the present work we only considered this evaluation method for comparison of PhAST screening performance.

Matrix Calculation

The main focus of this study lies on the systematic construction and evaluation of scoring schemes for PPPs using a set of bioactive reference compounds. The pharmacophore model employed in PhAST is presented in Table 1 and Table 2, the current scoring scheme is given in Table 3.

Dataset

We used a dataset of reference compounds compiled by Krier & Hutter for score calculation.⁸ This set was used for the construction of a score matrix. It contains molecules of 33 therapeutic classes. For evaluation of new scoring schemes, we performed retrospective screenings using the COBRA collection of drugs and lead compounds¹¹ (version 6.1, 8311 bioactive compounds; see Table 4 for a list of the selected targets). We removed duplicates and eliminated overlap between the Krier dataset and our COBRA library, resulting in a total of 1268 compounds for the Krier dataset distributed among compound classes as shown in Table 5. Both datasets are fairly similar with respect to their PPP composition (Table 6).

Methodology

PAM⁵ and BLOSUM⁶ matrices used in protein sequence alignment had been constructed based on the idea that scores should reflect the frequency of point mutations in proteins: Frequent events should receive low scores, and rare events should contribute to the overall alignment score by high score values. For this purpose, multiple sequence alignments of

1
2
3 closely related protein sequences were constructed, and the numbers of symbol occurrences in
4 gap-free regions used for the calculations of log-odds scores. We adopted this idea by using
5 PhAST-sequences generated from compounds of the same activity class (instead of closely
6 related protein sequences) and pairwise vertex assignments and symbol alignments (instead of
7 a multiple sequence alignment). The usage of multiple sequence alignments for estimation of
8 symbol alignment frequencies was not possible, as each PhAST-sequence would have to be
9 aligned in its original and inverted orientation. This is due to the PhAST canonization
10 algorithm which employs minimum volume embedding, where small modifications to the
11 molecular graph can invert the one-dimensional coordinate system. For a dataset with n
12 molecules we would have to create 2^n multiple sequence alignments in each iteration, which
13 is practically not feasible. Scores of matches and mismatches were calculated according to
14 Eq. (1):
15
16
17

$$18 \quad s(p_i, p_j) = \text{round} \left(c * \log_b \frac{h_{(p_i, p_j)}}{h_{p_i} * h_{p_j}} \right), \quad (1)$$

19
20
21
22
23 where c is constant, b the logarithm base, h_{p_i} the relative frequency of PPP type i ,
24 h_{p_j} the relative frequency of PPP type j and $h_{(p_i, p_j)}$ the relative frequency of the event of
25 alignment or assignment of PPPs of type i and type j . We used $c = 10$ and $b = 10$, as these
26 settings yield scores in a range that are comparable to our original score matrix. As default
27 frequency for assignment / alignment types not observed we used a value of 10^{-5} to avoid
28 calculating $\log(0)$.
29

30
31 Meaningful assignments or matches and mismatches in alignments between molecules
32 from the same activity class are only obtained if reference compounds have the same binding
33 mode at the same biological target. For both datasets used in this study the target binding
34 modes of most of the compounds are unknown. In studies specifically investigating
35 bioisosteric replacement the related problem of identifying related compound pairs is tackled
36 by computing a similarity measure between compounds in combination with a threshold t .¹⁶
37 Compound pairs exhibiting a similarity $> t$ are treated as if they had the same binding mode
38 and it is assumed that structural differences are caused by the exchange of bioisosters. We
39 adopted this idea and used MACCS keys¹⁷ and the Tanimoto coefficient¹⁸ as similarity index:
40 Only compound pairs with Tanimoto similarity < 0.98 were included in the process of score
41 calculation to exclude identical molecules and trivial analogues. For t as lower bound we
42 used 0, 0.5, 0.75, and 0.90. For each compound class presented in Table 5, we included each
43 unique pair of non-identical compounds exactly once.
44
45
46

47 *Assignment-based matrices*

48
49 The iterative similarity optimal assignment (ISOA) kernel⁷ generates assignments between
50 two labeled graphs A and B , assigning each vertex of the smaller structure to exactly one
51 vertex of the larger one. In a first step, ISOAK computes a similarity value for each vertex
52 pair (v_i^A, v_j^B) . The similarity of two vertices is influenced by two components. The first
53 component compares the isolated vertices based on their labels. For this purpose ISOAK uses
54 the Dirac kernel that returns '1' if two vertices have identical labels and '0' otherwise. The
55 second component of vertex similarity considers the environment of each vertex for similarity
56 assessment and returns high similarity values if these neighborhoods are similar. Recursive
57 measurement incorporates vertex similarities of neighboring vertices as well as a comparison
58 of the connecting edges. For edge comparison, a Dirac kernel based on the bond order labels
59 is applied. The recursive nature of the given vertex similarity definition is expressed by an
60 iterative computation, where vertex similarities of pairs of neighboring vertices incorporated

1
2
3 in the actual calculation of iteration i are taken from results of the previous iteration $i-1$. In
4 iteration 0, only direct vertex comparisons are employed and the neighborhood is ignored.
5 The final similarity of two vertices is expressed as a weighted sum of the two components,
6 where the influence of each component is controlled by parameter $0 \leq \gamma \leq 1$. Component 1
7 (vertex label) is weighted by $1-\gamma$, whereas component 2 (neighborhood) is weighted by γ .
8 As a result, high γ increase the influence of the topological graph neighborhood on vertex
9 comparison.

10
11 We applied the ISOA kernel to graphs of PPPs created from reference compounds. We
12 used the ISOA kernel with settings for γ : 0.25 (high influence of vertex label), 0.5 (equal
13 influence), and 0.75 (high influence of vertex neighborhood). Background frequencies of
14 PPPs necessary for the calculation of log-odds scores were determined from all PPPs involved
15 in assignments, and unassigned vertices were ignored. Combining three settings for γ with
16 four different thresholds for minimum similarity of molecules, we calculated 12 different
17 score matrices based on symbol and assignment frequencies following Eq. 1.

18
19 Gap penalties were optimized using grid-search, as it is difficult to choose them by
20 intuition.¹⁹ For each score matrix based on ISOAK assignments each penalty combination
21 starting from gap open penalty = -2 and gap extension penalty = -1 up to gap open penalty = -
22 28 and gap extension penalty = -27 with the gap extension penalty lower than the gap open
23 penalty was evaluated in a series of retrospective screenings (cf. Screening Protocol 1). Gap
24 penalties exceeding -28 seem unreasonable as no mismatch penalty exceeds this value.
25
26

27 28 *Alignment-based matrices*

29 BLOSUM matrices used for scoring matches and mismatches in protein sequence alignments
30 are constructed in an iterative process:⁶ First, scores are estimated from symbol- and symbol
31 alignment-frequencies in gap-free blocks of multiple sequence alignments of closely related
32 protein sequences. These scores are then used to recalculate the multiple sequence alignments,
33 yielding altered frequencies and new scores. This process is iterated three times. We adopted
34 this BLOSUM concept for the construction of score matrices for PPPs in PhAST: Based on
35 all pairwise alignments within each compound class fulfilling the similarity constraints we
36 determined background frequencies of symbols and alignment events, yielding a first matrix
37 of log-odds scores calculated according to Eq. 1 (gapped regions were excluded). Using this
38 score matrix, the same sequences were re-aligned, resulting in new frequencies and a new
39 score matrix. We iterated this process until convergence: When the actual matrix was
40 identical to any of the matrices generated in previous iterations, the process was terminated.
41
42

43 We performed the iterative construction of score matrices with fixed gap open penalty
44 = 5 and gap extension penalty = 1. We chose this combination because of its good
45 performance with our original score matrix.^{1,2} Using variable gap penalties as additional free
46 parameters is unnecessary because in this approach scores are adapted to the penalty
47 combination. For the initial start scoring scheme we used a primitive match = +2 mismatch =
48 -1 system and our original score matrix (Table 3). Combining these two basic scoring systems
49 with four different similarity thresholds for pair filtering resulted in a total of eight alignment-
50 based scoring matrices.
51
52

53 54 *Stochastic Optimization*

55 As an alternative approach to deducing alignment scores from reference alignments we
56 optimized score matrices in a multi-start stochastic optimization: 50 score matrices were
57 randomly initialized with uniformly distributed scores in $[-20,20]$. All matrices were
58 evaluated on the Krier dataset as training data, results were evaluated using the BEDROC
59 metric.¹² The matrix with highest averaged performance served as template for the generation
60 of 49 new score matrices. With probability 0.1 scores were modified by adding a uniformly
distributed number from in the interval $[-10,10]$. A fiftieth score matrix was again built

randomly with uniformly distributed scores. The template matrix was not part of the next cycle. This procedure was repeated for 100 iterations. The best performing matrix of each generation was evaluated in a set of virtual screenings on the COBRA library as test data (see Screening Protocol 1). We performed three independent optimization runs.

We chose targets with IDs 5, 6, 10, 13, 17, 23, 24, 29 and 31 from the Krier dataset (Table 5) for matrix evaluation by retrospective screening. This choice was guided by recommendations by Truchon and Bayly,¹² who suggest ratios of actives to decoys that avoid saturation effects in retrospective evaluation. BEDROC is mainly influenced by the early part of a ranked list. If the ratio between active and inactive compounds exceeds a certain threshold, and actives 'saturate' the early ranks, and BEDROC values become meaningless. We chose targets so that the saturation effect remained below 20%, calculated as described by Truchon and Bayly.¹² Sequence alignments were computed with fixed gap open penalty of 5 and gap extension penalty of 1, as these values yielded good results with the original score matrix (Table 3).

Screening Protocol 1

All score matrices were evaluated in a series of retrospective screenings. For each target (Table 4) each active was used once as query, resulting in 689 screenings. Each screening was evaluated with the (BEDROC) metric.¹² BEDROC scores were calculated with $\alpha = 20$, the suggested default value for evaluation.¹² We first evaluated screening performance for each target by averaging the corresponding BEDROC scores. Total retrospective performance was expressed as the mean of these averages. We used the mean of averages to give equal weight to each target although the COBRA library contains unequal numbers of actives for different targets. Best performing matrices from all three approaches presented in this study were compared to the basic +2 (match) -1 (mismatch) scoring scheme in combination with gap open penalty = 5 and gap extension penalty = 1 (baseline total averaged BEDROC = 0.28)

Significance Assessment

We compared the retrospective performance of PhAST with each new scoring matrix to the result obtained with the original matrix (Table 3). To assess whether differences in screening performance are statistically significant we used a paired permutation test¹³ that was recently found to be the most powerful available significance test for this purpose.²⁰ It has the null hypothesis that *virtual screening method P performs significantly better than method Q*. Assuming p and q are rank lists of actives resulting from the virtual screening methods, the null hypothesis requires that $BEDROC(p) > BEDROC(q)$. As each active has two ranks, one in p and one in q , new rank lists p^* and q^* can be created by swapping ranks in p with corresponding ranks in q for each active with a probability of 50%. This was repeated 10,000 times and the frequency of the event that $(BEDROC(p) - BEDROC(q)) < (BEDROC(p^*) - BEDROC(q^*))$ is the type I error rate for the null hypothesis, which was used as p -value for significance estimation. As significance levels we used 0.05 (5%) and 0.01 (1%).

As we performed the paired permutation test with 10^4 permutations, the lowest measurable p -value equals 10^{-4} and results if only for one out of 10^4 permutations $(BEDROC(p) - BEDROC(q)) < (BEDROC(p^*) - BEDROC(q^*))$.

Weighted PhAST

1
2
3 The flexible scoring system of PhAST allows for incorporation of *a priori* knowledge about
4 important PPPs into the alignment process. The idea is similar to the construction of a
5 position-specific score matrix,^{9,10} but with our approach it is constructed implicitly. Utilizing
6 a general score matrix, the alignment of symbol x_i from PhAST-sequence X (query
7 compound) and symbol y_j from PhAST-sequence Y (library compound) is scored with score
8 $s(x_i, y_j)$. If a weighting factor w_i is applied to symbol x_i the score $w_i * s(x_i, y_j)$ is computed
9 instead. The unweighted version of PhAST is identical to the application of weight 1 to every
10 position of X . We applied weighted PhAST to four peroxisome proliferator activated
11 receptor gamma (PPAR γ) agonists (Figure 1). For these molecules it is known that a negative
12 charge in the carboxyl group is critical for receptor activation.²¹ We used weights 1, 2, 3, 4, 5,
13 10, 15, and 20, where a weight value of 1 corresponds to 'no weighting', in combination with
14 the original PhAST scoring matrix (Table 3).¹

15
16 As an alternative, using explicit of match- and mismatch-scores for each position in a
17 query sequence would be possible with a position-specific scoring matrix. We chose implicit
18 weighting factors as our aim was the extension of PhAST to position-specificity in a simple
19 way. For explicit position-specific scores, each match- and mismatch-score has to be
20 determined, which is a more complex problem and might be not as intuitive as our simplistic
21 approach.
22
23
24
25
26

27 **Pharmacophore Elucidation**

28
29 Given sets of active and inactive compounds for a certain target, the idea of weighted PhAST
30 can be used for automated pharmacophore elucidation through interaction profiling for a
31 particular query structure. For this purpose we applied weights from the interval [-20,20] to
32 each position of a query PhAST-sequence and performed retrospective screening using the
33 dataset of known actives and inactives. If retrospective performance increases with a specific
34 weight at a particular position, then the increase in performance indicates relative greater
35 importance of this particular pharmacophoric feature. Positive weights identify important
36 potential pharmacophoric points. Negative weights indicate variable positions in PhAST-
37 sequences for which mismatches are tolerated. Applying the paired permutation test to ranked
38 lists received from weighted and unweighted screenings allows for the identification of
39 features responsible for significant differences. As example we used actives and decoys for
40 PPAR γ available in the COBRA dataset, as we know that for the four PPAR γ agonists
41 depicted in Figure 1 the only negative charge present in these molecules should be identified
42 as the most important feature.
43
44
45
46
47
48

49 **Screening Protocol 2**

50
51 Effects of weighted features on retrospective performance were evaluated by retrospective
52 screenings using the COBRA library. As the PPAR γ agonists used as queries with weighted
53 features were taken from this dataset, the query molecule was removed from the screening
54 library. We performed exactly one screening for each combination of a query and a weighting
55 scheme. The resulting ranked lists were evaluated using the BEDROC metric ($\alpha = 20$).¹²
56 Significance of improvements was assessed by computing *p*-values in a paired permutation
57 test.¹³
58
59
60

Library Preparation

Hähnke et al.

9

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

All molecules were protonated using the 'wash' function of MOE (Molecular Operating environment, version 2010.06, Chemical Computing Group Inc., Montreal, Canada). For similarity assessment between molecules we calculated MACCS keys¹⁷ and the Tanimoto coefficient¹⁸ as implemented in MOE.

For Peer Review

RESULTS AND DISCUSSION

Score Matrices

We deduced log-odds scores for the alignment of symbols representing potential pharmacophoric points (PPPs) from symbol frequencies observed in a reference set of compounds. Observed exchange frequencies of PPPs were estimated from pairwise kernel-based assignments of graphs of potential pharmacophoric points as well as pairwise global sequence alignments.

Assignment-based scoring

For each matrix created based on pairwise assignments of vertices between graphs of potential pharmacophoric points attained using the ISOA kernel we performed a grid-search for best performing gap penalties (Table 7). For all subsequent calculations, scoring matrices were only used in combination with the best performing set of gap penalties.

For all three values of γ , best retrospective performance was observed with a similarity threshold of 0.9 for the molecule pairs used as reference. Differences in averaged retrospective performance were marginal, but highest performance (BEDROC = 0.35) was observed with the matrix calculated from ISOA kernel assignments generated with $\gamma = 0.75$ and similarity threshold = 0.9, in combination with gap open penalty = 9 and gap extension penalty = 1. This score matrix is presented in Table 8. The superiority of the original PhAST score matrix with average retrospective performance of 0.40 is significant: We calculated the percentage of screenings per target in which PhAST with the original score matrix outperforms PhAST with the best performing ISOA kernel score matrix and *vice versa* (Table 9). On all targets and both significance levels PhAST with the original score matrix performed significantly and exclusively better to a higher percentage, with values above 50% for COX2, DHFR, FXA, and THR. Averaged over all targets this superiority manifests in a total of 67% (65%) for all screenings at 0.05 (0.01) significance level. Notably, the best performing ISOA kernel score matrix performs significantly better than simple scoring (averaged BEDROC = 0.28) on both significance levels in more than 50% of all screenings with only three exceptions: COX2 at 0.05 and 0.01, and PPAR γ at 0.01 (Table 9).

Score matrices created with different parameterizations of the ISOA kernel and similarity thresholds are not only similar judged from their retrospective performance but from the actual scores as well. The averaged absolute difference per score between the original PhAST score matrix and the best performing ISOA kernel score matrix was 9.02, which is even smaller than the highest averaged difference within the ISOA kernel matrices (9.69 between matrices calculated with $\gamma = 0.5$ (0.75) and similarity threshold 0.9 (0)) (see also the complete dissimilarity matrix in supplemental Table S1). For each kernel matrix, the remaining matrices calculated with the same similarity threshold are most similar, indicating that changing this parameter has a greater overall effect than kernel parameterization. The good agreement of matches between the original PhAST score matrix and the best performing ISOA kernel score matrix is reflected in a mean difference of match scores of 4.78 and 10.08 for mismatch scores. Taking relations of match scores into account both matrices agree as well: In both matrices the most frequent symbols (R, L, O) have the lowest scores for matches. On the other hand, rarely occurring symbols (P, N, E, Q) correspond to high match scores. Divergences are bigger for mismatch scores, with the maximum difference of 24 for the mismatch (A,Q) and (L,Q).

Alignment-based scoring

Figure 2 presents the development of retrospective performance for scoring matrices created through iterated alignment and determination of symbol- and symbol-alignment frequencies necessary for calculation of log-odds scores (see also supplemental Table S2 and Table S3).

At no point during this process a matrix was created with better retrospective performance than the original PhAST score matrix. Iteration processes starting from the original PhAST score matrix lost retrospective performance in each step. Although processes starting from the +2 (match) -1 (mismatch) scoring scheme yielded matrices with a slightly increased retrospective performance after the first step, after convergence the final matrices were always inferior. Irrespective of the starting conditions, all final matrices showed an averaged retrospective performance between BEDROC = 0.33 and 0.35 compared to BEDROC = 0.40 of the original PhAST scoring matrix. The best performing (converged) matrix created in this approach (starting from +2/-1 scoring and a similarity threshold of 0.9) with averaged retrospective performance of BEDROC = 0.34 is presented in Table 10. The superiority of the original PhAST score matrix is significant (Table 11): On five out of six targets (with exception of ACE) and on both significance levels PhAST with the original score matrix performed significantly better to a higher percentage, with values above 50% for COX2, DHFR, FXA and THR, with the opposite being true only for ACE at 0.01 significance level. Averaged over all targets this superiority manifests in 69% (66%) of all screenings at 0.05 (0.01) significance level. Still, the best performing iterated alignment score matrix performs significantly better than simple scoring (averaged BEDROC = 0.28): Except for COX2, simple scoring was significantly outperformed on each target in more than 50% of all screenings at the 0.05 significance level, and at 0.01 except for COX2 and PPAR γ .

The mean averaged absolute score difference between all matrices created through iterated alignment is 6.34, indicating that these matrices diverge more from each other than the matrices created using ISOAK assignments (see the full dissimilarity matrix in supplemental Table S4). The averaged difference per score between the original PhAST score matrix and the best performing matrix created in our iterated approach was 13.8, i.e. greater than the largest distance between any pair of these matrices. As for the ISOAK matrices, agreement between both matrices is best for match scores with an averaged match score difference of 4.4 compared to 16.2 for mismatch scores. The relation of match scores agrees also: Matches of L, R, and O get lowest scores, and matches between P, N, E, and Q receive high scores. Scores for mismatches concur less well, with the highest difference of 37 for the (A, L) mismatch.

Summarizing, both of our systematic approaches to generating scoring matrices for PPPs from a set of reference compounds yielded matrices that perform significantly better than simple +2/-1 scoring. Still, none of the resulting matrices exhibited better averaged retrospective performance than the original PhAST score matrix (Table 3). The best performing matrices from both approaches agree well with an averaged difference per score of 5.44. In both approaches highest performance resulted from using only closely related molecule pairs from the collection of reference compounds, indicating that the quality of the reference set has an influence on matrix properties. As none of the used datasets was built explicitly to contain only examples of observed bioisosterisms, this might present a possibility to further increase the quality of systematically constructed scoring matrices for molecular similarity assessment.

Stochastically optimized scoring

We performed stochastic optimization of match and mismatch scores using the averaged retrospective performance on nine targets from the Krier dataset as evaluation function. The best performing matrix from each generation was evaluated in a set of retrospective screenings on the COBRA library as test dataset. Results for three independent optimization runs are presented in Figure 3. For all three optimizations, retrospective performance steadily

1
2
3 increases on the training dataset. Retrospective performance on the test dataset increases
4 overall as well. For both datasets the increase flattens after 20 generations. At no time the
5 performance on the test dataset decreases strongly, suggesting that the optimization does not
6 greatly suffer from overfitting to the training data. The matrix with the highest retrospective
7 performance was created in run no. two in generation 85 yielding an averaged BEDROC of
8 0.38. Still, we chose the final matrix based on retrospective performance on the training
9 dataset for further experiments; for all optimization runs this was the best performing matrix
10 from the last optimization cycle (averaged BEDROC: 0.37) (Table 12).

11
12
13 Retrospective performance of the best stochastic scoring matrix was inferior to the
14 original PhAST score matrix (averaged BEDROC = 0.40). This difference is significant
15 (Table 13): For three targets (COX2, DHFR, THR) the best stochastic scoring matrix is
16 outperformed in more than 50% of all screenings at both significance levels with the opposite
17 being true only in one case (FXA). Averaged over all targets, this superiority manifests in
18 53% (51%) of all screenings at 0.05 (0.01) significance level, whereas the best stochastic
19 scoring matrix performs significantly better only in 35% (32%) of all screenings. On the other
20 hand, the best performing stochastic scoring matrix outperforms simple +2/-1 scoring in 84%
21 (81%) of all screenings at the 0.05 (0.01) significance level. Its averaged retrospective
22 performance (BEDROC = 0.37) is statistically significantly better compared to the best
23 ISOAK score matrix (BEDROC = 0.35) and iterated alignment score matrix (BEDROC =
24 0.34).

25
26
27 Compared to the matrices created using the ISOA kernel and iterated alignment,
28 matrices resulting from stochastic optimization are more diverse: the top-performing matrix in
29 the last cycle of run no. two has an averaged *per*-score difference to that of the first (third)
30 run of 16.9 (10.4). The top-performing matrices from the last optimization cycle of the first and
31 second run diverge by 15.2. The best-performing stochastic score matrix diverges from the
32 original PhAST score matrix to an even higher degree, as the averaged *per*-score difference
33 between those two matrices is 16.8. This value is mainly influenced by scores for matches, for
34 which alone the averaged difference is 30.8. With 13.4 the averaged difference for
35 mismatches is comparatively small. But even with the high divergence of match scores
36 relations between them agree fairly well: Matches for L, O, and R received the lowest score
37 (only the score for an (A,A) match is lower); and P, N, E, and Q are scored higher.

38
39
40 Summarizing, these results demonstrate that stochastic optimization yielded score
41 matrices that significantly outperform simple scoring and the score matrices attained using the
42 presented systematic approaches for score calculation. But even the best matrix resulting from
43 this approach is inferior to the original PhAST score matrix.

44
45
46 Despite the fact that retrospective performance differs significantly between matrices,
47 and scores for matches and mismatches differ highly in some cases, all matrices seem to
48 describe connatural relationships between PPPs. We brought scores for matches and
49 mismatches in sequential order for each of the four matrices (row-to-row concatenation in the
50 order given in this publication) and calculated Pearson's correlation coefficient (Table 14).
51 Similarity of score relations is highest between the best performing ISOA kernel and iterated
52 alignment matrices with a correlation of $r = 0.9$. The matrix with most dissimilar relations to
53 all other matrices is the original PhAST score matrix. With an averaged correlation coefficient
54 of $r = 0.68$ it still has scoring principles common to the other ones. The scores in the original
55 PhAST score matrix were determined based on chemical intuition and PPP frequencies.
56 Apparently, there are certain properties and relations between molecular features that were
57 recognized by chemists and computer-based methods alike.

58 59 60 Weighted PhAST

We evaluated the influence of the application of weights to known key interaction features in virtual screenings with PhAST using the example of PPAR γ agonist retrieval. For four test cases of PPAR γ agonists, a set of weights was applied to a negative charge known to be essential for activity and the resulting ranked lists were evaluated using the BEDROC metric (Table 15). In all cases hit retrieval was significantly improved. Each combination of query and PPP feature weights outperformed unweighted virtual screening. Best results were obtained with different weights for different queries. Highest retrospective performance averaged over all four PPAR agonists results from the application of a weight value of 10. For each query, the paired permutation test between the unweighted and the best performing weighted screening resulted in a p -value $< 10^{-4}$, meaning that the improvements are statistically significant. The outcome of this preliminary study demonstrates that by incorporation of knowledge about relevant PPPs virtual screening by PhAST can be improved. Which particular weighting scheme should be applied certainly depends on the target, the particular interaction type, and the dataset used for screening.

Pharmacophore Elucidation

Finally, we used weighted PhAST in a series of retrospective screenings for automated identification of important PPPs. Again, PPAR γ agonists served as an example. For our four test compounds the weights for each position resulting in highest retrospective performance are given in Table 16. Notably, up-weighting the sole negative charge present in the query compounds resulted in highest retrospective performance compared to each other combination of weight and position in the PhAST-sequences. The increase in retrospective performance is significant ($p < 10^{-4}$). This result proves that our approach can be used to identify critical PPPs.

Systematic application of weights revealed additional PPPs whose weighting significantly increased retrospective performance ($p < 10^{-4}$). For each query, we selected the four PPPs causing highest performance when emphasized with the identified best weight (Figure 4) and used these for retrospective screening. For query structure A (B) [C] {D} screening performance was again significantly improved from BEDROC = 0.24 (0.29) [0.12] {0.15} to 0.36 (0.54) [0.42] {0.39} with p -values $< 10^{-4}$. For query structure A, the improvement is smaller than with weighting of the negative charge alone (BEDROC = 0.40). Visual inspection of the molecule revealed that the acceptor functionality identified as a potentially important interaction is part of the same carboxyl group as the negative charge. So weighting this substructure twice might be redundant. When the weight of this acceptor was decreased to one and the screening was repeated with only the remaining three weights retrospective performance increased to BEDROC = 0.44 ($p < 10^{-4}$). Surprisingly, weighting of the acceptor functionality in the carboxyl group of query B did not cause a decrease in retrospective performance, and when omitted caused performance to decrease to BEDROC = 0.5. In this case the acceptor functionality weighted with -1 resulted in best performance, inverting the scores of the original PhAST score matrix (Table 3). Apparently, exchanges at this position are rewarded, contrary to query structure A where this feature was weighted with weight = 5, preferring a conservation of this functionality at this position. Apparently favoring exchange of functionality at this position is beneficial. Some additional potential pharmacophoric points of types L and O were suggested as 'important' interaction. This exposes the significance of increase in retrospective performance as a necessary but insufficient indication for the relative importance of a PPP.

CONCLUSIONS

In this study we investigated the influence of modified scoring schemes on our text-based virtual screening approach PhAST. The impact of this study is three-fold.

First, we proposed and validated three approaches for the generation of scoring systems for potential pharmacophoric points. These approaches can be applied to custom sets of reference compounds and yield scoring systems that are significant improvements to simple scoring (matches +2, mismatches -1). The two systematic methods calculate scores within minutes. Runtime of stochastic optimization is subject to parameterization. Besides their general applicability, these methods can be used in the future for rapid evaluation of modifications to the pharmacophore model employed in our screening method PhAST. Matrix quality might be increased by a set of reference compounds containing proven examples of bioisosterism, as we could show that retrospective performance increases if only very similar molecules are used for score calculations.

Second, we demonstrated the importance of knowledge about receptor-ligand interactions being incorporated into the virtual screening process. With our screening method PhAST as example we could show that such weighting of interactions significantly improves screening performance. Previously,³ we demonstrated that the overall performance of PhAST is comparable to other ligand-based virtual screening techniques. This increase in screening performance renders PhAST a potentially valuable tool for hit retrieval from very large compound collections.

Third, we demonstrated the pharmacophore elucidation capabilities of text-based virtual screening. Given a set of compounds with known pharmacological (in)activity, PhAST may be used to construct a set of optimal positional weights that indicate key interaction points common between active compounds. This set of weights can instantly be used for weighted prospective screenings. Expanding the protocol used for pharmacophore elucidation to the generation of a complete set of match- and mismatch-scores for a position instead of a weight applied to all scores might also provide reasonable suggestions for substitutions of functional groups in a molecule. This aspect will be subject to further investigations.

ACKNOWLEDGEMENTS

We are grateful to the Chemical Computing Group Inc. for generous software support.

REFERENCES

- 1 Hähnke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. *J Comput Chem* 2009, 30, 761.
- 2 Hähnke, V.; Rupp, M.; Krier, M.; Rippmann, F.; Schneider, G. *J Comput Chem* 2010, 31, 2810.
- 3 Hähnke V.; Klenner, A.; Schneider, G. *J Comput Chem* submitted.
- 4 Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis*; Cambridge University Press: Cambridge, 1998.
- 5 Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1978.
- 6 Henikoff, S.; Henikoff, J. G. *Proc Natl Acad Sci* 1992, 89, 10915.
- 7 Rupp, M.; Proschak, E.; Schneider, G. *J Chem Inf Model* 2007
- 8 Krier, M.; Hutter, M. C. *J Chem Inf Model* 2009, 49, 1280.

Hähnke et al.

15

- 1
2
3
4 9 Eidhammer, I.; Jonassen, I.; Taylor, W. R. Protein Bioinformatics; John Wiley & Sons
5 Ltd: West Sussex (England), 2004.
6 10 Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.;
7 Lipman, D. J. Nucleic Acids Res 1997, 25, 3389.
8 11 Schneider, P.; Schneider G. QSAR Comb Sci 2003, 22, 713.
9 12 Truchon, J. F.; Bayly, C. I. J Chem Inf Model 2007, 47, 488.
10 13 Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. BMC
11 Bioinformatics 2009, 10, 225.
12 14 Needleman, S. B.; Wunsch, C. D. J Mol Biol 1970, 48, 443.
13 15 Patani, G. A.; LaVoie, E. J. Chem Rev 1996, 96, 3147.
14 16 Sheridan, R. P. J Chem Inf Comput Sci 2002, 42, 103.
15 17 Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. J Chem Inf Comput Sci 2002,
16 42, 1273.
17 18 Jaccard, P. Bull Soc Vaudoise Sci Nat 1901, 37, 241.
18 19 Vingron, M.; Waterman, M. S. J Mol Biol 1994, 235, 1.
19 20 Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. Bioinformatics 2010, 26, 1348.
20 21 Zoete, V.; Grosdidier, A.; Michielin, O. Biochim Biophys Acta 2007, 1771, 915.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Legends to the Figures

Figure 1. Selected PPAR γ -agonists.¹¹ These four compounds were used as queries in weighted virtual screening and as test cases for automated pharmacophore elucidation with PhAST. The negative charge highlighted by circles is essential for PPAR γ activation.²¹

Figure 2. BEDROC performance of scoring matrices based on iterated alignment. O = original PhAST score matrix used for initial alignments, S = simple +2 (match) -1 (mismatch) scoring used for initial alignments, t = similarity threshold for aligned molecule pairs.

Figure 3. Performance of the stochastic score matrix optimization. For each iteration the retrospective performance on test and training dataset is given as averaged BEDROC score. A) first optimization run, B) second optimization run, C) third optimization run. Final averaged BEDROC scores on the test dataset are: A) 0.34, B) 0.37, C) 0.36.

Figure 4. Automated pharmacophore elucidation with PhAST. For each marked atom of the four PPAR γ agonists the weight of the corresponding pharmacophoric feature is given. Larger values indicate potentially greater importance.

Hähnke et al.

17

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60**TEXT FOR GRAPHICAL ABSTRACT**

Various scoring schemes assessing the similarity of potential pharmacophoric points in bioactive compounds were generated by algorithmic optimization, and compared to a scheme based on chemical intuition. We demonstrate that screening performance of text-based similarity searching is significantly increased by position-specific weighting of ligand-receptor interaction sites, and suitable sets of weights can be generated fully automatically.

For Peer Review

Table 1. Potential pharmacophoric points used in PhAST and their corresponding symbols.

possible interactions	symbol
hydrogen bond acceptor	A
charge positive	P
charge negative	N
lipophilic	L
aromatic	R
hydrogen bond acceptor, hydrogen bond donor	E
hydrogen bond acceptor, polar	Q
hydrogen bond acceptor, hydrogen bond donor, polar	U
no possible interactions	O

Table 2. MQL queries defining pharmacophoric points in PhAST. Symbols are assigned to atoms used in queries from left to right. Queries are used in the given order from top to bottom.

MQL query	PPP symbols
c	R
n	R
*[charge<0]	N
*[charge>0]	P
C(=O)-O-H	O;N;E
P(=O)-O-H	O;N;E
S(=O)-O-H	O;N;E
N[allHydrogens=0&totalConnections=3]	Q
N[allHydrogens=1&totalConnections=3](-C')-C'	U
N[allHydrogens=2&totalConnections=3]-C'	U
N[allHydrogens=1&totalConnections=2]=C'	E
N[allHydrogens=0&totalConnections=2](=C')-C'	A
O-H	E
C=O	O;A
C[!bound(~N)&!bound(~O)]~*[C F Cl Br I S]	L
Cl	L
Br	L
I	L
S[!bound(~N)&!bound(~O)]~*[C H]	L

Table 3. Original PhAST score matrix for matches and mismatches of potential pharmacophoric points.

	A	E	L	N	O	P	Q	R	U
A	8	2	2	-1	-2	-4	4	-4	-2
E		12	-4	-9	-4	-6	-4	-9	0
L			2	-2	-2	-2	-4	1	-6
N				10	-2	-6	-7	-4	-10
O					2	-2	-4	-4	-6
P						10	6	-5	4
Q							14	-9	6
R								3	-13
U									16

Table 4. Targets in the COBRA library version 6.1 used for retrospective virtual screenings. Shown are abbreviations used in this study as well as the number of active compounds. The total number of molecules in the COBRA library is 8311.

Target	Abbreviation	No. Actives
Angiotensine-converting Enzyme	ACE	34
Cyclooxygenase 2	COX2	136
Dihydrofolat-reductase	DHFR	64
Factor Xa	FXA	228
Peroxisome-proliferator activated receptor γ	PPAR γ	44
Thrombin	THR	183
Total		689

Table 5. 31 therapeutic classes of the Krier dataset used for score matrix calculation. For each class the number of compounds and the number of unique pairs with similarity above similarity threshold t measured by MACCS keys and the Tanimoto coefficient is shown. As additional constraint pairwise similarity has to be below 0.98 to exclude trivial analogues from score matrix calculation.

ID	Therapeutic class	No. Compounds	No. Pairs ($t = 0.0$)	No. Pairs ($t = 0.5$)	No. Pairs ($t = 0.75$)	No. Pairs ($t = 0.90$)
1	ACE inhibitors	43	902	742	128	16
2	anabolic steroids	51	1261	1092	592	144
3	androgens	39	726	697	338	120
4	angiotensin II-antagonists	25	300	272	23	1
5	antiarrhythmics (class III)	17	135	41	7	0
6	barbitals	23	252	250	103	13
7	benzodiazepams	97	4653	2825	203	33
8	beta-blockers	50	1224	1047	217	9
9	calcium channel blockers	30	435	259	52	9
10	carbonic anhydrase inhibitors	8	28	25	3	0
11	antifungals (Conazoles)	54	1424	1002	119	28
12	COX inhibitors	73	2628	520	77	24
13	dazoles	17	136	56	9	2
14	floxacin	39	740	677	270	51
15	histamine H1-antagonists	28	378	208	12	2
16	histamine H2-antagonists	26	325	132	15	3
17	HIV protease inhibitors	18	153	113	21	6
18	leukotriene antagonists	58	1653	221	7	0
19	local anesthetics (Caines)	64	2011	1257	135	21
20	nitrofuranes	29	405	374	40	2
21	penicillines and derivatives	163	13196	11420	1329	78
22	phosphodiesterase IV inhibitors	11	55	6	1	0
23	pramines	22	231	181	41	7
24	antiulcers (Prazoles)	18	153	87	27	2
25	progestogens	59	1688	1678	739	107
26	reverse transcriptase inhibitors	66	2145	1189	215	26
27	serotonin antagonists	25	300	164	14	1
28	sulfonamides	54	1427	1373	286	28
29	tetracyclines	18	152	152	86	24
30	anticoagulants	29	406	112	8	0
31	tyrosine kinase inhibitors	14	91	49	4	0
Total		1268	39613	28221	5121	757

Table 6. Relative frequencies of potential pharmacophoric point symbols in datasets. Shown are the percentages of each symbol in PhAST-sequences created from all molecules of the respective compound collection.

Symbol	COBRA	Krier
A	4.95	6.44
E	1.44	1.37
L	19.65	24.38
N	1.22	1.75
O	24.63	26.35
P	1.80	1.72
Q	1.58	1.99
R	41.61	33.49
U	3.11	2.49

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 7. Retrospective performance of score matrices for potential pharmacophoric points calculated based on ISOAK assignments. For all combinations of α in the ISOA kernel and the similarity threshold for molecule pairs used as reference the best performing combination of gap penalties and averaged retrospective performance are shown. t = similarity threshold for molecule pairs, GO = gap open penalty, GE = gap extension penalty.

ISOAK γ	t	GO	GE	$\bar{\phi}$ BEDROC
0.25	0.00	9	2	0.3349
	0.50	7	1	0.3353
	0.75	9	1	0.3396
	0.90	9	1	0.3505
0.50	0.00	9	2	0.3343
	0.50	7	1	0.3345
	0.75	9	1	0.3399
	0.90	9	1	0.3505
0.75	0.00	8	1	0.3329
	0.50	7	1	0.3359
	0.75	8	1	0.3386
	0.90	9	1	0.3506

Table 8. Best performing score matrix calculated based on ISOAK assignments.

	A	E	L	N	O	P	Q	R	U
A	12	1	-15	-4	-11	-11	-20	-23	-20
E		17	-19	-4	-12	-6	-5	-26	-8
L			3	-14	-11	-19	-28	-8	-15
N				19	-14	-4	-13	-10	-13
O					6	-11	-25	-13	-13
P						20	-12	-12	-2
Q							18	-24	-3
R								7	-14
U									18

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 9. Comparison of the best performing ISOAK score matrix to the original PhAST score matrix and simple +2 (match) -1 (mismatch) scoring. Shown are percentages of retrospective screenings PhAST employing one score matrix outperforms PhAST in combination with the other one at 0.05 (0.01) significance level.

	∅	ACE	COX2	DHFR	FXA	PPAR _γ	THR
ISOAK best	72 (69)	94 (91)	11 (9)	92 (91)	79 (78)	64 (59)	91 (88)
Simple	20 (19)	0 (0)	85 (83)	3 (3)	14 (14)	16 (14)	4 (3)
ISOAK best	19 (17)	26 (26)	13 (12)	6 (3)	28 (28)	32 (27)	8 (8)
PhAST original	67 (65)	41 (41)	85 (85)	88 (86)	66 (65)	39 (34)	84 (80)

For Peer Review

Table 10. Best performing score matrix based on iterated alignment.

	A	E	L	N	O	P	Q	R	U
A	11	1	-35	10	-14	-18	-20	-31	-20
E		16	-30	-14	-13	-5	-15	-26	-16
L			3	-28	-23	-26	-29	-28	-18
N				17	-24	-11	-13	-24	-13
O					7	-12	-25	-36	-13
P						20	-12	-22	-12
Q							18	-25	-14
R								7	-17
U									18

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 11. Comparison of the best performing score matrix calculated based on iterated alignment to the original PhAST score matrix and simple +2 (match) -1 (mismatch) scoring. Shown are percentages of retrospective screenings PhAST employing one score matrix outperforms PhAST in combination with the other one at 0.05 (0.01) significance level.

	∅	ACE	COX2	DHFR	FXA	PPAR _γ	THR
Iterated best	69 (67)	94 (91)	14 (13)	86 (84)	82 (80)	52 (45)	89 (86)
Simple	20 (19)	0 (0)	79 (79)	6 (5)	13 (12)	18 (14)	4 (3)
Iterated best	18 (16)	35 (29)	12 (12)	5 (3)	29 (27)	23 (18)	6 (6)
PhAST original	69 (66)	35 (26)	86 (86)	95 (92)	66 (66)	43 (41)	85 (83)

For Peer Review

Table 12. Best performing score matrix resulting from stochastic optimization.

	A	E	L	N	O	P	Q	R	U
A	8	-13	-15	1	-3	-32	-10	-33	0
E		76	-35	4	5	-40	30	5	-18
L			21	-24	-4	-20	-17	8	2
N				67	-31	-19	-9	-20	-3
O					23	0	-12	6	2
P						63	14	-8	-10
Q							27	4	8
R								17	2
U									52

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 13. Comparison of the best performing score matrix attained from stochastic optimization to the original PhAST score matrix, simple +2 (match) -1 (mismatch) scoring and the best performing score matrices calculated based on ISOAK assignments and iterated alignment. Shown are percentages of retrospective screenings PhAST employing one score matrix outperforms PhAST in combination with the other one at 0.05 (0.01) significance level.

	Mean	ACE	COX2	DHFR	FXA	PPAR γ	THR
Stochastic best	84 (81)	88 (88)	62 (57)	91 (91)	89 (88)	77 (68)	96 (96)
Simple	10 (10)	3 (3)	33 (32)	5 (3)	7 (6)	11 (11)	3 (3)
Stochastic best	35 (32)	32 (24)	28 (26)	0 (0)	68 (67)	48 (43)	35 (34)
PhAST original	53 (51)	41 (32)	68 (65)	98 (97)	26 (25)	27 (27)	60 (58)
Stochastic best	54 (50)	35 (21)	85 (85)	14 (14)	69 (68)	43 (41)	74 (74)
ISOAK best	31 (30)	35 (32)	12 (10)	73 (73)	24 (23)	23 (20)	22 (20)
Stochastic best	54 (52)	18 (15)	80 (79)	20 (19)	72 (69)	57 (52)	75 (75)
Iterated best	32 (29)	56 (47)	13 (11)	64 (61)	22 (20)	18 (14)	20 (20)
Iterated best	30 (26)	47 (35)	35 (31)	5 (5)	41 (36)	16 (14)	35 (33)
ISOAK best	47 (42)	18 (15)	41 (35)	78 (69)	48 (46)	52 (43)	45 (42)

Table 14. Comparison of score matrices by Pearson correlation coefficient.

	Original	ISOAK	Iterated	Stochastic
Original	1	0.76	0.67	0.62
ISOAK		1	0.90	0.72
Iterated			1	0.71
Stochastic				1

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 15. Retrospective performance of weighted PhAST. Shown are BEDROC scores calculated for screenings of PPAR γ -agonists with upweighted key interaction, the difference between unweighted and the maximum of weighted performance as well as the p-values calculated for the improvement from unweighted to the best performing weighted screening.

weight	Compound				Ø BEDROC
	A	C	B	D	
1	0.24	0.12	0.29	0.15	0.20
2	0.32	0.15	0.40	0.19	0.26
3	0.37	0.18	0.44	0.23	0.30
4	0.39	0.22	0.45	0.26	0.33
5	0.40	0.25	0.45	0.29	0.35
10	0.40	0.30	0.44	0.32	0.37
15	0.40	0.32	0.42	0.33	0.37
20	0.40	0.32	0.40	0.34	0.37
BEDROC (unweighted)	0.24	0.12	0.29	0.15	
BEDROC (max weighted)	0.40	0.32	0.45	0.34	
p-Value	< 10 ⁻⁴	< 10 ⁻⁴	< 10 ⁻⁴	< 10 ⁻⁴	

Table 16. Results of the pharmacophore elucidation for four PPAR γ -agonists. Shown are PhAST-sequences, weights resulting in highest retrospective performance, corresponding BEDROC score and significance estimation for performance increase as well as the BEDROC scores of the corresponding unweighted screenings (shown in parentheses). S = symbol corresponding to a potential pharmacophoric point, W = weight resulting in highest retrospective performance for this position, B = BEDROC score obtained with the given weight, P = p-value indicating the significance of increase in retrospective performance of the weighted screening performed with the given weight at this position; if no p-value is given, no weight could increase retrospective performance compared to the unweighted screening; “-” indicates p-values below 10⁻⁴. Bold positions were used in a combined retrospective screening.

	Structure A (BEDROC 0.24)				Structure B (BEDROC 0.29)				Structure C (BEDROC 0.12)				Structure D (BEDROC 0.15)			
	S	W	B	P	S	W	B	P	S	W	B	P	S	W	B	P
O	-1	0.26	-		R	1	0.29		L	20	0.14	0.00	R	14	0.17	0.00
O	-1	0.25	0.03		R	1	0.29		R	7	0.12	0.02	R	20	0.16	0.01
O	-1	0.25	0.07		R	0	0.30	0.01	R	11	0.12	0.42	R	20	0.16	0.00
L	0	0.25	-		R	1	0.29		R	0	0.12	0.09	R	13	0.16	0.00
O	3	0.25	0.14		R	0	0.30	0.24	R	0	0.12	0.06	R	-1	0.16	0.06
N	12	0.40	-		R	1	0.29		R	-1	0.12	0.13	O	20	0.19	0.00
A	5	0.29	-		R	0	0.30	0.08	R	4	0.12	0.00	R	12	0.17	0.00
O	-1	0.26	0.01		O	-2	0.32	0.00	L	-1	0.13	0.07	Q	2	0.15	0.36
O	-1	0.28	-		O	-1	0.32	0.00	O	16	0.16	0.00	O	19	0.17	0.00
O	0	0.25	0.08		R	10	0.32	0.00	O	14	0.15	0.00	O	19	0.17	0.00
L	20	0.30	-		Q	0	0.32	0.00	R	5	0.12	0.39	O	19	0.17	0.00
R	2	0.24	0.37		O	0	0.30	0.01	R	4	0.12	0.02	R	10	0.16	0.01
R	3	0.25	0.02		O	17	0.32	0.00	R	4	0.12	0.03	R	12	0.16	0.00
R	3	0.24	0.09		O	17	0.32	0.01	R	5	0.12	0.03	R	12	0.16	0.00
R	0	0.25	0.21		R	9	0.30	0.23	R	6	0.12	0.06	R	12	0.16	0.00
R	0	0.24	0.39		R	6	0.30	0.03	R	6	0.12	0.02	R	12	0.16	0.00
R	0	0.24	0.51		R	6	0.30	0.13	L	20	0.15	0.00	R	12	0.17	0.00
R	0	0.26	0.01		R	6	0.30	0.03	O	5	0.12	0.08	L	20	0.18	0.00
O	11	0.28	-		R	5	0.30	0.06	O	4	0.12	0.03	O	19	0.16	0.08
R	-1	0.27	-		R	8	0.30	0.14	U	0	0.16	0.00	O	20	0.17	0.00
L	10	0.26	0.01		L	20	0.34	0.00	R	20	0.12	0.03	U	0	0.20	0.00
R	2	0.24	0.46		O	0	0.29	0.46	A	-10	0.13	0.10	R	1	0.15	
R	2	0.24	0.46		O	0	0.32	0.00	N	20	0.32	0.00	A	8	0.16	0.06
R	3	0.24	0.16		O	-1	0.32	0.00	R	0	0.12	0.34	N	20	0.34	0.00
O	9	0.26	0.07		A	-1	0.35	0.00	R	1	0.12		R	0	0.16	0.17
R	0	0.24	0.31		N	5	0.45	0.00	O	0	0.12	0.21	R	0	0.16	0.00
L	15	0.28	-		O	4	0.33	0.00	R	1	0.12		O	6	0.16	0.14
R	0	0.25	0.01		L	-1	0.32	0.00	R	0	0.12	0.39	R	0	0.16	0.00
R	0	0.24	0.03		O	-1	0.33	0.00	R	0	0.12	0.15	R	0	0.16	0.02
R	0	0.25	-		O	-1	0.32	0.00	A	20	0.14	0.00	R	0	0.16	0.00
R	0	0.25	0.13		O	-1	0.33	0.00	R	0	0.12	0.07	A	20	0.17	0.03
R	1	0.24							R	0	0.12	0.29	R	-1	0.17	0.01
R	2	0.24	0.37						R	0	0.12	0.36	R	-1	0.18	0.01
									R	2	0.12	0.31	R	-1	0.18	0.00
									R	2	0.12	0.47	R	-1	0.16	0.14
									R	1	0.12		R	-1	0.17	0.05
													R	0	0.16	0.04

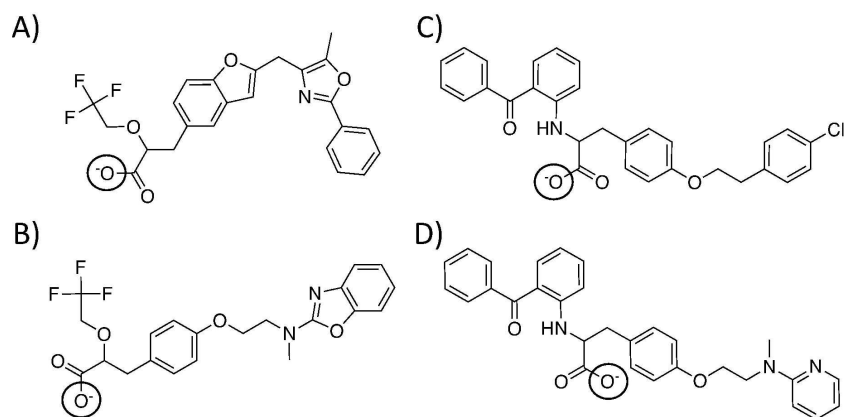


Figure 1
160x81mm (600 x 600 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

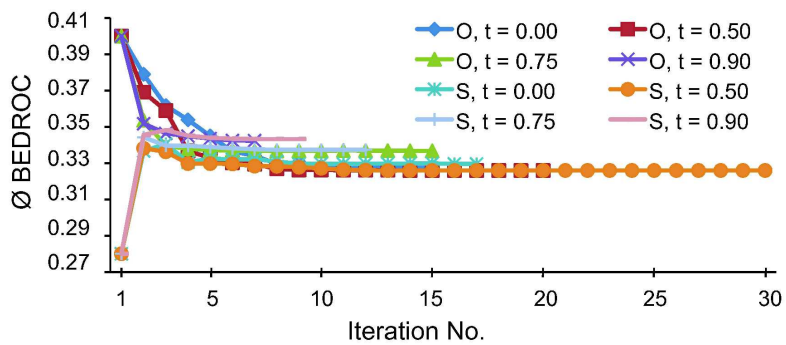
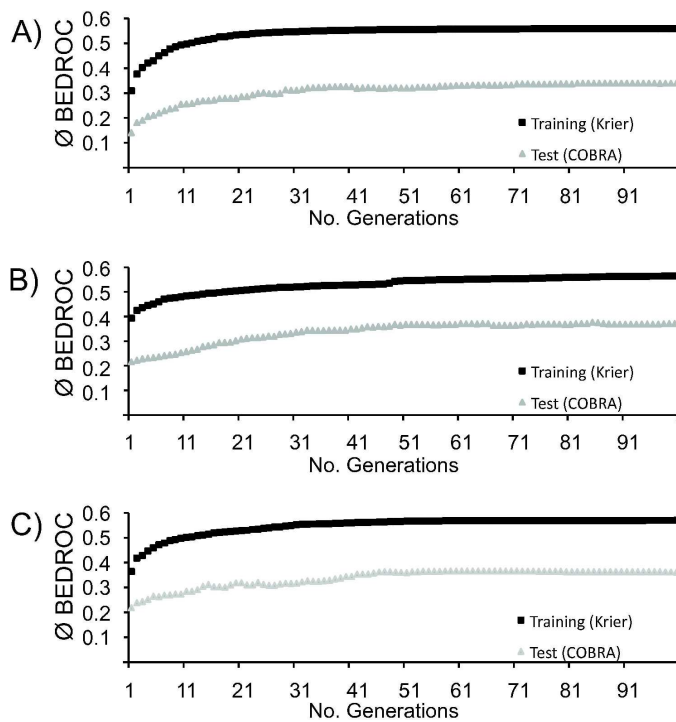
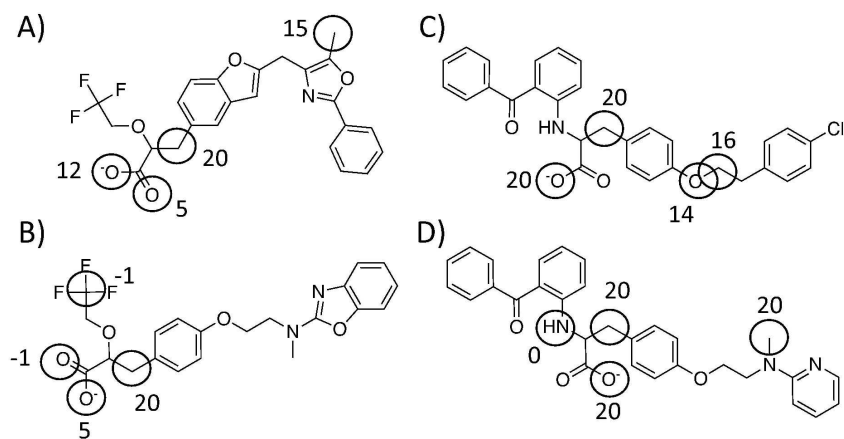


Figure 2
160x86mm (600 x 600 DPI)



160x137mm (600 x 600 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



160x86mm (600 x 600 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	A	E	L	N	O	P	Q	R	U
A	8	2	2	-1	-2	-4	4	-4	-2
E		12	-4	-9	-4	-6	-4	-9	0
L			2	-2	-2	-2	-4	1	-6
N				10	-2	-6	-7	-4	-10
O					2	-2	-4	-4	-6
P						10	6	-5	4
Q							14	-9	6
R								3	-13
U									16

70x55mm (600 x 600 DPI)

review

Table S2. Development of retrospective performance during iterated alignment. Shown is the averaged retrospective performance on the COBRA dataset, iterated alignment was performed on the Krier dataset. t = similarity threshold for aligned molecule pairs. The original PhAST score matrix was used as scoring scheme for the first alignment step, gap penalties were fixed at gap open penalty 5 and gap extension penalty 1.

Iteration	$t = 0.00$	$t = 0.50$	$t = 0.75$	$t = 0.90$
1	0.4001	0.4001	0.4001	0.4001
2	0.3790	0.3691	0.3536	0.3518
3	0.3617	0.3590	0.3396	0.3464
4	0.3539	0.3367	0.3379	0.3449
5	0.3450	0.3330	0.3373	0.3436
6	0.3365	0.3303	0.3370	0.3425
7	0.3341	0.3295	0.3367	0.3424
8	0.3297	0.3269	0.3368	
9	0.3305	0.3263	0.3368	
10	0.3282	0.3263	0.3368	
11	0.3280	0.3261	0.3370	
12	0.3280	0.3261	0.3369	
13	0.3280	0.3261	0.3369	
14	0.3280	0.3260	0.3368	
15	0.3280	0.3260	0.3368	
16		0.3260		
17		0.3260		
18		0.3260		
19		0.3260		
20		0.3260		

Table S3. Development of retrospective performance during iterated alignment. Shown is the averaged retrospective performance on the COBRA dataset, iterated alignment was performed on the Krier dataset. t = similarity threshold for aligned molecule pairs. Simple +2 (match) -1 (mismatch) scoring was used as scoring scheme for the first alignment step, gap penalties were fixed at gap open penalty 5 and gap extension penalty 1.

Iteration	$t = 0.00$	$t = 0.50$	$t = 0.75$	$t = 0.90$
1	0.2801	0.2801	0.2801	0.2801
2	0.3367	0.3382	0.3443	0.3456
3	0.3406	0.3363	0.3397	0.3482
4	0.3305	0.3298	0.3396	0.3452
5	0.3325	0.3298	0.3389	0.3440
6	0.3321	0.3296	0.3382	0.3435
7	0.3322	0.3284	0.3377	0.3433
8	0.3310	0.3283	0.3374	0.3433
9	0.3296	0.3278	0.3374	0.3433
10	0.3297	0.3272	0.3374	
11	0.3296	0.3264	0.3374	
12	0.3297	0.3262	0.3374	
13	0.3298	0.3259		
14	0.3298	0.3260		
15	0.3299	0.3260		
16	0.3296	0.3260		
17	0.3296	0.3260		
18		0.3260		
19		0.3260		
20		0.3260		
21		0.3260		
22		0.3260		
23		0.3260		
24		0.3260		
25		0.3260		
26		0.3260		
27		0.3260		
28		0.3260		
29		0.3260		
30		0.3260		

Appendix D

Authors: Zander, J.
Hartenfeller, M.
Hähnke, V.
Proschak, E.
Besier, S.
Wichelhaus, T. A.
Schneider, G.

Title: Multistep Virtual Screening for Rapid and Efficient
Identification of Non-Nucleoside Bacterial Thymidine Kinase
Inhibitors

Publication Year: 2010

Journal: Chemistry – a European Journal
Volume 16
Pages 9630-9637



Multistep Virtual Screening for Rapid and Efficient Identification of Non-Nucleoside Bacterial Thymidine Kinase Inhibitors

Johannes Zander,^[b] Markus Hartenfeller,^[a] Volker Hähnke,^[a] Ewgenij Proschak,^[c] Silke Besier,^[b] Thomas A. Wichelhaus,^[b] and Gisbert Schneider*^[a]

Abstract: Antimicrobial activity of trimethoprim/sulfamethoxazole (SXT) against *Staphylococcus aureus* (*S. aureus*) is antagonized by thymidine, which is abundant in infected or inflamed human tissue. To restore the antimicrobial activity of SXT in the presence of thymidine, we screened for small-molecule inhibitors of *S. aureus* thymidine kinase with non-nucleoside scaffolds. We present the successful ap-

plication of an adaptive virtual screening protocol for novel antibiotics using a combination of ligand- and structure-based approaches. Two consecutive rounds of virtual screening and in vitro testing were performed that resulted in

Keywords: antibiotics · drug design · drug resistance · molecular modeling · protein models

several non-nucleoside hits. The most potent compound exhibits substantial antimicrobial activity against both methicillin-resistant *S. aureus* strain ATCC 700699 and nonresistant strain ATCC 29213, when combined with SXT in the presence of thymidine. This study demonstrates how virtual screening can be used to guide hit finding in antibacterial screening campaigns with minimal experimental effort.

Introduction

Staphylococcus aureus (*S. aureus*) causes multiple diseases ranging in severity from minor skin infections to life-threatening conditions, such as endocarditis, pneumonia, and sepsis.^[1] Methicillin-resistant *S. aureus* (MRSA) has been widespread and has become a serious pathogenic bacterium, leading to high morbidity and mortality.^[2,3] MRSA is not only resistant to treatment with β -lactams, but often also to other antibiotics such as aminoglycosides, macrolides, lincosamides, and fluoroquinolones, because many MRSA strains possess a multidrug resistant genotype. Moreover, the ap-

pearance of vancomycin and linezolid resistance limited options for therapy against MRSA.^[4,5] This evolution points to an urgent need for new anti-MRSA compounds and for the optimization of established ones with high antimicrobial activity.

Folic acid antagonists, such as trimethoprim/sulfamethoxazole (SXT), possess a wide antimicrobial spectrum and show good antimicrobial activity against *S. aureus* including MRSA.^[6,7] These bioactive agents inhibit different enzymatic steps of the folic acid pathway leading to cessation of the bacterial synthesis of deoxythymidine monophosphate (dTMP) by thymidylate synthase. However, several bacterial species including *S. aureus* possess an alternative pathway for synthesis of intracellular dTMP by uptake of extracellular thymidine and subsequent intracellular phosphorylation to dTMP. Thus, the effect of folic acid antagonists can be antagonized by a high extracellular thymidine concentration as detected in tissues with necrotic cells such as pus and sputum from cystic fibrosis patients.^[8–10] Indeed, there are several reports of unsuccessful treatment with folic acid antagonists, supposedly due to elevated thymidine concentrations in human tissues containing necrotic cells.^[9,11,12]

We recently showed that, in the presence of thymidine, simultaneous inhibition of the folic acid pathway by SXT and the bacterial thymidine kinase (TK; EC 2.7.1.21) by nucleoside analogues, especially halogenated 2'-deoxyuridine derivatives, results in synergistic antimicrobial activity against

- [a] M. Hartenfeller,* V. Hähnke, Prof. Dr. G. Schneider
Eidgenössische Technische Hochschule (ETH)
Institute of Pharmaceutical Sciences
Wolfgang-Pauli Strasse 10, 8093 Zurich (Switzerland)
Fax: (+41) 44-633-1379
E-mail: gisbert.schneider@pharma.ethz.ch
- [b] Dr. J. Zander,* Dr. S. Besier, Prof. Dr. T. A. Wichelhaus
Institute of Medical Microbiology and Infection Control
Hospital of Goethe-University
Paul-Ehrlich-Strasse 40, 60596 Frankfurt/Main (Germany)
- [c] Dr. E. Proschak
Institute of Pharmaceutical Chemistry
LIFF/OSF, Max-von-Laue-Strasse 9
Goethe-University Frankfurt, 60348 Frankfurt (Germany)
- [*] Both authors contributed equally to this work.

S. aureus.^[10] Halogenated 2'-deoxyuridine derivatives such as 5-chloro-2'-deoxyuridine (5-CldU) and 5-iodo-2'-deoxyuridine (5-IdU) have been shown to inhibit bacterial TK.^[13,14] However, nucleoside analogues can be associated with cytotoxicity when phosphorylated to triphosphates and incorporated into DNA, thereby leading to single-strand breaks.^[15,16] Screening for non-nucleoside analogues as potential thymidine kinase inhibitors is therefore of particular interest for the development of novel antibiotics.

This study was aimed at 1) screening for non-nucleoside analogue inhibitors of *S. aureus* thymidine kinase by multistep virtual screening, and 2) determining the in vitro activity of these thymidine kinase inhibitors against *S. aureus* in combination with SXT in the presence of thymidine.

Results and Discussion

Substances that interact with viral and human thymidine kinases have been studied for many decades and several compounds have been found that exhibit high antiviral or anticancer activity.^[17,18] In contrast, inhibitors of bacterial thymidine kinases have not attracted much attention in antibacterial research.^[10,15,19,20] In most bacteria intracellular dTMP can be synthesized by two different pathways, which suggests combinations of bioactive agents inhibiting both pathways simultaneously.^[10] Thymidine kinase inhibitors impair the salvage pathway for dTMP, which is initiated by thymidine kinase catalyzing the transfer of a gamma-phosphate group from adenosine-5'-triphosphate (ATP) to thymidine.^[21] Folic acid antagonists inhibit different enzymatic steps of the bacterial synthesis of methylenetetrahydrofolate, an essential cofactor of thymidylate synthase for generation of dTMP from deoxyuridine monophosphate (dUMP). Simultaneous inhibition of both pathways therefore results in an intracellular lack of dTMP^[22] and synergistic antimicrobial activity in the presence of thymidine.^[10]

Comparative protein model: Here we used a virtual screening protocol to find potential thymidine kinase inhibitors with non-nucleoside structures. A crucial step of our screening protocol comprised automated docking of selected compounds into a homology model of *S. aureus* thymidine kinase (*SaTK*). Several bacterial thymidine kinases can be crystallized, such as thymidine kinases from *S. aureus* (*SaTK*, PDB identifier: 3e2i), *Ureaplasma urealyticum*, *Bacillus cereus*, and *Bacillus anthracis*.^[13,20,23] As a crystal struc-

ture of *SaTK* in the presence of a natural ligand (thymidine) is not known, we used the structure of *Bacillus anthracis* thymidine kinase (*BaTK*, PDB identifier: 2j9r, resolution: 2.7 Å)^[20] as template for this purpose as the best available model. A sequence alignment between *BaTK* and *SaTK* exhibits sequence identity of 63% overall and 100% in the thymidine binding-site residues (Figure 1). Consequently,

<i>S. aureus</i> (ATCC 700699/ATCC 29213)	MYETYHSGWIEICITGSMFSGKSEELIRRLRRGIYAKQ
<i>B. anthracis</i> (PDB: 2j9r)	SH_MYLINQNGWIEVICGSMFSGKSEELIRRVRRTOFAKQ
<i>B. anthracis</i> (complete)	MGSSHHHHHSSGLVPRGSHMYLINQNGWIEVICGSMFSGKSEELIRRVRRTOFAKQ
<i>S. aureus</i> (ATCC 700699/ATCC 29213)	KVVVFKPAIDDRYHKEKVVSHNGNAIEAINISKASEIMTHNLTVNDVIGIDEVQFFD
<i>B. anthracis</i> (PDB: 2j9r)	HAIIVFKPCVKAVPVSAKIDIFKHITTEEMDVIAIDEVQFFD
<i>B. anthracis</i> (complete)	HAIIVFKPCIDNRYSEEDVVSHNGLKVKAVPVSAKIDIFKHITTEEMDVIAIDEVQFFD
<i>S. aureus</i> (ATCC 700699/ATCC 29213)	DEIVSIVEKLSADGHRVIVAGLDMDFRGEPFEMPMPKLMVSEQVTKLQAVCAVCGSS
<i>B. anthracis</i> (PDB: 2j9r)	GDIVEVVQVLNARGYRVIVAGLDQDFRGLPFGQVQQLMAIAEHVTKLQAVCSACGSP
<i>B. anthracis</i> (complete)	GDIVEVVQVLNARGYRVIVAGLDQDFRGLPFGQVQQLMAIAEHVTKLQAVCSACGSP
<i>S. aureus</i> (ATCC 700699/ATCC 29213)	SSRTQRLINGKPAKIDDPILVGNESYEPRCRAHHIVAPS
<i>B. anthracis</i> (PDB: 2j9r)	ASRTQRLIDGEPAAFDDPILVGNESYEPRCRHCHAVPTKQ
<i>B. anthracis</i> (complete)	ASRTQRLIDGEPAAFDDPILVGNESYEPRCRHCHAVPTKQ

Figure 1. Sequence alignment of thymidine kinases. The first two sequences represent the alignment that was used for the homology model (63% sequence identity). PDB entry 2j9r of *B. anthracis* thymidine kinase misses some parts of the complete sequence (highlighted by black boxes and white letters) in the complete protein sequence, third line. A continuous gap was inserted at the corresponding position. Complete sequence identity of binding pocket residues (gray boxes) can be observed.

the resulting homology model (*SaTK*) shows excellent structural agreement with the template (*BaTK*), especially in the thymidine binding site (Figure 2). A continuous sequence stretch from *BaTK* comprising 17 residues is missing in the template structure. This part is predicted to form a helix in the homology model of *SaTK*. TKs of ATCC 29213 and ATCC 700699 have perfect sequence identity. This justifies employing one homology model for both proteins.

We explicitly did not perform docking studies on an existing crystal structure of *SaTK* (PDB identifier: 3e2i).^[23] The need for a homology model regardless of an existing structure of the target protein is rationalized by the fact that the structure of *SaTK* has been crystallized in its *apo* form (i.e., no bound thymidine). It was shown that TKs of several microorganisms undergo substantial structural changes in a loop region forming the upper part of the binding pocket upon thymidine binding.^[24] This renders the existing X-ray structure of the *SaTK* in its *apo* form unsuitable for docking efforts. It is therefore not surprising that a comparison between the homology model of *SaTK* and the respective crystal structure of the *apo* form exhibits a relatively high root-mean-square deviation (RMSD) of 3.9 Å. This finding originates from 1) a deviation in the position of the loop region of *SaTK* that depends on the missing ligand binding and 2) the fact that a part of the sequence corresponding to the one missing in the template structure of *BaTK* is also absent in the structure of the *SaTK apo* form (Figure 3).

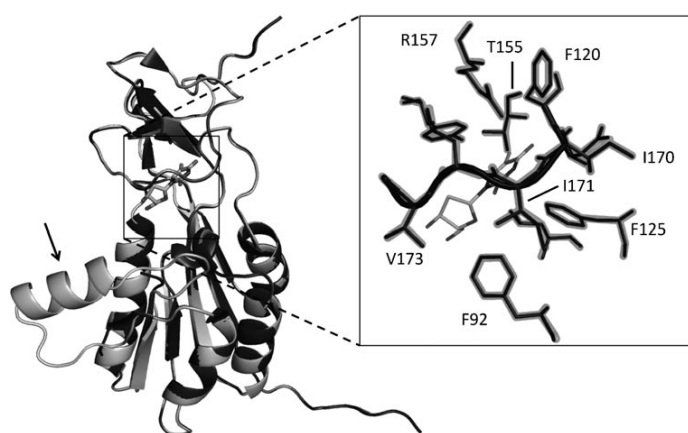


Figure 2. Binding-site model of *SaTK*. Left: Comparison of a homology model of *S. aureus* thymidine kinase and the template structure of *B. anthracis* thymidine kinase (PDB entry: 2j9r, chain A), together with bound native ligand thymidine. The missing part of the template (cf. Figure 1) is predicted to form a helix (arrow) flanked by two loop regions. Right: Perfect alignment between amino acid side chains of the model (transparent) and the template (solid). Identifiers of selected pocket residues of the model and a short stretch of the backbone (sketched) are shown for orientation.

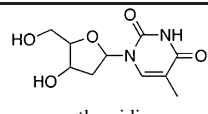
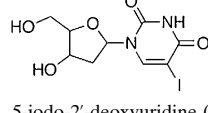
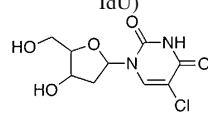


Figure 3. Comparative "homology" model of *SaTK*. Comparison of the homology model of *S. aureus* thymidine kinase (light gray) and an existing X-ray *apo* structure of the same protein (dark gray, PDB entry: 3e2i). Structural difference can be found mainly in the position of the loop defining the upper part of the binding cavity upon ligand binding (arrow). Bound glycerol (not shown) does not populate the thymidine binding pocket in structure 3e2i. As within the structure of thymidine kinase of *B. anthracis* that was used as template for homology modeling, an equivalent part of structure 3e2i is missing (dashed circle).

Reference ligands: For our ligand-based screening efforts we used 5-chloro-2'-deoxyuridine (5-CldU, **3**) with a minimal inhibitory concentration (MIC) of 0.0625 mgL⁻¹ against both *S. aureus* strains when combined with SXT in the presence of thymidine (**1**). The same MICs were determined for 5-iodo-2'-deoxyuridine (5-IdU, **2**), another ligand of bacteri-

al thymidine kinase.^[14] SXT alone in the presence of thymidine showed MIC values of >128 mgL⁻¹ against both *S. aureus* strains (not shown). 5-CldU and 5-IdU were chosen as reference ligands because halogenated 2'-deoxyuridine derivatives have recently been reported as thymidine kinase inhibitors showing significantly improved antimicrobial activity against *S. aureus* when combined with SXT in the presence of elevated thymidine concentrations.^[10] Moreover, Kosinska and co-workers showed that thymidine kinase from *Ureaplasma urealyticum* exhibits pronounced phosphorylation activity with 5-CldU as substrate.^[13] In a first study, we re-docked the natural ligand thymidine and the screening reference 5-CldU to obtain a reference value for the assessment of docking scores and to evaluate the performance of our docking protocol. Notably, automated ligand docking was able to reproduce the binding pose of thymidine. Thymidine and 5-CldU achieved favorable comparable docking scores of 37 and 38, respectively (higher docking scores suggest better ligand binding; Table 1).

Table 1. Reference compounds and values. Minimal inhibitory concentration (MIC) values measured for both *S. aureus* strains, and docking scores.

Structure	MIC [mgL ⁻¹]		Docking score (ASP)
	ATCC 700699	ATCC 29213	
<p>1</p>  <p>thymidine</p>	-	-	37
<p>2</p>  <p>5-iodo-2'-deoxyuridine (5-IdU)</p>	0.0625	0.0625	38
<p>3</p>  <p>5-chloro-2'-deoxyuridine (5-CldU)</p>	0.0625	0.0625	38

Virtual screening protocol: We followed a stepwise virtual screening protocol (Figure 4). A diverse screening library containing approximately 557 000 readily available compounds from two different suppliers was prepared. The first virtual screening step consisted of a rigorous reduction of the screening library ("negative design") by similarity analysis of pool compounds with the reference ligand 5-CldU (Table 1). For this purpose an in-house implementation of a self-organizing map (SOM)^[25] was employed to map the screening pool (represented in a high-dimensional space spanned by uncorrelated molecular descriptors) to a two-dimensional (2D) regular grid, as described.^[26] The SOM allowed for the identification of a cluster of 912 compounds

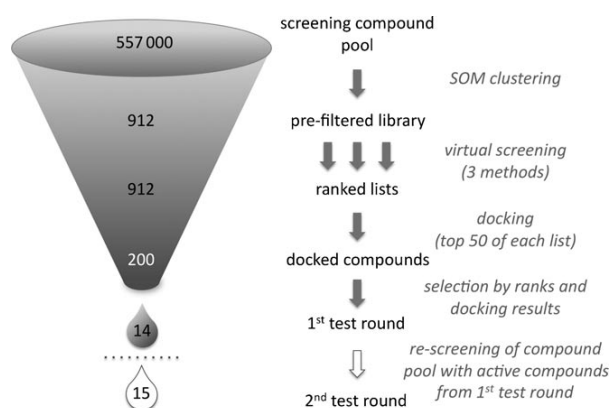


Figure 4. Virtual screening protocol. The second test round was performed on the complete screening compound library with the best hits from the first screening round.

that exhibit high similarity to the reference compound (“positive design”).

These candidate ligands were considered for the next screening steps. Three ligand-based screening techniques—each one focusing on a different aspect of ligand similarity—were applied on this small pre-filtered compound collection with respect to the same reference ligand (5-ClIdU):

- 1) The pharmacophore alignment search tool (PhAST)^[27] compares molecules by aligning strings of pharmacophoric feature types devised from their 2D representation.
- 2) Pseudoreceptor point similarity (PRPS)^[28] computes pseudoreceptor representations of molecules based on three-dimensional (3D) conformations.
- 3) ShaEP^[29] calculates a similarity score by comparing 3D conformers with respect to spatial overlap of shape and electrostatic potentials.

PhAST was applied in two different modes of structure canonization (for more information, see the Experimental Section) resulting in a total of four individual screening runs. Each method provided us with a sorted list of the remaining 912 screening compounds, ranked according to the scoring

schemes of the methods. Molecules ranked among the top 50 of each individual list were subsequently docked into a homology model of *S. aureus* thymidine kinase. Compounds from the top scoring ranks with plausible docking poses yielding high docking scores and hydrogen bridges similar to the reference ligands were considered for further investigation. We selected and ordered 14 compounds, which were tested in vitro for their biological activity on *S. aureus* thymidine kinase. A bacterial whole-cell assay was chosen to see whether virtual screening can cope with antibacterial activity without explicitly predicting this property. Out of the 14 tested compounds, seven compounds (**4–10**) exhibit antimicrobial activity against *S. aureus* strain ATCC 700699 and *S. aureus* ATCC 29213 when combined with SXT in the presence of thymidine (Table 2). None of these compounds had any intrinsic antimicrobial activity (data not shown). The fact that 50% of the 14 compounds chosen for in vitro screening showed antimicrobial activity when combined with folic acid antagonists argues for an effective first screening round. Based on the findings of the first screening two parallel strategies were applied to select compounds in a second screening round:

- 1) A second pseudoreceptor model using our software PRPS was employed to screen the complete compound

Table 2. Results of the first round of virtual screening and in vitro tests. MIC values represent the median of three experiments.

Structure	MIC [mg L ⁻¹]		Docking score (ASP)	Virtual screening rank			
	ATCC 700699	ATCC 29213		P1 ^[a]	P2 ^[b]	PRPS	ShaEP
	128	128	36	3	2	–	–
	128	128	38	11	41	–	–
	128	128	39	17	18	–	–
	128	>128	31	–	–	2	–
	128	128	26	–	–	6	–
	32	64	42	–	8	5	–
	32	64	41	–	–	–	18

[a] PhAST with Isomap canonization. [b] PhAST with Prabhakar canonization (cf. Experimental Section).

library ($\approx 557\,000$ molecules) again. The model was built from all seven active compounds found in the first screening round. Reference compounds were aligned according to their docking poses.

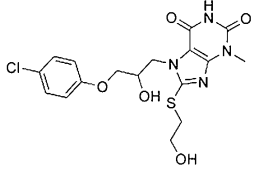
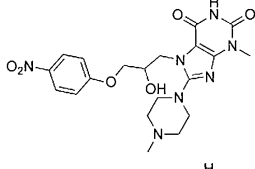
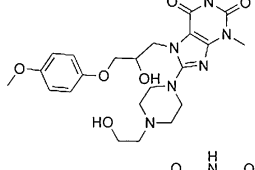
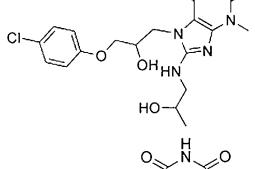
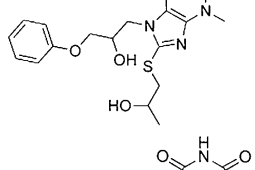
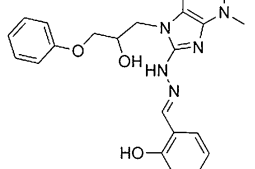
- 2) Compound **10** contains a catechol moiety that is buried deep inside the thymidine binding site according to the docking hypothesis. This “head group” is of particular interest as it is structurally distinct from nucleosides. Therefore, we performed a substructure search for compounds featuring this head group.

The top-scoring 100 molecules of the PRPS screening with the second pseudoreceptor model were docked into the homology model, and seven compounds were selected for testing, from which six compounds (**11–16**) exhibit antimicrobial activity in combination with SXT. Compound **16** has higher activity than the best compounds **9** and **10** found in the first screening round (Table 3). In addition, we retrieved 50 compounds containing the head group identified as promising in round one and docked them into the homology model. Compounds **17–24** were selected for testing according to plausibility of generated poses, high docking scores, and structural variations of the “tail group”. All eight substances exhibit the desired effect (hit rate 100%) with six compounds showing improved MIC values with respect to compounds of test round one. The most potent compound, **24**, exhibits a MIC value of 0.25 mg L^{-1} on both *S. aureus* strains when combined with SXT in the presence of thymidine, which is only fourfold less potent than 5-ClIdU and 5-IdU (Table 4). Again, docking of **24** suggests that the head group is buried in the binding pocket while the methylquinoline tail group interacts with the protein surface outside the cavity (not shown).

Compound **24** has a rather poor ligand efficiency^[30] [$LE = -\ln(\text{MIC})/(\text{no. of non-hydrogen atoms})$] of 0.06. For both reference compounds, 5-IdU and 5-ClIdU, we obtained an LE value of 0.16. Although the primary aim of this work was to identify non-nucleoside inhibitors of SaTK, the motivating findings suggest that there is room for further optimization with respect to both binding affinity and molecular mass.

Five compounds (**19**, **20**, **21**, **23**, **24**) exhibit intrinsic antimicrobial activity. MICs of these compounds in the presence of thymidine against *S. aureus* strains ATCC 29213 and ATCC 700699 are given in Table 5. MICs are substantially higher than those obtained in combination with SXT. The fact that the substances tested in this study showed no or only weak intrinsic antimicrobial activity is consistent with mainly thymidine kinase inhibition. It is known that some thymidine kinase inhibitors such as 5-fluoro-2'-deoxyuridine also inhibit thymidylate synthase and as a consequence have intrinsic antimicrobial activity.^[15] Future studies aiming at hit-to-lead structure optimization should use direct bacterial thymidine kinase inhibition assays to verify thymidine kinase being the target of these non-nucleoside antibiotics.

Table 3. Screening results of the second PRPS model based on the active compounds of the first round. MIC values represent the median of three experiments.

Structure	MIC [mg L^{-1}]		Docking score (ASP)
	ATCC 700699	ATCC 29213	
	128	128	45
	128	128	41
	128	128	41
	64	128	42
	32	128	43
	16	16	45

Conclusion

Our study demonstrates that multistep virtual screening can help identify bioactive substances from a large screening compound pool with limited experimental effort. Rapid focusing on promising candidate structures was possible, so that inhibitors of bacterial thymidine kinase with non-nucleoside scaffolds were identified. These inhibitory compounds exhibit moderate to high antimicrobial activity when combined with folic acid antagonists in the presence of thymidine, and provide rich opportunity for further optimization. Notably, at least two subsequent screening rounds were

Table 4. Results of substructure screening. The dihydroxyphenyl head group is preserved in all active molecules. MIC values represent the median of three experiments.

	Structure	MIC [mgL ⁻¹]		Docking score (ASP)
		ATCC 700699	ATCC 29213	
17		128	> 128	37
18		128	128	39
19		16	8	43
20		8	8	46
21		4	2	41
22		4	2	45
23		1	1	38
24		0.25	0.25	39

Table 5. Intrinsic antimicrobial effect of non-nucleoside analogues. MIC values were measured in the presence of thymidine (200 µg L⁻¹) and absence of SXT against both *S. aureus* strains. Values are medians of three experiments.

	MIC [mgL ⁻¹]	
	ATCC 700699	ATCC 29213
19	128	128
20	64	128
21	64	64
23	16	32
24	32	32

required to yield potent hits. The trick was to use information gained about the structuring of the chemical space spanned by the screening compound pool for “adaptive” optimization based on iterative learning.^[31] We suggest explo-

ration of the full potential of adaptive multistep and multi-method virtual screening in early drug discovery projects,^[32] which might speed up the transition from biological target validation to chemical hit and lead structure optimization.

Experimental Section

Strains and genetic sequence determination of bacterial thymidine kinase: *S. aureus* strain ATCC 700699 is resistant to methicillin (MRSA) and exhibits reduced susceptibility to vancomycin.^[33] The genetic sequence of its thymidine kinase-encoding *tdk* gene was published in 2001 as part of the whole genome sequence.^[34] Methicillin-susceptible *S. aureus* strain ATCC 29213 serves as a quality-control strain for antibiotic susceptibility testing.^[35] The chromosomal *tdk* gene of *S. aureus* strain ATCC 29213 was amplified by polymerase chain reaction (PCR) with forward primer P1 (5'-GCGAT-TATGTTTTGAAAAGGTGG-3') and reverse primer P2 (5'-GTTCGIATCTTCTTCTACAA-TATC-3'). The nucleotide sequence of the *tdk* gene of *S. aureus* ATCC 29213 was determined by cycle sequencing using an ABI PRISM DNA sequencer (Applied Biosystems, Foster City, USA).

Compound library: Virtual screening was performed with a structurally diverse set of compounds from supplier catalogues of Specs (v01/2009, Specs, Delft, The Netherlands) and Asinex Gold and Platinum collections (v11/2008, Asinex, Moscow, Russia). Protonation states of all compounds were standardized (“washed”) using the “wash” function of MOE (v2008.10, Chemical Computing Group, Montreal, QC, Canada). Single three-dimensional conformations for each screening compound were computed with the software CORINA (v3.2, Molecular Networks, Erlangen, Germany).

Self-organizing map: The reference compound 5-ClDU was added to the screening library before the calculation of all 184 2D descriptors of MOE for each molecule. Principal component analysis^[36] revealed that 95% of the variance in the dataset could be explained using the 40 first principal components, so these uncorrelated descriptors were used for representing the screening compound library. We used an implementation of the self-organizing map (SOM)^[25] algorithm to further reduce the dimensionality of the dataset.^[26] The SOM performed a nonlinear mapping from the original descriptor space (here: 40-dimensional) on a two-dimensional map. Each molecule is assigned to one of the receptive fields (clusters) of the SOM. We used a SOM with a topology of 20×30 neurons (600 receptive fields) organized as a torus. The SOM was trained in 5×10⁶ cycles. The parameter defining the decay of weight update during training was initialized with 1. The initial width of the Gaussian neighborhood function was 5. Distances were calculated as the Euclidean distance. From

the trained SOM we selected the 912 compounds assigned to the neuron containing the reference compound for the virtual screening process.

PhAST: The pharmacophore alignment search tool (PhAST) is a string-based approach to virtual screening.^[27] It reduces each molecule to an unambiguous linear representation describing its pharmacophoric features—called ‘PhAST-sequence’—in three steps: 1) each non-hydrogen atom in the structure graph is replaced by a potential pharmacophoric point symbol; hydrogen atoms are removed; 2) vertices of this pharmacophoric feature graph are canonically labeled, and 3) vertex symbols are concatenated into a string in increasing order of their canonical labels. For virtual screening, both the screening compound collection (‘library’) and the query molecules were converted and the resulting PhAST sequences were compared using pairwise global sequence alignment.^[37] As a result, molecular similarity values are computed from the pairwise alignments, which were used for the retrieval of molecules with similar pharmacophoric features from a compound database. PhAST distinguishes between nine different potential pharmacophore points: positive charge; negative charge; aromatic; lipophilic; hydrogen-bond donor; hydrogen-bond donor and acceptor; hydrogen-bond acceptor and positive charge; hydrogen-bond donor and acceptor and positive charge; no interaction. The original version of PhAST uses the algorithm of Weininger et al.^[38] for canonization. In this work we employed the algorithm by Prabhakar and Balasubramanian^[39] (referred to as ‘PhAST Prabhakar’) and the Isomap algorithm^[40] (referred to as ‘PhAST Isomap’). PhAST Prabhakar was used with gap open penalty=5 and gap extension penalty=1; PhAST Isomap with gap open penalty=8 and gap extension penalty=1. With all versions of PhAST the published standard score matrix was used.^[27] In contrast to the original version of PhAST, we calculated the alignment score normalized to the alignment length as a similarity measure between aligned sequences instead of sequence identity. These modifications were shown to be superior to the original approach.^[41]

PRPS: Pseudoreceptor point similarity (PRPS) is a virtual screening tool bridging receptor- and ligand-based screening techniques.^[28] Starting from a 3D conformation of a ligand, PRPS projects potential interaction points into the surrounding space mimicking a surrounding ‘idealized’ receptor pocket. Location of interaction points depends on known preferred distances and angles of the respective hypothetical interaction, assumed to be possible at this position of the ligand. The type of an interaction point (hydrogen-bond donor, hydrogen-bond acceptor, π stacking (‘aromatic’)) is complementary to the respective potential pharmacophoric point of the ligand. The spatial arrangement of generated interaction points is then transformed into an alignment-invariant representation as a cross-correlation descriptor. PRPS compares two molecules by calculating the Euclidian distance between their descriptor representations. A PRPS model can be computed for a single ligand or for a set of multiple ligands. In the latter case the model is built based on an alignment of all compounds, and projected interaction points are weighted by the number of molecules that projected them to the same location.

ShaEP: ShaEP is a tool for 3D ligand-based virtual screening that evaluates the similarity between two molecules by means of spatial overlap in volume and calculated electrostatic potential fields.^[29] Rigid body alignment of the molecules is performed to optimize overlaps. Ligand flexibility can be addressed implicitly by not only comparing a single conformation of both molecules but instead by performing an exhaustive pairwise comparison of conformation ensembles. For ShaEP screenings, up to 10 conformations of both the reference ligand and each screening compound were generated using the stochastic conformer generation routine of MOE. Partial charges for every conformation were calculated according to the MMFF94 parameter set available in MOE. Only the highest score of all pairwise comparisons was considered for the final ranking of screening compounds.

Homology model: A comparative protein model (‘homology model’) of SaTK was built using the web service of Swiss Model^[42,43] in automated mode. The crystal structure of *Bacillus anthracis* thymidine kinase (PDB identifier: 2j9r, chain A) served as template. The query sequence was derived from *S. aureus* ATCC 700699 thymidine kinase (access number: NP_372643).

Automated ligand docking: Docking experiments were performed using the software GOLD^[44] with the ASP scoring function. Residues F92, L116, D119, F120, F125, T155, R157, I170, I171, L172, V173, G174, and Y179 defined the binding site. Initial 3D conformations of docked compounds were calculated by CORINA prior to docking.

Microdilution assay: Minimal inhibitory concentrations (MICs) of the potential different thymidine kinase inhibitors alone and in combination with SXT against *S. aureus* strain ATCC 700699 and *S. aureus* ATCC 29213 in the presence of thymidine were determined according to Clinical and Laboratory Standards Institute (CLSI) guidelines with some modifications.^[34] Therefore, a bacterial suspension (95 μ L, exponential growth phase) of *S. aureus* strains (ca. 5×10^5 cells mL⁻¹) in cation-adjusted Mueller–Hinton broth (Becton, Dickinson and Company, Sparks, USA) supplemented with thymidine (200 μ g L⁻¹; Sigma–Aldrich, Munich, Germany) and with or without trimethoprim/sulfamethoxazole (40 mg) in a ratio of 1:19 (both Sigma–Aldrich) was added to each well of a 96-well microtiter plate (Greiner, Monroe, USA). A solution (5 μ L) of different potential thymidine kinase inhibitors in various dilutions was added to each well (range of final concentrations: 0.03125 to 128 mg L⁻¹). After 20 h of incubation at 37°C, MICs were determined. Experiments were performed in triplicate.

Acknowledgements

We thank Denia Frank and Simone Schermuly for excellent technical assistance. The Chemical Computing Group Inc. (Montreal, Canada) is thanked for generous software support.

- [1] F. D. Lowy, *N. Engl. J. Med.* **1998**, *339*, 520–532.
- [2] T. Kitahara, Y. Aoyama, Y. Hirakata, S. Kamihira, S. Kohno, N. Ichikawa, M. Nakashima, H. Sasaki and S. Higuchi, *Int. J. Antimicrob. Agents* **2006**, *27*, 51–57.
- [3] G. Y. Zuo, G. C. Wang, Y. B. Zhao, G. L. Xu, X. Y. Hao, J. Han, Q. Zhao, *J. Ethnopharmacol.* **2008**, *120*, 287–290.
- [4] F. W. Goldstein, *Clin. Microbiol. Infect.* **2007**, *13*, 2–6.
- [5] S. Tsiodras, H. S. Gold, G. Sakoulas, G. M. Eliopoulos, C. Wrennsten, L. Venkataraman, R. C. Moellering, M. J. Ferraro, *Lancet* **2001**, *358*, 207–208.
- [6] M. Adra, K. R. Lawrence, *Ann. Pharmacother.* **2004**, *38*, 338–341.
- [7] S. A. Grim, R. P. Rapp, C. A. Martin, M. E. Evans, *Pharmacotherapy* **2005**, *25*, 253–264.
- [8] S. Besier, J. Zander, E. Siegel, S. H. Saum, K. P. Hunfeld, A. Ehrhart, V. Brade, T. A. Wichelhaus, *J. Clin. Microbiol.* **2008**, *46*, 3829–3832.
- [9] R. A. Proctor, *Clin. Infect. Dis.* **2008**, *46*, 584–593.
- [10] J. Zander, S. Besier, H. Ackermann, T. A. Wichelhaus, *Antimicrob. Agents Chemother.* **2010**, *54*, 1226–1231.
- [11] J. S. Lockwood, H. M. Lynch, *JAMA J. Am. Med. Assoc.* **1940**, *114*, 935–939.
- [12] C. M. MacLeod, *J. Exp. Med.* **1940**, *72*, 217–232.
- [13] U. Kosinska, C. Carnrot, S. Eriksson, L. Wang, H. Eklund, *FEBS J.* **2005**, *272*, 6365–6372.
- [14] P. Voytek, P. K. Chang, W. H. Prusoff, *J. Biol. Chem.* **1972**, *247*, 367–372.
- [15] C. Carnrot, S. R. Vogel, Y. Byun, L. Wang, W. Tjarks, S. Eriksson, A. J. Phipps, *Biol. Chem.* **2006**, *387*, 1575–1581.
- [16] E. Galanis, R. Goldberg, J. Reid, P. Atherton, J. Sloan, H. Pitot, J. Rubin, A. A. Adjei, P. Burch, S. L. Safgren, T. E. Witzig, M. M. Ames, C. Erlichman, *Ann. Oncol.* **2001**, *12*, 701–707.
- [17] E. De Clercq, *Methods Find. Exp. Clin. Pharmacol.* **1980**, *2*, 253–267.
- [18] W. H. Prusoff, M. S. Chen, P. H. Fischer, T. S. Lin, G. T. Shiau, R. F. Schinazi, J. Walker, *Pharmacol. Ther.* **1979**, *7*, 1–34.
- [19] C. Carnrot, R. Wehelie, S. Eriksson, G. Boelske, L. Wang, *Mol. Microbiol.* **2003**, *50*, 771–780.

- [20] U. Kosinska, C. Carnrot, M. P. Sandrini, A. R. Clausen, L. Wang, J. Piskur, S. Eriksson, H. Eklund, *FEBS J.* **2007**, *274*, 727–737.
- [21] L. Stryer in *Biochemie* (Ed.: L. Stryer), Spektrum der Wissenschaft, Heidelberg, **1990**, pp. 627–653.
- [22] J. Zander, S. Besier, S. H. Saum, F. Dehghani, S. Loitsch, V. Brade, T. A. Wichelhaus, *Infect. Immun.* **2008**, *52*, 2183–2189.
- [23] R. Lam, K. Johns, K. P. Battaille, V. Romanov, K. Lam, E. F. Pai, N. Y. Chargadze, unpublished results.
- [24] D. Segura-Peña, J. Lichter, M. Trani, M. Konrad, A. Lavie, S. Lutz, *Structure* **2007**, *15*, 1555–1566.
- [25] a) T. Kohonen, *Biol. Cybern.* **1982**, *43*, 59–69; b) J. Zupan, J. Gastteiger, *Neural Networks for Chemists. An Introduction*, Wiley-VCH, Weinheim, **1999**.
- [26] a) P. Schneider, G. Schneider, *QSAR Comb. Sci.* **2003**, *22*, 713–718; b) P. Schneider, Y. Tanrikulu, G. Schneider, *Curr. Med. Chem.* **2009**, *16*, 258–266; c) S. Renner, M. Hechenberger, T. Noeske, A. Böcker, C. Jatzke, M. Schmuker, C. G. Parsons, T. Weil, G. Schneider, *Angew. Chem.* **2007**, *119*, 5432–5435; *Angew. Chem. Int. Ed.* **2007**, *46*, 5336–5339.
- [27] V. Hähnke, B. Hofmann, T. Grgat, E. Proschak, D. Steinhilber, G. Schneider, *J. Comput. Chem.* **2009**, *30*, 761–771.
- [28] a) Y. Tanrikulu, E. Proschak, T. Werner, T. Geppert, N. Todoroff, A. Klenner, T. Kottke, K. Sander, E. Schneider, R. Seifert, H. Stark, T. Clark, G. Schneider, *ChemMedChem* **2009**, *4*, 820–827; b) Y. Tanrikulu, G. Schneider, *Nat. Rev. Drug Discov.* **2008**, *7*, 667–677.
- [29] M. J. Vainio, J. S. Puranen, M. S. Johnson, *J. Chem. Inf. Model* **2009**, *49*, 492–502.
- [30] A. Hopkins, C. Groom, A. Alex, *Drug Discovery Today* **2004**, *9*, 430–431.
- [31] a) G. Schneider, M. Hartenfeller, M. Reutlinger, Y. Tanrikulu, E. Proschak, P. Schneider, *Trends Biotechnol.* **2009**, *27*, 18–26; b) G. Schneider, S.-S. So, *Adaptive Systems in Drug Design*, Landes Bioscience, Georgetown, **2002**; c) T. I. Oprea, *Curr. Opin. Chem. Biol.* **2001**, *5*, 384–389.
- [32] a) T. I. Oprea, H. Matter, *Curr. Opin. Chem. Biol.* **2004**, *8*, 349–358; b) H. Eckert, J. Barorath, *Drug Discovery Today* **2007**, *12*, 225–233; c) G. Schneider, *Nat. Rev. Drug Discovery* **2010**, *9*, 273–276.
- [33] K. Hiramatsu, H. Hanaki, T. Ino, K. Yabuta, T. Oguri, F. C. Tenover, *J. Antimicrob. Chemother.* **1997**, *40*, 135–136.
- [34] M. Kuroda, T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa, I. Kobayashi, L. Cui, A. Oguchi, K. Aoki, Y. Nagai, J. Lian, T. Ito, M. Kanamori, H. Matsumaru, A. Maruyama, H. Murakami, A. Hosoyama, Y. Mizutani-Ui, N. K. Takahashi, T. Sawano, R. Inoue, C. Kaito, K. Sekimizu, H. Hirakawa, S. Kuhara, S. Goto, J. Yabuzaki, M. Kanehisa, A. Yamashita, K. Oshima, K. Furuya, C. Yoshino, T. Shiba, M. Hattori, N. Ogasawara, H. Hayashi, K. Hiramatsu, *Lancet* **2001**, *357*, 1225–1240.
- [35] CLSI. Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard, 8th ed., M7A8, Clinical and Laboratory Standards Institute, Wayne, PA, **2008**.
- [36] K. Pearson, *Philos. Mag.* **1901**, *2*, 559–572.
- [37] S. B. Needleman, C. D. Wunsch, *J. Mol. Biol.* **1970**, *48*, 443–453.
- [38] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- [39] Y. S. Prabhakar, K. Balasubramanian, *J. Chem. Inf. Model* **2006**, *46*, 52–56.
- [40] J. B. Tenenbaum, V. de Silva, J. C. Langford, *Science* **2000**, *290*, 2319–2323.
- [41] V. Hähnke, M. Rupp, M. Krier, F. Rippmann, G. Schneider, *J. Comput. Chem.* **2010**, DOI: 10.1002/joc.21574.
- [42] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, *Bioinformatics* **2005**, *22*, 195–201.
- [43] Swiss Model: <http://swissmodel.expasy.org/> (accessed 25.02.2010).
- [44] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.

Received: May 17, 2010
Published online: July 20, 2010

18 - Curriculum Vitae

Name:	Volker Dirk Hähnke
Date of Birth:	05.11.1981
Place of Birth:	65929 Frankfurt am Main, Germany
Civil Status:	unmarried
Education:	
01/2010-01/2011	Finalization of Ph.D. studies at the Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, with Prof. Dr. Gisbert Schneider.
03/2008-12/2009	Ph.D. studies at the Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University, Frankfurt am Main, Germany, with Prof. Dr. Gisbert Schneider, supported by a full Ph.D. scholarship awarded by Merck KGaA, Darmstadt (Germany).
10/2002 – 11/2007	Studied Bioinformatics at Johann Wolfgang Goethe-University, Frankfurt am Main, Germany. University degree: Diploma in Bioinformatics. Average Grade: 1.0 (on a scale between 4.3 and 1.0, with 1.0 being best). Major subjects: <ul style="list-style-type: none">▪ Bioinformatics (Prof. Dr. Gisbert Schneider, Dr. Dirk Metzler)▪ Evolutionary Algorithms (Priv.-Doz. Dr. Markus Nebel)▪ Genetics (Prof. Dr. Pascal von Koskull-Döring)▪ Structure of Biomolecules (Prof. Dr. Joachim Engels)
07/2001 – 04/2002	Community service at German Red Cross, Wiesbaden, Germany.
07/1992 – 06/2001	Heinrich-von-Brentano secondary school, Hochheim, Germany. School diploma: University entrance diploma. Average grade: 1.2 (on a scale between 4.0 and 1.0, with 1.0 being best). Subjects: Biology, Social Studies, German, Music.